Received: 20 May 2022 | Accepted: 1 November 2023

DOI: 10.1002/berj.3934

ORIGINAL ARTICLE

BERJ BERA

# How robust are socio-economic achievement gradients using PISA data? A case study from Germany

**John Jerrim** | **Laura Zieger**

UCL Social Research Institute, London, UK

**Correspondence**
John Jerrim, UCL Social Research Institute. 55-59 Gordon Square, Bloomsbury, London, WC1H 0NU, UK.
Email: j.jerrim@ucl.ac.uk

**Funding information**
European Research Council

## Abstract

Large-scale international achievement studies such as PISA have been widely used to study how educational inequality compares across countries. Yet the various different biases that may affect these estimates are often not considered or are poorly understood. In this paper we draw upon the total survey error framework to provide a case study of the potential biases affecting estimates of the socio-economic achievement gaps using PISA data from Germany. The results illustrate how procedural and measurement errors have a substantial impact upon estimates of socio-economic achievement gradients in Germany, including how it compares with other countries. This leads us to conclude that estimates of socio-economic achievement gaps using the PISA data for Germany do not seem to be particularly robust. More generally, we argue that better articulation and reporting of such challenges with comparing socio-economic achievement gaps using large-scale international assessment data such as PISA are needed.

**KEYWORDS**
inequity and social justice, social class

## INTRODUCTION

International large-scale assessments (ILSAs), such as the Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA), aim to collect information about the cognitive skills of students around the world. For

**Key insights**

- This paper addresses the issue of potential bias that may impact estimates of socio-economic achievement gaps using PISA data.
- Estimates of socio-economic achievement gaps using the PISA data for Germany do not seem to be particularly robust.
- It is likely that ignoring the potential bias in the German PISA data is likely to lead to underestimation of differences in achievement between parental education groups.

more than 20 years they have had considerable influence on education policy and international education debates (Breakspear, 2012; Hopkins et al., 2008).

These studies do not only measure students' cognitive skills, but also administer background questionnaires. As a result, ILSAs have also been widely used to analyse the association between different background variables and student achievement across countries and over time (e.g. Davoli & Entorf, 2018; OECD, 2019). The influence of family background on educational achievement has been widely researched by academic economists, educationalists and sociologists (e.g. Davis-Kean, 2005; Guryan et al., 2008; Ludeke et al., 2021; Pishghadam & Zabihi, 2011; Yeung et al., 2002) as well as its potential mediating role between social origin and destination (e.g. Breen & Jonsson, 2005). This includes a substantial body of work investigating how such intergenerational patterns vary across countries, using data from PISA.

However, when conducting cross-national comparisons of educational inequality, one must be aware of the limitations with the data available. As early as the 1940s, research started to engage with the 'total survey error' (TSE) framework. This aims to evaluate the 'usefulness' and 'meaningfulness' of sample survey-based research (such as ILSAs) by comprehensively considering various sources of error and bias (Groves & Lyberg, 2010). Total survey error includes issues surrounding how well the sample reflects the underlying population of interest, encompassing survey non-response, missing data (amongst respondents) and coverage, as well as sampling error. It also captures errors that are due to the actions of respondents (e.g. inaccurate reporting, misunderstanding), the survey instrument or data processing (e.g. data entry)—see Biemer (2010) for details. All these issues could, in turn, impact estimates of socio-economic achievement gaps and how they differ across countries (Billiet & Matsuo, 2012).

This paper uses the TSE framework to highlight how such different sources of bias may impact estimates of socio-economic achievement gaps when using ILSAs such as PISA. In doing so, we focus on six factors: (i) school and student non-response; (ii) non-coverage of the target population; (iii) item non-response to information on socio-economic background; (iv) measurement error in the socio-economic background variable(s); (v) the derivation of PISA scores; and (vi) processing of the data to fit an international framework. Whereas previous research has investigated several of these aspects of ILSAs in isolation (e.g. Heine et al., 2017; Rutkowski, 2011; Rutkowski & Rutkowski, 2013), few studies have provided a comprehensive review of them all. This paper adds this to an existing literature via a case study of the PISA 2012 data from Germany, providing a comprehensive review of possible errors that may affect socio-economic group comparisons in international studies. We thus aim to foster greater awareness and understanding of potential sources of bias when using such data to investigate socio-economic inequality in educational achievement.

## TOTAL SURVEY ERROR AND SOCIO-ECONOMIC GROUP COMPARISONS

This paper is situated in the total survey error framework, which gauges the quality of sample survey data (Schnepf, 2018). Total survey error has been defined in different ways, and multiple typologies now exist. Our analysis is guided by the description of Groves and Lyburg (2010, p. 850), who note:

> 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, operationalise constructs that are the target of the measurement, obtain responses to the measurements, and summarise the data for estimating some stated population parameter.

