European Commission

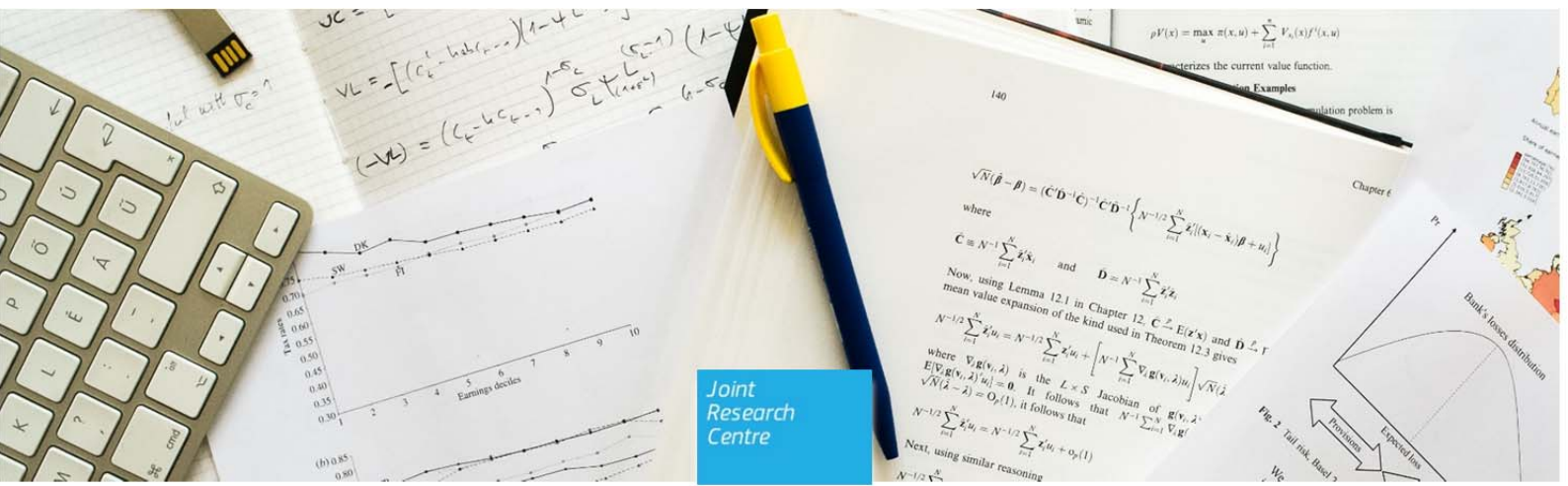
JRC TECHNICAL REPORTS

Insights into survey errors of large scale educational achievement surveys

Schnepf, Sylke V.

2018

JRC Working Papers in Economics and Finance, 2018/5



Contact information
Name: Schnepf, Sylke V.
Email: sylke.schnepf@ec.europa.eu

JRC Science Hub
https://ec.europa.eu/jrc

JRC111734

PDF ISBN 978-92-79-85795-9 ISSN 2467-2203 doi:10.2760/219007

Luxembourg: Publications Office of the European Union, 2018

How to cite this report: Schnepf, S.V., *Insights into survey errors of large scale educational achievement surveys*, JRC Working Papers in Economics and Finance, 2018/5, doi:10.2760/219007

INSIGHTS INTO SURVEY ERRORS OF LARGE SCALE EDUCATIONAL ACHIEVEMENT SURVEYS

Sylke V. Schnepf

European Commission, Joint Research Centre, Ispra, Italy

August 2018

Abstract

While educational achievement surveys revolutionised research on education cross-nationally, the surveys have been repeatedly subject of heated debate since first results were published. This paper reviews existing research examining the design and methodology of educational achievement surveys. Results are reported by allocating them to the specific survey error component of achievement estimates they address. Different error components from the design, collection, processing and analysis of survey data constitute the total survey error, which is an error difficult to quantify but important for assessing the overall accuracy of the surveys' achievement estimates. The review shows that there are many reasons to assume that the total survey error associated with countries' educational achievement estimates is likely to be inflated by other errors besides the standard error reported by survey organisers. Given the policy relevance of the surveys' estimates, policy makers and the research community would greatly benefit from survey organisers providing more transparency on the different potential errors of educational achievement estimates. Without this information the debate about the fitness of educational achievement data for policy making is unlikely to dissolve.

JEL Codes: I20, I21, C83

Keywords: educational achievement surveys, survey methodology, survey errors, PISA, TIMSS, PIRLS.

Disclaimer: The views expressed are purely those of the writer and may not under any circumstances be regarded as stating an official position of the European Commission.

Acknowledgements

The discussion in this paper builds on the author's joint research with her colleagues over the past 15 years, especially John Micklewright, Luisa Araujo, Giorgina Brown, Gabriele Durrant, Pedro N. Silva, Andrea Saltelli and Chris Skinner.

1 Introduction

The 'Programme for International Student Assessment' (PISA) and other educational achievement surveys revolutionised research on education cross-nationally. PISA, the most prominent survey, was launched in 2000 and focuses on educational achievement of 15 year-olds. It is run every three years in a large and growing number of countries (72 countries in 2015) by the OECD. Other surveys comprise the 'Trends in International Mathematics and Science Study' (TIMSS) focusing on 4th and 8th graders and the 'Progress in International Reading Literacy Study' (PIRLS) looking at primary school children only. The typical design of educational achievement surveys involves collecting a representative sample of schools at a first stage and then pupils within schools at a second stage.

All of these achievement surveys sample students, measure their educational achievement with a battery of questions aiming to capture school curriculum (TIMSS) or life skills (PISA and PIRLS) and collect additional information on the students covering their socio-economic background and attitudes, but also depending on year and survey in-depth information on their school, their teachers and even sometimes their parents. These cross-national data have enriched educational research since they provide opportunities to investigate educational achievement in an unprecedented way cross-nationally. Most importantly, PISA results have become highly influential for policy formulation impacting on education policy design in many European countries (Schnepf and Volante, 2017).