This has led us to identify six issues that may affect estimates of socio-economic achievement gaps. We do not claim that this completely captures *all* aspects of total survey error in all ILSAs, but that these represent six key sources of bias that may impact the results.

### Coverage of the (target) population

Within any sample-based study the target population must be defined. Although this may seem trivial at first, it can have important implications for investigations of socio-economic inequalities. For instance, in some countries not all students at a given age are enrolled in school, potentially meaning under-representation of disadvantaged and lower achieving groups (Education Datalab, 2017). Moreover, it is often the case that some groups within the target population are excluded owing to pre-defined reasons such as accessibility or feasibility—most notably students with special educational needs (Jerrim, 2021; Anders et al., 2020). This then has the potential to impact socio-economic group comparisons. This becomes even more complex in cross-national comparative research, given that exclusion rates (including of SEN pupils) differ drastically between countries (LeRoy et al., 2019).

### Survey (unit) non-response

In large, cross-national studies of students, non-response can occur at the school level and the student level. If such non-response occurred completely at random, socio-economic group comparisons would not be biased. Yet this is unlikely to hold. For instance, Heine et al. (2017) showed that non-participation in PISA in Germany is related to family background and prior achievement at both the student- and school-levels. Thus, if disadvantaged pupils disproportionately drop out of the sample, this may lead to biased estimates of socio-economic comparisons.

### Item 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. This 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 and Cortés (2012) found that

students missing socio-economic status data differed systematically from those with complete data, potentially biasing socio-economic group comparisons.

## Constructing and operationalising a comparable group measure

To reliably compare socio-economic differences across countries, it is vital that the primary covariate of interest (family background) carries the same meaning in each nation. This is, however, a difficult task. Take, for example, the case of parental education—a key measure of family background. 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.

Although considerable time and effort have been invested into building international classifications (such as the International Standard Classification of Education (ISCED); Organisation for Economic Co-operation and Development, 1999) and developing measures for use in an international context (e.g. chapter 16 in OECD, 2014) this only partially resolves issues of cross-national comparability. National qualifications are often misclassified within the ISCED framework, as they do not easily fit into such international classification schema (Schneider & Kogan, 2008). This has been further complicated by extensive recoding of the ISCED schema across different versions. Previous research by Rutkowski and Rutkowski (2013) has also shown how there is poor cross-national comparability for other socio-economic background measures, such as the PISA Economic, Social and Cultural Status index.

## Measurement error in socio-economic background

Measures of socio-economic background face further challenges when participants act as proxy respondents. This is common in some ILSAs such as PISA, where students provide information about their parents. Misreporting of parental characteristics (e.g. education levels) hence becomes a problem which, 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 previous cross-national research illustrating how this may lead to underestimation of parental education achievement gaps across countries (Jerrim & Micklewright, 2014).

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

International large-scale assessmentss only have limited time available to test students. They thus employ a rotated test design, meaning that 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. In reality, this is essentially a form of multiple imputation. Specifically, students' answers to the administered test questions, along with their responses to the background questionnaire (including their socio-economic status), are used to estimate student achievement (Mislevy et al., 1992; von Davier et al., 2009). Importantly, this means that measures of socio-economic status (such as parental education) have a direct influence on the construction of ILSA test scores. Previous research has

shown how measurement error in background variables (such as socio-economic status) can introduce bias into this process (Rutkowski, 2014).

## DATA AND METHODS

### Data

We use PISA 2012 data from Germany to illustrate how the aforementioned issues can impact estimates of academic achievement between parental education groups. This setting was chosen owing to the rich information available for this country in the PISA 2012 cycle. Specifically, PISA 2012 was the last cycle to administer our measure of interest—highest parental education—in both student and parental questionnaires. Moreover, we have also accessed a country-specific version of the PISA 2012 data for Germany (Prenzel et al., 2015), which includes more fine-grained responses to the parental education questions than have been published in the publicly available PISA dataset (OECD, n.d.).[i] This allows us to investigate whether the coding of parental education into the ISCED framework impacted estimates of socio-economic disparities in student achievement in Germany.

### Sampling design in PISA 2012

The target population in PISA is 15-year-olds enrolled in at least grade 7. Consequently, some 15-year-olds—such as those who are home-schooled, have been permanently excluded or have repeated many grades—are not covered. Furthermore, countries are allowed to exclude up to 5% of students/schools from their sample owing to: (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 test is not available. Countries are also allowed to specify a fifth national specific criteria for excluding certain schools or students from the sample, including remoteness of geographical regions, certain language groups owing to political reasons and students with certain special educational needs (OECD, 2014, chapter 4).