At the same time, educational achievement surveys have been repeatedly the subject of heated debate since first results were published (i.e. Prais 2003, Brown at al 2007, Hopmann et al, 2009; Kreiner and Christensen 2014, Fernandez-Cano 2016, Goldstein 2017, Wiseman and Waluyo 2017). The debate was not restricted to academics but also covered in the media. In 2014 an open letter (Meyer and Zahedi, 2014) to *The Guardian* suggested skipping the 2015 round of PISA due to grave concern about its deficiencies. The letter was jointly signed by approximately 80 academics, public school district administrators, parents and teachers and initiated correspondence with the OECD.

The current criticism of educational achievement surveys focuses on many different potential problems and is unstructured. This paper reviews existing research examining the design and methodology of educational achievement surveys. Using the view of a survey methodologist, existing research results are structured by allocating them to the specific survey error of different stages of the survey estimate production: the design, collection, processing, and analysis of survey data. Thereby, survey error refers to the deviation of a survey response from its underlying true value. The deviation can be due to the variance or bias component any survey error entails.

Different error components constitute the total survey error, which is an error difficult to quantify but important to consider for judging on the overall accuracy of the achievement estimate obtained from educational achievement surveys. While the discussion below is generally framed around the most policy relevant survey PISA, the discussion is mostly applicable to educational achievement surveys in general due to similar survey methodologies used.

2 Possible survey errors of cross-national educational achievement surveys

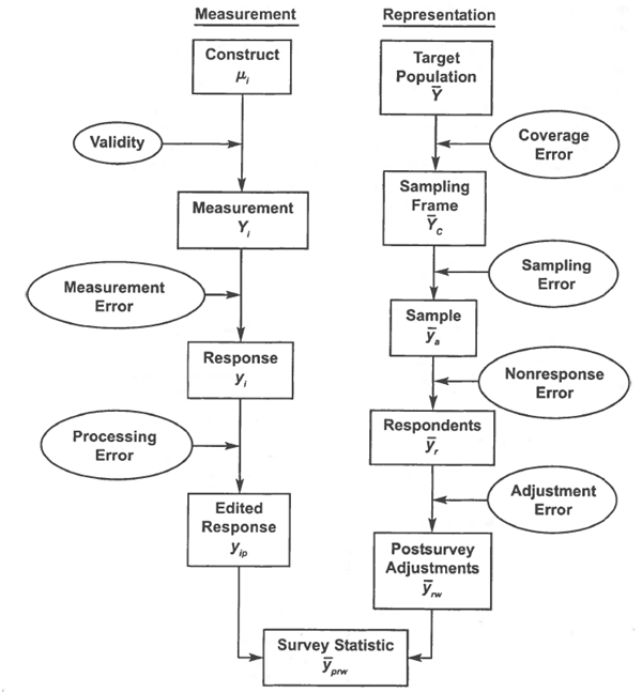
Survey methodologists tell us, that the quality of any survey depends on two aspects: the achieved quality of the measurement of the construct it aims to capture and the achieved representativeness of the sample for the population the survey aims to describe (Groves et al. 2009). Measurement and representativeness have both several quality components to meet as illustrated in Figure 1 taken from Groves et al (2009, p.48). On the measurement side, the extent to which the measure fails to reflect the underlying construct (lack of validity), whether the measure departs from the 'true' value (measurement error) and possible processing mistakes of responses can contribute to the total survey error. Regarding representativeness, coverage error (an exclusion of individuals from the target population that should be included and/or an inclusion of individuals not covered in the target population), sampling error, non-response error and any kind of adjustment errors need to be taken into account.

It is common practice, that regardless of these different potential sources for survey errors to arise, any kind of analysis reports only the variance component of the sampling error. In contrast to other errors, the sampling error is relatively easy to calculate. Following this practice, also educational achievement organisers provide only standard errors for educational achievement estimates. As a consequence, all other error components are neither considered nor discussed once results are interpreted or used for policy design. Possible bias of these errors is assumed to be equal to zero. Survey methodological issues that relate to these errors are generally shifted to long technical reports. This paper provides insights on possible bias due to other errors besides sampling error reviewing existing literature on methodological issues of educational achievement surveys.

The following section 2.1 discusses existing literature on potential methodological caveats of educational achievement surveys that relate to the measurement side of the survey estimate (validity, measurement and processing error). Section 2.2 investigates possible survey errors deriving from survey data not being representative for the population the survey aims to investigate. Given the complex survey design of educational achievement surveys, the analysis of these surveys requires specific caution. Further errors of

estimates can therefore derive during the analysis stage. Three examples of analysis traps are presented in Section 2.3. Section 2.4 discusses the total survey error which comprises all the different error components considered before. Section 3 concludes.

Figure 1: Stages at which survey errors can arise



Source: Groves, R., Floyd, J., Couper, P., Lepkowski, J., Singer, E. & Tourangeau, R. (2009) *Survey Methodology* (Hoboken, Wiley), p. 48.

2.1 Potential survey errors of educational achievement estimates deriving from measuring performance

Validity

Validity refers to the extent to which a measure reflects an underlying construct. PISA aims to assess the construct 'how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today's knowledge societies' cross-nationally (OECD, 2004, p. 12). Below, the problem of validity is discussed from two different angles: a) whether the ambitious PISA construct can be measured across countries which vary greatly in their culture and economic development and b) whether the design of the survey could wrongly lead to capturing other students' characteristics besides skills.

Can PISA measure cross-national skills?

The PISA construct is based on the assumption, that life skills needed to function in knowledge societies are the same for all countries covered in the survey. In 2015

countries like Singapore, Germany, the United States, Peru and Trinidad and Tobago all appear together in one league table. Obviously, countries participating in PISA differ greatly in terms of their cultures and level of economic development. The assumption on equal life skills needed in these societies is therefore rather uncertain and raises legitimate questions about how appropriate it is to rank these countries in a single table (Araujo et al 2017).

This links to concerns that country comparability in PISA is only achieved by ignoring the great diversity across the participating countries. Until PISA 2012, the modelling applied for constructing achievement scores assumed that each question (or item) had a specific 'difficulty' to be answered and that this 'difficulty' was exactly the same across countries (see discussion on Rasch models below). Items that did not fit this modelling assumption and therefore showed 'poor psychometric characteristics in more than ten countries ('dodgy' items)' (OECD, 2012, p. 148) could be deleted. Generally this was only a small number of items (OECD 2016a, Annex A5). However, these items could be regarded as those reflecting cultural bias (Goldstein, 2017). Their removal from the set therefore could be seen to obscure differences between countries that might otherwise demonstrate greater heterogeneity on varying educational dimensions.

Probably as a response to this criticism, the new PISA 2015 design (discussed in OECD 2016a, Annex A5) uses models that allows the difficulty of items to vary between countries to a limited extent. Jerrim et al (2018) show that the choice of modelling with country fixed 'difficulty' parameters or with more flexible 'difficulty' parameters does not change the results of educational achievement score estimates at the country level once the focus is on OECD countries and one single year only. While this is an encouraging result from the research community using OECD data, it would be very interesting to see this exercise repeated for all countries, including about half of the PISA countries that are less affluent than OECD countries. These are the ones where the modelling is likely to be most problematic. OECD (2016a, Annex A5) is very limited in the discussion on the possible results of the recent change of its model.

Furthermore, if we acknowledge that education is a multidimensional phenomenon as argued by Goldstein (2017) and other education specialists (Hopman et al 2007) it would be important to have a clear definition of this single dimension captured with the PISA educational achievement score. The construct remains rather vague.

Does PISA only measure skills?

If we agree with survey organisers and users of educational achievement surveys, that skills can be measured and compared across countries that differ greatly in many aspects, the question arises as to what extent the PISA test design measures the construct of achievement. Meyerhoefer (2007) reports that items used in PISA do not

only measure educational achievement but also a student's ability to comply with the test structure.

A similar problem could derive from a change of the assessment mode of PISA in 2015. Pre-2015 children sat paper tests. In 2015 PISA organisers introduced test delivery with computers. Could the introduction of computer tests possibly lead to a measure that captures more computer skills than the construct of skills aimed to be measured? The OECD (2016a) concludes that mode effects are negligible so that results can be compared across paper- and computer-based modes.

Measurement error

Measurement error appears if the measure does not fit the 'true' value it aims to capture. It is tricky to avoid measurement error in a cross-national survey. The first challenge is to create items that are culturally neutral; the second is to translate these items into other languages. In order to achieve the desired neutrality, the OECD uses a variety of mechanisms to make sure that wording and translation do not impact on the results. Moreover, the OECD generally runs trials before implementing the final PISA questionnaire. Currently, not much is known about this process.

Test questions have not often been scrutinised by the research community, so that not much is known about possible measurement errors. Sjoberg (2007) states pupils might not give their best performance in answering especially long survey items given that they have no incentives to do so. Pupils' willingness to comply can also differ across countries which potentially could lead to bias of achievement measures impacting on cross-national differences found.

More openness and transparency on the part of the OECD about the results of trials and the consequent choice of items would help potential users of the results to judge their reliability.

Processing error: item response models

Data processing errors derive from flawed editing, data entry and coding. This happens after the data has been collected and when it is transformed into a data set used for the analysis.

Once students have answered a battery of achievement questions, these answers need to be summarised in an estimate of a person's 'proficiency' for each subject (math, reading, science, etc.) measured in the survey. This is generally done by using item response (IR) models. The achievement scores are therefore derived data and very different to a measure of i.e. percent of right answers.

Recently, Jacob and Rothstein (2016) and Jacob (2016) raise doubts about the use of item response models for evaluating and comparing educational achievement between groups. '*The scores that the models produce are generally not unbiased measures of student ability, and may not be suitable for many secondary analyses that economists would like to perform*' (Jacob and Rothstein 2016, p. 86). In their article they cover a number of problems of item response models not discussed here.

Independent of possible fundamental problems of item response models, the question arises as to whether the choices made over the method of derivation have an appreciable impact on the surveys' results. Are results on league table rankings and variation within countries sensitive to the choice of item response models?

In order to discuss this question the following introduces into the basic idea of item response models and discusses which models are currently in use.

Educational achievement surveys use generally uni-dimensional IR models. For PISA 2015 the item response model was changed compared to its previous rounds. This was due to 'concerns over the insufficiencies of the Rasch model' (OECD 2016a, p. 142, Kreiner and Christensen 2014) which was previously used. In detail, in prior PISA cycles (2000 to 2012) the so called Rasch model or one parameter item response model was applied for dichotomous outcome items (for more response categories a partial credit model was used, which is an extension of the Rasch model). The one parameter model allows for differences in the degree of difficulty of each question (αi) (which as discussed above was assumed to be the same for all countries in previous rounds). It measures students' proficiency (θ) in the following way, whereby i refers to the question and j to the student.

One parameter model: $p_{ij}(\text{correct answer}) = 1/[1+\exp(-(\theta_j - \alpha_i))]$

The same model was used in TIMSS but only for the 1995 round.

Since 2015, however PISA data is based on the two parameter model for dichotomously scored responses (and the generalised partial credit model for other items (OECD 2016a)). The added parameter β refers to the power of a question to discriminate between individuals with high and low ability.

Two parameter model: $pij(\text{correct answer}) = 1/[1+\exp(-\beta i(\theta j - \alpha i))]$

How do other achievement surveys scale the many responses to different items? PIRLS since its beginning in 2001 and TIMSS since 1999 use also the two-parameter model but only for items not deriving from multiple choice but with just two response options and a partial credit model for response items with more than two response options. For other dichotomous outcomes deriving from multiple-choice items, a three

parameter model is used. Compared to the two-parameter model it allows in addition for the probability that the answer is simply guessed. Formally, the models give the probability of a correct answer to question i by student j as:

Three parameter model: $p_{ij}(\text{correct answer}) = \gamma_i + (1 - \gamma_i) / [1+\exp(-\beta_i(\theta_j - \alpha_i))]$

where, γ is the probability that the answer to a question is guessed.

The use of varying item response models by different survey organisers leads to the question of how different models impact on the results found.