Once the sampling criteria are defined, students are selected into PISA using a two-stage procedure. Schools are first selected with probability proportional to size and then students are randomly sampled within schools. Two replacement schools are also drawn for each sampled school in case they refuse to participate (this is essentially a form of imputation). See OECD, 2014, chapter 4) and Prenzel et al. (2013, chapter 10) for further details.

### Test design

In ILSAs, students only get asked a fraction of all of the test questions. 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 were 13 item clusters. As mathematics was the focus of PISA 2012, seven of the 13 item clusters addressed mathematics with only three each about science and reading. While each booklet contained at least one cluster of mathematics items, most booklets only included two of three domains. Consequently only 40% of the students answered questions in all three domains. The data for each student is thus mostly missing by design (OECD, 2014, pp. 30, 31).

## Measurement of socio-economic status

While socio-economic background has multiple facets, within this paper we focus on parental education. This is because parental education is commonly used in secondary analyses of ILSA (e.g. Martins & Veiga, 2010; Yang-Hansen & Gustafsson, 2016)—as well as the social stratification literature more generally. The data we have available is also particularly well suited to studying this socio-economic indicator.

The publicly available PISA 2012 data contains students' answers about their mother's and father's education. Using this information, the OECD has derived maternal and paternal ISCED levels. Highest parental education is computed by combining responses across mothers and fathers. In contrast to the student questionnaire, the parent questionnaire only asks about education levels equivalent to ISCED level 3A and above. Moreover, the parent education categories included in the parent and student questionnaires differ, meaning they cannot be directly compared using the publicly available PISA data.

However, more fine-grained information about parental education is available in the German national version of the data (available upon request from Prenzel et al., 2015). This has the major advantage that the same questions and response options were posed to students and their parents. Yet the pre-computed highest parental education variable (in ISCED levels) in the German data does not make use of the more detailed information available. Rather, it classifies highest parental education into the same (fewer and broader) categories as the international version. The mapping of the German parental education questions into the ISCED levels used in PISA is outlined in the PISA technical report for Germany (Mang et al., 2018, pp. 173–176).

## Method of plausible value computation

As noted above, not all students answer questions in all three core PISA domains. Yet every student is assigned a mathematics, reading and science score via a complex statistical procedure. Specifically, background variables (including parental education) are used in combination with the cognitive items to derive student achievement distributions (OECD, 2014). Five 'plausible values' are then computed following five steps:

- First, item difficulties are determined using an item response theory model based on a common sample (e.g. including subsample of almost all countries in PISA 2012).
- Second, information from the background questionnaires is reduced to a smaller set of variables via a principal components analysis.
- Third, student achievement distributions are estimated using a 'latent regression' model, combining students' responses to the test questions and their background characteristics.
- Fourth, 'plausible values' are randomly drawn from each student's achievement distribution. These are 'imputations' for unobserved (latent) student achievement (Mislevy, 1991).
- Finally, the plausible values are transformed onto the common PISA scale.

In the following section we analyse the impact that including highest parental education in the plausible value computation has upon estimates of socio-economic achievement gaps (see section 4.6). As a result, we focus on one specific aspect of the aforementioned process. Specifically, we conduct three different versions of the third step:

- 0. No background variables are included in the latent regression.
- 1. Highest parental education—based on student responses—is included in the latent regression.

- 2. Highest parental education—based on parental responses—is included in the latent regression.

Estimates of the socio-economic achievement gap using these three different approaches are then compared.

## ANALYSIS

All analyses are conducted using the 'intsvy' package (Caro & Biecek, 2017) within 'R'. This fully accounts for the complex PISA sample and test design.

### Coverage of the (target) population

The German target population in PISA 2012 consists of 798,136 fifteen-year-olds. However, owing to a combination of school exclusions, students being withdrawn from the sample or being deemed ineligible, the non-coverage rate of this target population was around 2.6%. Although we do not have direct evidence on the characteristics of these students (owing to a lack of data) special educational needs (SEN) is one of the key reasons why such exclusions are made. Given that socio-economically disadvantaged students are more likely to have SEN, it seems likely that those who have been excluded will be disproportionately from disadvantaged backgrounds.

Nevertheless the non-coverage rate for Germany is smaller than that for other countries. Indeed, in eight countries (Canada, Denmark, Estonia, Luxemburg, Norway, Sweden, UK and the USA) it exceeds the maximum 5% limit specified within the PISA technical standards (yet has still managed to pass to OECD's data adjudication process). This suggests that, while non-coverage is unlikely to be a major concern in Germany, this may not hold true elsewhere. Indeed, there is a risk that countries with high levels of exclusions will underrepresent the proportion of students from disadvantaged backgrounds. However, without further data, it is difficult to know whether this is likely to lead to an upward or downward bias in estimates of socio-economic achievement gaps.

### Survey non-response

In total, 99% of the 233 sampled schools participated in Germany (after replacements were included).[ii] All three non-participating schools were private schools, which are disproportionately attended by children from advantaged socio-economic backgrounds (Lohmann et al., 2009). The final student response rate was 93% (OECD, 2014, p. 185).

Such response rates are comparatively high by international standards. School response rates vary internationally between 78% in the USA and 100% in several countries, while student response rates vary from 81% in Canada to over 99% in Viet Nam. The risk of non-response bias affecting socio-economic achievement gaps is hence lower in Germany (owing to higher overall response rates) than in many other Western countries.

### Non-response to the background questionnaires

The aforementioned response rates refer to students taking the PISA test. Yet there could be additional non-response owing to non-completion of the PISA background questionnaires.

Of the 16 German states, five made the whole student questionnaire mandatory, one made only parts of it mandatory, while in 10 states the questionnaire was voluntary (and required parental approval). Response rates differed between states as a result (Prenzel et al., 2013: 33). Overall, 14% of the German student sample did not complete the background questionnaire.[iii] When combined with the overall student response rate (93%) reported in the section above, almost a fifth (19%) of the originally sampled students did not complete both the background questionnaire and the PISA test. The response rate to the parental questionnaire in Germany was much lower (58%).

Figure 1 compares the distribution of highest parental education in PISA 2012 with the distribution of highest household education in the German socio-economic panel (Socio-Economic Panel, 2019)—having made the two datasets as comparable as possible.[iv] There are clearly large differences in the distributions. This is most prominent for category 3 (ISCED 3B/C) which is mainly consists of apprenticeships in Germany. In PISA, only 12% of the students and 11% of the parents indicated that ISCED 3B/C was the highest parental education level. In contrast, the most comparable category in the socio-economic panel forms 37% of the sample, which is broadly consistent with figures from the 2012 German statistical yearbook (Statistisches Bundesamt, 2012, chapter 3). Together, this provides a clear