The most recent PISA report (OECD 2016a, Annex A5) stays relatively vague. OECD organisers computed country means of previous PISA rounds using the new 2015 scaling approach. Then correlations of country means under alternative scaling approaches were calculated. The report concludes: '*The high correlations reported in this table for the years 2006, 2009 and 2012 (all higher than 0.993, with the exception of reading in 2006, for which the correlation is 0.985) indicate that the relative positon of countries on the PISA scale is hardly affected by the changes introduced in 2015 in the scaling approach*'. While this information is very valuable and reassuring, checking the robustness of models for country means only is very limited, since it is the periphery of the achievement distribution that is most effected by the choice of item response models (Jacob 2016). A correlation of country ranking on inequality measures by different item response model choices is however not provided by the survey organisers.

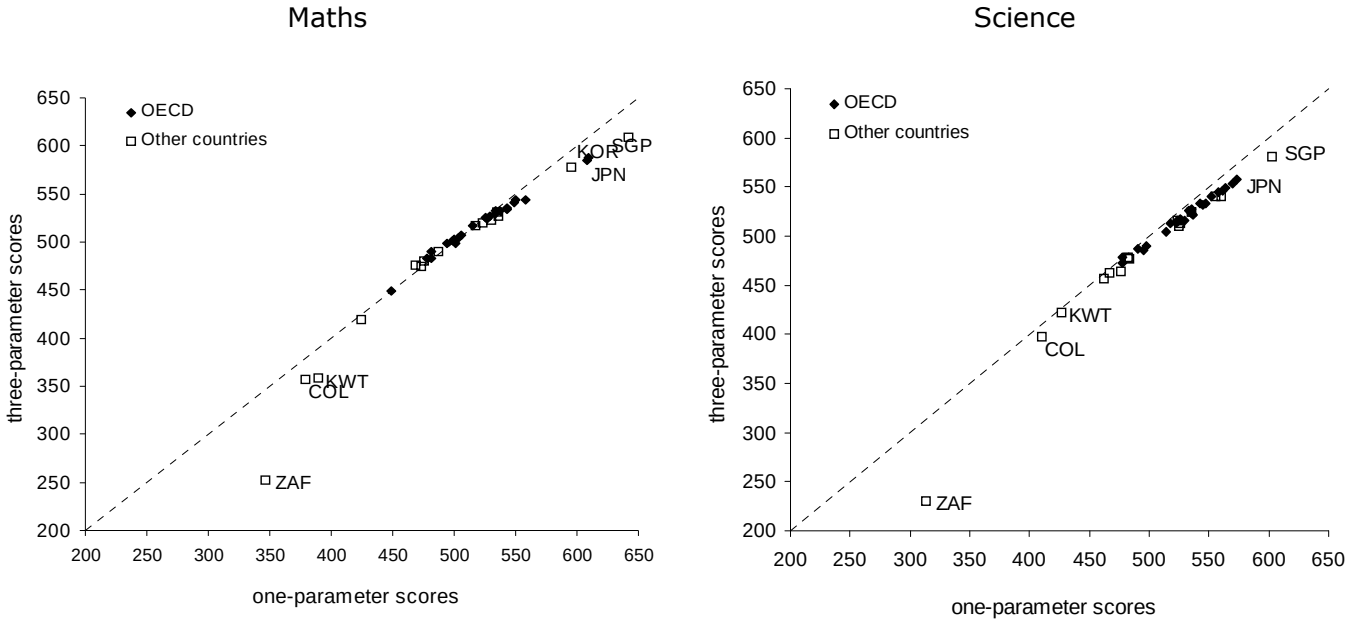
For most researchers, it is rather impractical to estimate different scaling models themselves for evaluating the robustness of survey organisers' item response model choice on educational achievement results reported. Jerrim et al (2018) approached the task investigating whether the new changes to the scaling method for PISA 2015 impact on mean, standard deviation and different percentiles of countries' achievement distribution. However, given the computational complexity involved they limited their focus on OECD countries and the year 2015. Their results imply that the change to another item response model and further alterations of the scaling model implemented with PISA in 2015 did only lead to trivial impacts on cross-country comparisons.

These results stand in contrast to an earlier study by Brown et al (2007) who examined the robustness of item response models by comparing the impact of the change from the one parameter model used in TIMSS 1995 to the three parameter model in TIMSS 1999 (whereby for PISA the main change is from a one-parameter model to a two-parameter model). In order to make results comparable over time, TIMSS organisers provided retrospectively achievement scores estimated with the three parameter model for 1995 data. As a consequence, for TIMSS 1995 data, achievement scores are available

that were estimated with two different item response models using exactly the same underlying 'raw data', which are the initial points each child scored on the test.

Brown et al (2007) exploited the data to see how results concerning countries average achievement and educational inequalities changed depending on IR model choice. They showed that the correlation between the derived scores produced from the IR model and the raw scores is lower for the three-parameter model. The extent of the change of achievement distributions varies from country to country.