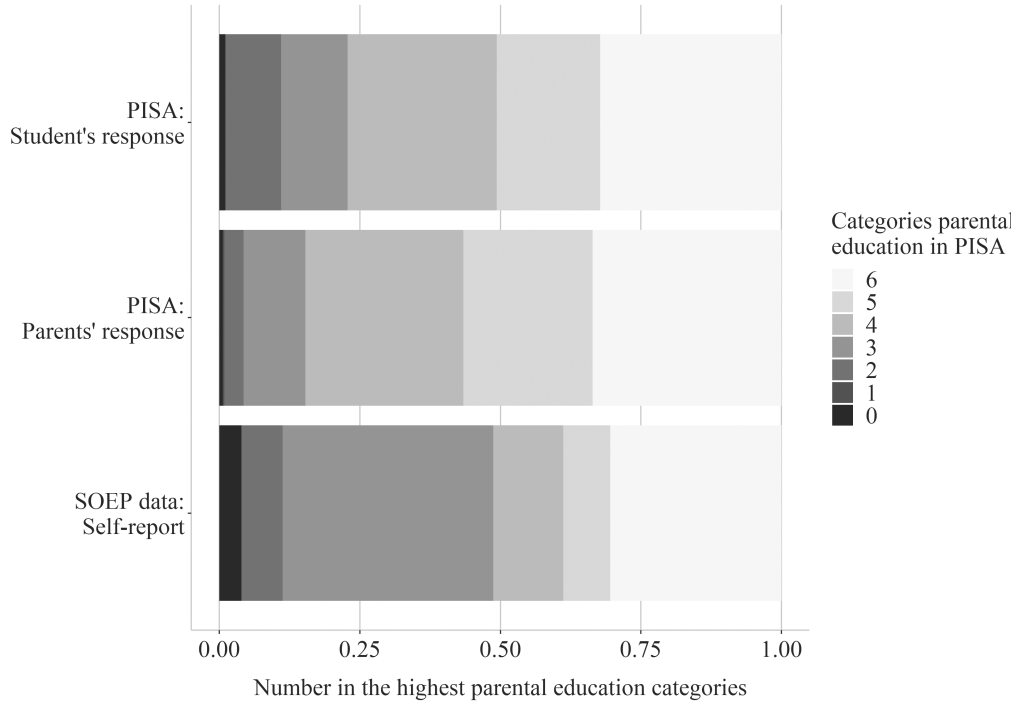


**FIGURE 1** The distribution of highest parental education in Germany based on (i) students' answers in the PISA data set, (ii) parents' answers in the PISA data set and (iii) household data in the German socio-economic panel dataset. Categories: 0 = below ISCED 1; 1 = ISCED 1; 2 = ISCED 2; 3 = ISCED 3C, 3B; 4 = ISCED 3A, 4; 5 = ISCED 5B; 6 = ISCED 5A, 6. Sample size students = 3936. Sample size parents = 2832. The student sample is based on 3936 students from the German PISA 2012 dataset (with the parental education variable coded as described in Section 4.3). 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 Section 4.3). The SOEP data sample is displaying highest education in 19,629 households, which were selected according to the following criteria: (i) person was not interviewed as child or youth of the household; (ii) person was born between 1952 and 1982; and (iii) 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.

indication that the PISA data for Germany may underrepresent the proportion of German students from ISCED 3B/C parental education backgrounds.

## Constructing and coding of parental education

To compare parental education achievement gaps across countries, an internationally comparable measure is needed. In PISA 2012, the ISCED97 schema was used. Specifically, respondents were asked about their German qualifications, which were then recoded into this scale, following the procedures outlined in Mang et al. (2018, pp. 173–178).

Processing and coding errors in such situations are known as procedural errors in the total survey error framework. We believe that two such procedural errors were made when parental education was converted into ISCED levels in Germany.[v] First, according to Mang et al. (2018) the education level of '*Abgeschlossene Lehre—Abschluss an einer Handelsschule oder ein vergleichbarer* Abschluss im Ausland' (Completed apprenticeship—from a commercial school or a comparable qualification abroad) was planned to be coded as ISCED level 3B/C in the PISA database. However, this category was actually not incorporated into the ISCED parental education schema in PISA; those holding this qualification seem to have been ignored. Second, the education level '*Abschluss an einer Fachoberschule / Berufsschule/Berufsf*' (Graduated from a technical college/vocational school/vocational school) was—according to Mang et al. (2018)—supposed to be included within ISCED level 3A/4. However, in the international PISA dataset, it appears that those holding this qualification were placed into ISCED level 3B/C instead. Importantly, these suspected errors are present in the PISA 2012 public use PISA data files for Germany, downloadable from the OECD website.

The distribution of parental education in the German PISA dataset is hence rather different in places than it would have been had these suspected procedural errors not been made. Specifically, the proportion of parents recorded as holding an ISCED level 3B/C qualification would increase substantially, while the proportion recorded as holding ISCED level 2 would roughly halve.

Figure 2 presents average PISA mathematics scores for each parental education group before and after we have corrected for these suspected procedural errors. The greatest difference can be observed for ISCED level 2 (category) where, after correcting for the suspected procedural errors, average mathematics scores are 29 points lower. In contrast, average PISA mathematics scores for the ISCED 3B/C group (category 3) are 12 points higher. Moreover, the estimated parental education achievement gap—as measured by the difference in mean scores between the ISCED level 5 or 6 and ISCED level 2 and below groups—is around 30 points higher once the suspected procedural errors have been corrected (a 92-point difference rather than a 62-point difference). Putting this into comparative perspective, the parental education gap in Germany moves from well below the OECD average (77 points) to substantially above it once this suspected error has been recoded.

## Item non-response

For around a fifth of students, no data on highest parental education is available in Germany from students—amongst the highest out of any country (the OECD average is just 4%). This is due to a combination of (i) participants not returning the questionnaire and (ii) participants skipping the parental education question. Figure 3 illustrates how this differs between parents and students. Far fewer parental background questionnaires were returned, but almost all that were (98%) provided a response to the parental education question. For students,

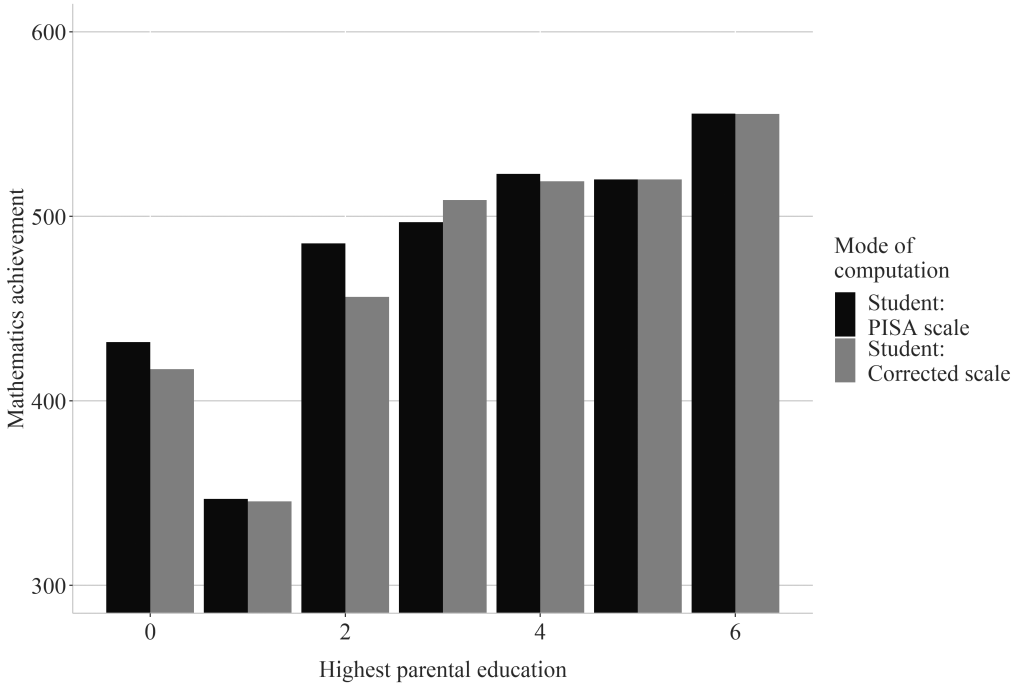


**FIGURE 2** Average PISA mathematics scores by parental education group before and after correcting for suspected procedural errors. Categories: 0 = below ISCED 1; 1 = ISCED 1; 2 = ISCED 2; 3 = ISCED 3C, 3B; 4 = ISCED 3A, 4; 5 = ISCED 5B; 6 = ISCED 5A, 6. Group means were computed using plausible values and accounting for sample weights.

there is more information available overall (highest parental education is observed for 79% of the sample), but where information on parental education is missing, this is more likely to be due to students skipping the question about parental education.

It is also the case that students/parents who do not respond to the questions about parental education tend to be academically weaker. For instance, the average PISA mathematics score of those students who reported information about parental education was 522. This is around 40–50 points higher (equivalent to an effect size of 0.4–0.5) than their peers who did not provide such information—owing to them either not completing the specific question asking about parents' education or not completing the background questionnaire at all.[vi] Similar patterns emerge with respect to the parental questionnaire as well. This is a strong indicator that missing parental education data does not occur at random. Unfortunately, without further information, it is not possible to establish whether this issue leads to an upward or downward bias in estimated parental education achievement gaps.

## Measurement error: Agreement of parents and students

As students and parents both answered questions about parental education, it is possible to establish the level of agreement in their responses. Table 1 hence presents a cross-tabulation of parent and student reports. Out of the 2658 cases where information is available from both parties, around half selected the same category. Approximately a fifth of students reported a higher parental education level than their parents, with around a third reporting it to be lower. 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—indicating