Figure 2: Comparison of medians of one-parameter and three-parameter values

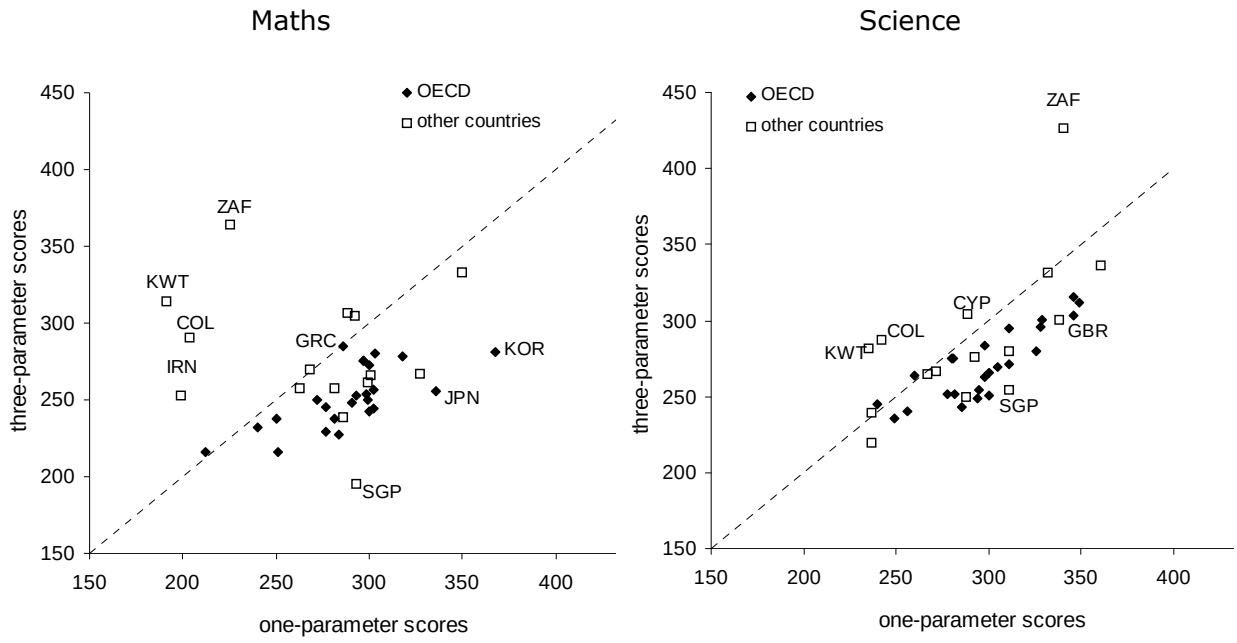


Note: the correlations of one- and three-parameter medians are 0.98 for maths (1.00 for OECD countries) and 0.97 for science (0.99 for OECD countries).
Source: Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007) International surveys of educational achievement: how robust are the findings? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(3), p. 638.

Figure 2 and 3 used from Brown et al (2007) compare the results between the two item response models which are based on identical raw data. Any derivation from the 45 degree line is due to the change in item response models. Figure 1 shows that the medians of achievement scores are very highly correlated for both maths and science. This is similar to what is found in OECD (2016a) and Jerrim et al (2018) for the switch from PISA one parameter to PISA two parameter item response model. Results however are very different once the focus is on the difference between the 95th and 5th percentiles, a measure of inequality in educational achievement. Taking all countries into account, for maths the correlation between the two sets of values is essentially zero (0.03), for science it is much better with 0.67. However, if the focus is only on OECD

countries, the correlation of educational inequalities are much higher (0.70 for maths and 0.85 for science). As a consequence and in contrast to the median, the cross-country pattern of educational inequality is therefore far from robust to the choice of IR model for TIMSS.

Figure 3: Comparison of P95-P5 of one-parameter and three parameter values



Note: the correlations of one- and three-parameter values of P95-P5 are 0.03 for maths (0.70 for OECD countries) and 0.67 for science (0.85 for OECD countries).

Source: Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007) International surveys of educational achievement: how robust are the findings? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(3), p. 639.

In contrast, results by Jerrim et al (2018) show correlations as high as 0.99 also for inequality measures for PISA data. There are two probable explanations for the differences in the results on robustness. First, Jerrim et al (2018) do not include OECD countries. As TIMSS results suggest, it could be mainly less prosperous countries leading to item response model choices not being robust. An indication for a similar trend could be that pupil level correlations between the different PISA IR models are lowest for those less affluent OECD countries (Turkey, Chile and Mexico) covered by Jerrim et al (2018). Second, for PISA the scaling method changed from a one to a two parameter model, while for TIMSS the one parameter model was mainly replaced by a three parameter model including an additional parameter on 'guessing'. It could well be that this latter guessing component drove the results found for TIMSS. Nevertheless, no clear

conclusions can be drawn. Item response model choice and its impact on country rankings remain unclear.

Concluding from this and given that researchers cannot replicate results based on the raw data easily, educational achievement organisers need to provide comprehensive and in-depth sensitivity analyses of their choice of items, item response models and other assumptions on the results they produce in a clear and accessible form to the research community (Araujo 2017, Schnepf and Volante 2017).

2.2 Potential survey errors of educational achievement estimates deriving from a possible lack of representativeness of the sample

Besides estimates' accuracy being influenced by measurement issues of the survey design, also discrepancies in the representativeness of the survey data can cause bias. This is the case if the resulting sample drawn is not representative for the target population.

Target population

In general, educational achievement surveys target school pupils of a specific age or grade. PISA's target population is 15-year-old students attending educational institutions in grade 7 and higher. The focus on school children derives from a practical aspect: in order to achieve a representative sample a sampling frame needs to be available. The sampling frame must contain a list of all individuals in the target population. Survey organisers sample first from a list of all schools teaching the pupils of their target population in the country. For the selected schools the sampling frame is simply the list of all students in the target population attending the school. If out of school children were considered in the target population, representative sampling would be difficult, since a comprehensive list of all out of school children is difficult to obtain.

However, the choice to focus on school children only can be problematic for countries where the number of children out of school is high. Indeed, the OECD sheds light on this issue providing a 'Coverage Index 3' in its technical reports, which depicts the number of the targeted school population by PISA expressed as a percentage of the population of all 15 year olds by country. While for PISA 2015 the Coverage Index 3 was relatively close to 100% for affluent countries, PISA covered for example only about 50% of the same age population in Vietnam, around 60% in Mexico and 70% in Turkey, Peru and Brazil. (OECD 2016b) Low coverage of the population of 15 year olds is associated with a number of problems. First, the examination of school children only seems too narrow a focus for evaluating a country's education system. This is especially important given that current literature generally does not discuss the Coverage Index 3. (Spaul

2017) Second, the often used figure on the percent on children who are not functionally literate (the percent of individuals below PISA level 2) is difficult to interpret if the number of same age children not covered in the survey is considerable. Third, the target population varies over time if the number of out of school children changes. Spaul (2017) shows for Turkey that PISA targeted only 36% of the 15 year old population in 2003 compared to 70% in 2015. He draws the conclusion that this change in the coverage of the population has a substantial effect on results comparing achievement across time and achievement gaps by socio-economic background: '*Perhaps unsurprisingly, the analysis showed that when PISA ineligible 15-16 year olds are accounted for the gaps between rich and poor are bigger than was previously thought and the improvements over time are larger than traditionally reported by the OECD*.'

While the focus on school pupils is the most practical one in order to achieve a representative sample of pupils, it is important to remember that the interpretation of the data for countries with low values of the Coverage Index 3 is limited to school children only. As Spaul (2017) shows, conclusions on time improvements and gaps between rich and the poor are not very sensible in this context. The out of school children are mainly the poor. In addition, any evaluation of an education system would need to consider how well it channels all children into education.

Coverage error: exclusions from the target population

Exclusions from the target population can lead to coverage error. The coverage error depends on a) the percent of the target population not covered in the sampling frame and b) the differences in achievement between the covered and non-covered population. PISA organisers allow omitting students with special educational needs and newly arrived immigrants from the target population. This has raised some concern (i.e. Wuttke 2007), because some countries excluded more students than the five percent threshold set by PISA organisers. Clearly this can lead to coverage error: achievement estimates are likely to be upwards biased for those countries with high exclusion rates since exclusion criteria are generally associated with lower achievement. To the knowledge of the author, there is no research available that examines the different practices by country in detail and estimates the impact of the exclusion on countries' mean achievement estimates and positon in the league tables.

Non-response error

Another crucial issue is potential non-response bias of educational achievement scores. Achievement survey organisers use similar thresholds for limiting possible non-response bias. For example, in PISA organisers set a threshold of 85% for school response and 80% for student response which need to be met for avoiding further

investigation of the data quality. Such thresholds, however, are no guarantee that non-response bias will be negligible since besides the non-response rate the pattern of response impacts on non-response bias. Low response may result in little bias if respondents and non-respondents are similar. On the other hand, high response can still yield high non-response bias if the group of respondents and non-respondents are very different. Despite this, only PISA countries who do not meet the response threshold are required to examine the non-response pattern. In line with this arguable choice, the PISA reports provide information on the extent of school and student response by country, but no information on cross-national differences in response patterns. The latter are very important, since if response bias differs between countries, country ranking results will be sensitive to these biases. This point is even more striking, since the OECD weight provided to the research community aims to correct for non-response bias at the school level but does not do so for non-response patterns at the student level.

Micklewright et al (2012) non-response in the England 2000 and 2003 PISA data. Using merged PISA and administrative school data, the authors exploit rich auxiliary information on respondents' and non-respondents' cognitive ability that are highly correlated both with response and the learning achievement that PISA aims to measure. They show that for both 2000 and 2003 England data, students with lower ability are less likely to agree sitting the PISA test. For both years, the overall achievement score for English students is therefore upwards biased. They then construct a generalised regression weight, that accounts for differences between the composition of the PISA sample of responding pupils and the composition of the population from which the sample is drawn. This weight can be used to estimate the extent of response bias for England.

Table 1 is an extract of Table 8 taken from Micklewright et al (2012) for PISA 2000. The design value provides the sample value just correcting for different selection probabilities of the sample. The OECD value provides the estimate once the OECD weight is applied which corrects for school but not student non-response. The Greg value provides estimates applying the generalised regression weighting, taking population characteristics into account. It is very obvious, that for all three achievement measures non-response bias (the difference between the 'true' and the 'estimated' PISA score) is huge, leading to a considerable upwards bias of reported results for England. OECD weights do little to correct for the biases found. This reflects the lack of adjustment in the OECD weights for the pattern of pupil response, which is the principal source of bias.

Table 1: Estimates of characteristics of distribution of PISA test scores using different weights, 2000

Weight	Maths	s.e.	Reading	s.e.	Science	s.e.
Mean						
Design	531.3	4.02	525.7	4.18	535.8	4.37
OECD	531.0	4.41	525.0	4.70	535.3	4.84
GREG	516.8	1.59	510.5	1.59	521.3	1.76
% < PISA level 2						
OECD	n.a.	n.a.	12.43	1.06	n.a.	n.a.
Propensity	n.a.	n.a.	14.18	1.23	n.a.	n.a.
GREG	n.a.	n.a.	15.68	0.72	n.a.	n.a.
Differences between means						
Design – GREG	14.5	3.83	15.2	3.88	14.5	4.01
Differences between % < level 2						
Design – GREG	n.a.	n.a.	-3.73	0.71	n.a.	n.a.

Source: Micklewright, J., Schnepf, S. V., and Skinner, C. J. (2012) Non-response biases in surveys of school children: the case of the English PISA samples, *Journal of the Royal Statistical Society. Series A (General)*, selected results from Table 8 p. 931.

The bias is considerable being two to three times bigger than the published standard error. However, once effects of the bias on the country ranking in the PISA league tables are considered England's position shifts downward by only a small number of places.

Since PISA 2000 and 2003, England's school and student response has improved considerably. Nevertheless, to the knowledge of the author, a similar exercise of examining non-response bias for more recent educational achievement survey data and other countries is not available.

Certainly, similar examinations to those conducted for England are only possible in those countries that keep a register of pupils and have information on their test scores deriving from national tests sat close to the timing of the educational achievement survey. However, this would be possible in a number of countries like for example the US, Portugal, Sweden, the Netherlands, Estonia and Italy.

A cross-country examination of non-response patterns and possible bias would be beneficial due to the following four reasons. First, it would overcome uncertainties of country rankings due to possible non-response bias. Second, it would start a discussion whether non-response by students is an important factor that should be considered to be included in the creation of weights for educational achievement data. Third, if one knew which school and student characteristics are associated with response, the sample design of the data could take this into consideration thereby improving the representativeness of the sample. However, with the exception of England (Durrant and Schnepf, 2017) there is scant information on non-response patterns for achievement surveys across countries. Fourth, since item response models use characteristics of students to estimate their achievement score, response bias could possibly impact on these estimates as well.

2.3 Potential survey errors of educational achievement deriving from the analysis stage

The analysis of educational achievement data and the interpretation of its results need to take the specific design of the surveys and the complex methodology for creating achievement scores into account. However, for many applied researchers and economists the caveats of the methodological data design and the resulting requirements for analysis are far from obvious (Jacob and Rothstein 2016, Jerrim et al 2017). The analysis stage of the educational achievement survey data could therefore be a further source for errors of estimates to arise.

This section discusses potentially flawed analyses based on educational achievement data using three examples. First, research results from standard econometric models can be flawed if the complex design of multiple imputations of achievement scores for different subject domains is not taken into account. Second, causality conclusions cannot be easily drawn from the data. Third, the measurement of peer effects is problematic.