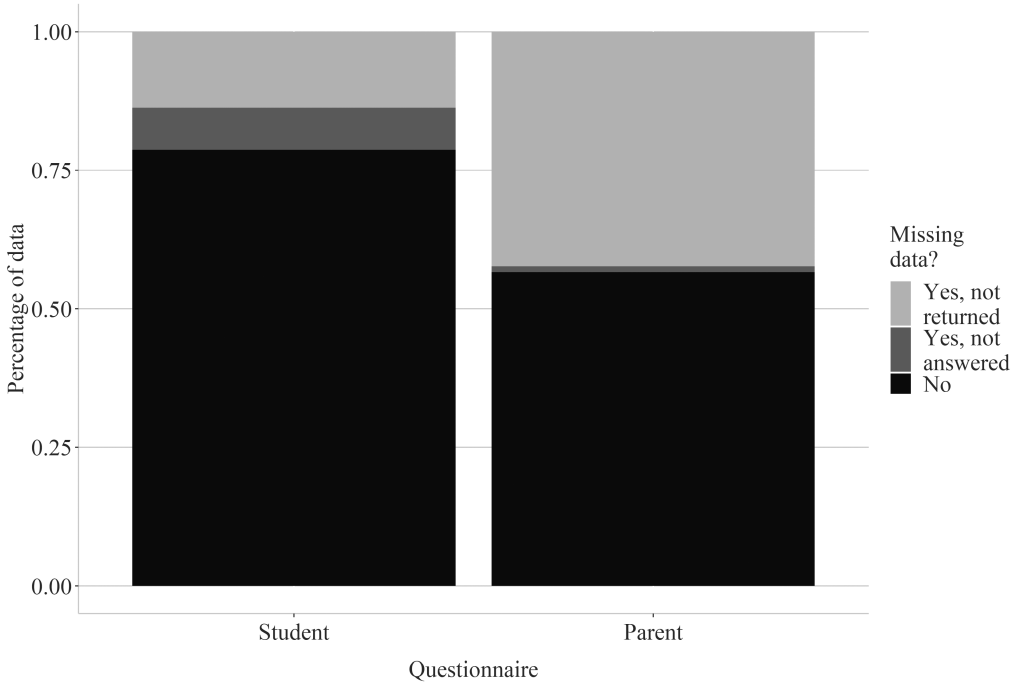


**FIGURE 3** Missing highest parental education data in the student and parent background questionnaire. The student sample is 5001 with 381 items of non-response and 684 survey non-responses. The parental sample is 5001 with 53 items of non-response and 2116 survey non-responses. 'Yes, not returned' refers to where the student/parent questionnaire was not completed (survey non-response). 'Yes, not answered' is where the student/parent questionnaire was completed, but the questions about parental education were skipped (item non-response). 'No' refers to where parental education data is available.

**TABLE 1** Cross-tabulation of student and parent reports of highest parental education.

| | | Student response | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** | **Missing** |
| 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 | |
| | Missing | 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.

Grey shading refers to the diagonal where parent and student responses agree.

'moderate' levels of agreement. The consistency of parent and student reports of parental education levels is thus modest, at best.

After restricting the sample to participants with valid information on highest parental education from both students and parents, the mathematics achievement gap is estimated to be

89 points using students' responses. However, the analogous figure using parental reports is much larger—116 points. This is a substantial difference—approximately a quarter of a standard deviation. Assuming that parents report their own education level more accurately than their children, this suggests that the gap in academic achievement between parental education groups is severely underestimated when using students reports. Note that, when reporting on socio-economic achievement gaps across countries, the OECD use information reported by students rather than by their parents. Student reports are also the primary source of information about socio-economic background that secondary analysts of PISA draw upon as well (Jerrim & Micklewright, 2014).

## Plausible value computation

As noted in sections 2.6 and 3.5, the derivation of PISA test scores ('plausible values') uses information from the student background questionnaire (including parental education) via a latent regression model. Here, we investigate how including or excluding different information on highest parental education from this model impacts estimates of parental education gaps in achievement. Specifically, Table 2 illustrates how these vary when three different latent regression model specifications are used to generate PISA scores:

0. No conditioning.
1. Student reports of highest parental education as the only conditioning variable.
2. Parent reports of highest parental education as the only conditioning variable.

## Estimates are reported in terms of effect sizes

The parental education achievement gap is smallest in specification 0, when no conditioning is applied. This is as expected, as it is well established in the psychometric literature that group differences will be attenuated if the variable of interest (highest parental education) is not included in the latent regression model (Mislevy, 1991; Wu, 2005). Yet there are also differences between specifications (1) and (2), suggesting that the results are also impacted by *whose* reports of parental education are used in the latent regression model (students or parents). For instance, focusing on the figures in the right-hand column, the estimated achievement gap increases from 1.17 to 1.28 standard deviations, depending on whether parents or student reports are used in the latent regression. In other words, using

**TABLE 2** The estimated parental education achievement gap using different conditioning models and grouping variables.

| | Grouping variable | |
|---|---|---|
| **Conditioning model** | **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 done based on either the students' response or parents' response.

student rather than parental reports of mothers and fathers' education level in the latent regression model may lead to a downward bias in estimated parental education achievement gaps.

## CONCLUSIONS AND RECOMMENDATIONS

International large-scale assessments such as PISA have become an important part in national and international education debates. It is therefore of great importance that the data underlying ILSAs has a solid foundation. Yet, this is not always guaranteed. Bias can creep into the data at different points in the collection, coding and analysis process, from sampling through to the derivation of the PISA scores. 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. Most previous studies hence fail to provide an overarching picture of the various issues that could impact the results. This paper has started to fill this gap in the literature. Applying the total survey error framework to the PISA 2012 data from Germany, we document different ways that bias may be introduced into estimates of socio-economic status achievement gaps.