Analyses comparing achievement across different subject domains

In order to reduce pupils' response burden to PISA, they are not required to answer the entire set of survey questions used for creating achievement scores. Instead, pupils answer questions of randomly assigned booklets; sometimes they answer only questions on the key domain of the PISA survey. Nevertheless, the final data base provides achievement scores for all students on all domains. This is achieved by multiple imputations which take students' and school dummies and their answers to their booklet

into account. As a consequence, researchers need to be careful once they want to draw conclusions from comparing pupils' achievements between different domains.

Jerrim et al (2017) discuss this in great detail for an analysis which uses a standard econometric procedure of individual fixed-effects for comparing each pupils' performance between different subjects. While this model theoretically can measure within—pupil variation, applied to the complex data constructed from educational achievement surveys the results can be driven largely by variation deriving from the survey's method of imputation. Jerrim et al (2017, p. 57) conclude that '*some fairly standard econometric approaches should only be applied to these data with caution, and require an additional set of important robustness tests. More generally, a key lesson from this paper is that the statistical techniques required to robustly analyse resources such as PISA are perhaps more complicated than first meets the eye*.' (The same paper also discusses the importance of using PISA weights once handling educational achievement data, a topic important and applicable to all survey data.)

Causality

There is general agreement in academia that the main limitation of PISA is its reliance on cross-sectional data. In contrast to longitudinal data which follows the same students over the course of their school careers, cross-sectional data comprises a different sample of students for each round. Therefore, scholars (e.g. Goldstein 2017) have repeatedly argued that PISA should not be used for the purpose of drawing specific policy implications for improving education systems. Nevertheless, claims of causal relations based on cross-sectional achievement data are common in the research community. Given the limitation of cross-sectional data, Cordero and Cristobal (2017) describe strategies to still derive causal inferences from educational achievement surveys by using counterfactual impact evaluation. Nevertheless, the question remains whether the research community and policy makers would not be better served by having at least one of the large cross-sectional cross-country surveys (PISA, TIMSS, PIRLS) being replaced by a survey with longitudinal design.

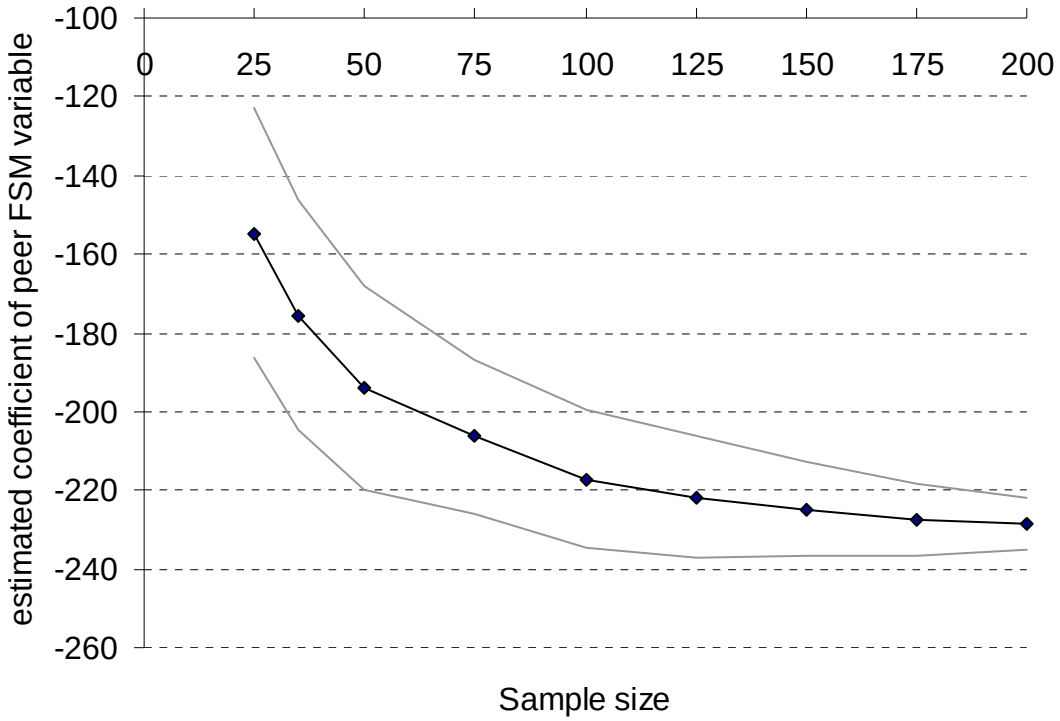
Measurement of peer effects

One part of education research investigates the impact of a student's peers on his/her achievement results. Educational achievement surveys have been used to investigate the so called 'peer effect' cross-nationally (e.g. OECD 2007 (Chapter 5), Entorf and Lauk, 2008, Schneeweis and Winter-Ebmer 2007). However, the survey's design means that only a small random sample of peers (generally around 30) is observed for each individual (in contrast to all peers). The summary statistic of peer attributes is based on the survey data and hence subject to sampling variation. This

generates measurement error. As a result, the estimated explanatory 'peer group' coefficients is subject to downwards attenuation bias in an OLS regression with the dependent variable 'achievement score'. The problem has been recognised before (i.e. Ammermüller and Pischke 2009).

Micklewright et al (2012) were able to quantify the extent of the bias in peer group estimates obtained. Using English administrative data merged with PISA 2003 data, they could estimate the peers' socio-economic background (based on receipt of free school meals, a state benefit for low income households) using PISA and calculate the true value using population data. Results show substantial attenuation bias when measuring peer receipt using just the peers present in the survey data. Using Monte Carlo simulations, Figure 4 shows how the peer group coefficient changes depending on the sample size of the peers. The bias increases non-linear as peer sample sizes fall. The attenuation bias is about one third in the peer group coefficient with the sample size of 35 students implied by PISA's survey design.

Figure 4: Monte Carlo simulation of the effect of changing within-school sample size on the estimate of the peer group FSM coefficient



Note: the graph shows how the peer group effect estimate changes depending on the sample size of peer group.