Six different issues have been investigated, encompassing coverage rates, survey non-response, procedural errors, item non-response, measurement error and the latent regression underlying the generation of PISA scores. Our analyses show that some—although not all—of the above can impact measurement of socio-economic achievement gaps. For instance, a suspected procedural error in converting information gathered on maternal and paternal education into ISCED levels was found to reduce parental education achievement gaps in Germany by around 30 PISA test points As a result, many parents were falsely classified as holding only a low qualification level, when really their education level was higher. The achievement gap increased from 62 to 92 PISA points (around 0.3 standard deviations). On the other hand, we find that using student reports of mothers' and fathers' education (rather than parental reports) leads to an underestimation of the parental education achievement gap in Germany by around 25 test points (0.25 standard deviations). Hence, our overall conclusion is that estimates of socio-economic achievement gaps using the PISA data for Germany do not seem to be particularly robust. Although it is not possible to combine all sources of potential bias to empirically estimate their joint impact, we believe it to be likely that ignoring them will probably lead to underestimation of differences in achievement between parental education groups in Germany.

Two key recommendations follow. First, better communication of such issues is needed. The limitations about the background data collected in PISA should be more thoroughly investigated and caveats discussed. Second, we believe that many researchers are likely to stumble across such issues during the course of their study. As a result, they either decide to use another variable or briefly note it somewhere in their paper. Yet it would be of great communal benefit if this knowledge was shared with other users. One possible way this could be done is via a log or database on the PISA homepage. 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 paper aims to show different sources of bias and is set in the framework of total survey error, we do not claim to look at all potential sources of bias. This study serves as a comprehensive review and wants to shed light on some infrequently discussed statistical properties that can potentially influence estimates of socio-economic achievement gaps. We have also focused upon a single national case study to facilitate in-depth investigation. It is likely that the influence of the various different aspects investigated will differ in other

national settings. Nevertheless, we believe that this paper highlights the importance of conducting such investigations and how they can be systematically approached.

## FUNDING INFORMATION

The authors are part of the European Training Network OCCAM. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 765400. For more information, see https://etn-occam.eu.

## CONFLICT OF INTEREST STATEMENT

Nothing to declare.

## DATA AVAILABILITY STATEMENT

The German PISA data we use are available by application from 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.

## ETHICS STATEMENT

The research was conducted under BERA ethical guidelines.

## ENDNOTES

[i] The parental education questions in Germany included more options than the publicly available PISA 2012 dataset suggests, with Germany's national qualifications condensed to fit the ISCED framework.

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

[iii] The German report states a return rate of 82%, yet the official and publicly downloadable data contain information for more cases.

[iv] Highest education was calculated for the adults, born between 1952 and 1982, of each household where an adult had 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.

[v] These are the conclusions we have reached from comparing parental education variables in the German PISA dataset with the recoded information on parental education included in the international dataset. Although we cannot verify this for certain, we believe this to be the most likely explanation.

[vi] The average PISA mathematics score of students who did not complete the student questionnaire (i.e. survey non-response) was 486. Those who completed the survey, but skipped the questions asking about parental education, achieved an average score of 473.

## REFERENCES

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 Asseessment Evaluation and Accountability*, *33*(1), 229–249.

Biemer, P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

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. https://doi.org/10.1007/978-1-4614-3876-2_10

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*, *71*, 1–31. 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.

Caro, D. H., & Biecek, P. (2017). intsvy: An R Package for analyzing international large-scale assessment data. *Journal of Statistical Software*, *81*(7), 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.

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/

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

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

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

Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK? *Review of Education*, *9*(3), e3270.

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

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.

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. https://doi.org/10.1037/pspp0000314

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

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

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

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

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

OECD. (2019). *PISA 2018 results (volume II) where all studentes can succeed*. OECD.

OECD. (n.d.). *Programme for international student assessment 2012 (PISA 2012) [Data set]*. https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

Organisation for Economic Co-operation and Development. (1999). *Classifying educational programmes: Manual for ISCED-97 implementation in OECD countries*. Paris: OECD.

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

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. https://www.pisa.tum.de/fileadmin/w00bgi/www/Berichtsbaende_und_Zusammenfassungungen/PISA_2012_EBook_ISBN3001.pdf

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]*. IQB—Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_PISA_2012_v5

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

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

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* (p. 5). JRC Working Papers in Economics and Finance, 2018/5. https://doi.org/10.2760/219007

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

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

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.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.

Yang-Hansen, K., & Gustafsson, J.-E. (2016). Determinants of country differences in effects of parental education on children's academic achievement. *Large-Scale Assessments in Education*, *4*(1), 11.

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.

**How to cite this article:** Jerrim, J. & Zieger, L. (2023). How robust are socio-economic achievement gradients using PISA data? A case study from Germany. *British Educational Research Journal*, *00*, 1–16. https://doi.org/10.1002/berj.3934