Source: Micklewright, J., Schnepf, S. V., and Silva, P. N. (2012) Peer effects and measurement error: the impact of sampling variation in school survey data (evidence from PISA), *Economics of Education Review*, *31*(6), Figure 4 on p. 1141.

As a consequence, caution is needed when estimating peer effects with educational achievement data, but attenuation bias should be bigger in countries where schools are less socially segregated and hence where peer groups are less homogenous. (Micklewright et al 2012).

2.4 In sum: the potential total survey error of educational achievement estimates

Potential survey errors deriving from the measurement (Section 2.1), the representation (Section 2.2) and the analysis stage (examples given in 2.3) contribute to the so called 'total survey error'. The total survey error provides a measure of overall accuracy of a survey estimate taking all possible errors into account. The total survey error is conventionally expressed by the mean squared error. Unfortunately, the mean squared error can generally not be computed because this would require an error free estimate. It is defined as the square of the bias plus the variance. The quadratic term in the formula of the mean-squared error shows that if the bias component of an error increases, it can quickly inflate the total survey error. A big mean squared error indicates that the total survey error is big as well. This is problematic since it limits the accuracy of inferences that can be drawn from the estimate like for example whether results for one country differs from that of another.

The previous discussion of errors deriving from measurement, representation and analysis shows that there is a considerable reason to assume that errors on the measurement side like validity (deriving from a questionable fit of measure to underlying construct) and measurement and processing error (deriving from the choice of items and item response models) and errors on the representation side like coverage error (deriving from exclusions from the target population) and non-response errors are unlikely to be negligible. For example, as discussed above the non-response error was about two to three times higher than the standard error for the English PISA sample in 2000 and 2003. Even though this error is likely to be overestimated for current rounds of England (nothing can be said about other countries due to the lack of research on the topic) this result is concerning especially since non-response is just one component of the total survey error. Obviously, not much can be said about the size of the other survey errors discussed.

The standard error is the only error of achievement estimates published in educational achievement reports. This follows the usual practice for presenting research results. Nevertheless, the considerations above show that the standard error could very well be only a small part of the total survey error of educational achievement surveys. This questions furthermore the presentation of educational achievement results by ranking countries in a league table in survey reports and the media. The lack of

transparency on the data collection and model choices does not allow guaranteeing that countries' positions are indeed determined by their students' educational achievement and not influenced by the size of total survey errors resulting from the survey design.

Given the considerable impact of the surveys' results on education policy, the survey organisers should increase transparency and discuss possible error sources in greater detail and provide estimates of their extent. While admittedly this is difficult for some areas (validity), it would be well possible to examine non-response bias, coverage error and measurement errors to some degree. Furthermore, greater detail on the robustness of achievement results to the choice of item response models would serve the research community and policy makers.

3 Conclusions

While educational achievement surveys revolutionised research on education cross-nationally, the surveys have been repeatedly subject of heated debate since first results were published. This paper focused on possible errors deriving from the survey methodology implemented by organisers of educational achievement surveys. The main focus was on the choice of educational achievement measures, target population, item choice, item non-response model choice, educational achievement scale used and non-response. It was shown that there are many reasons to assume that the total survey error associated with countries' educational achievement estimates is likely to be inflated by other errors besides the reported standard error.

Given the policy relevance of the surveys' estimates, policy makers and the research community would greatly benefit from survey organisers providing more transparency on the different potential errors of educational achievement estimates. This would include information on how the modelling choices impact on the results and in-depth examinations of representativeness of country data. Without this information, the generation of the data and its accuracy is not transparent and as such justifies questions on fitness of educational achievement data for policy making.

References

Ammermueller, A. and Pischke, J.-S. (2009) Peer effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study, *Journal of Labor Economics* 27(3): 315-48.

Araujo, L., Saltelli, A., & Schnepf, S. (2017) Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development*, *19*(1), 20.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007) International surveys of educational achievement: how robust are the findings? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(3), 623-646.

Cordero, J and Cristobal, V. (2017) Causal inference on education policies: a survey of empirical studies using PISA, TIMSS and PIRLS, *Journal of Economic Surveys*.

Durrant, G., and Schnepf, S. (2017). Which schools and pupils respond to educational achievement surveys? A focus on the English PISA sample. *Journal of the Royal Statistical Society A*.

Entorf, H and Lauk, M (2008) Peer Effects, Social Multipliers and Migrants at School: An International Comparison, *Journal of Ethnic and Migration Studies* 34(4): 633-645.

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

Goldstein, H. (2004). International comparison of student attainment: some issues arising from the PISA study, *Assessment in Education*, 11, 319-330.

Goldstein, H. (2017) Measurement and Evaluation Issues with PISA, in Louis Volante (ed.), *The PISA Effect on Global Educational Governance*, Routledge.

Groves, R., Floyd, J., Couper, P., Lepkowski, J., Singer, E. & Tourangeau, R. (2009) *Survey Methodology* (Hoboken, Wiley).

Hopmann, S.,Brinek, G. and Retzl, M. (Eds) 2009, *PISA according to PISA*, University of Vienna Press, Vienna

Jacob, J (2016) Student test scores: How the sausage is made and why you should care, *Economic Studies at Brookings, Evidence Speaks Reports*, Vol 1(25).

Jacob, B. and Rothstein, J. (2016) The Measurement of Student ability in Modern Assessment Systems, *Journal of Economic Perspectives* 30:3, 85-108.

Jerrim, J.; Lopez-Agudo, L.; Marcenaro-Gutierrez, O. and Shure, N. (2017) What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review.*

Jerrim, J., Parker, P., Choi, A., Chmielewski, A., Sälzer, C. And Shure, N. (2018) How robust are cross-country comparisons of PISA scores to the scaling model used? *Educational Measurement: Issues and Practice*.

Kreiner, S. and Christensen, K (2014) Analyses of model fit and robustness. A new look at the PISA scaling model underlying rankings of countries according to reading literacy, *Psychometrica* 79(2): 210-231.

Meyer, H.-D. and Zahedi, K. (2014) An open letter: to Andreas Schleicher, OECD, Paris; *Global Policy Institute*, 5 May and Guardian, 6 May, available at: www.globalpolicyjournal.com/blog/05/05/2014/open-letter-andreas-schleicher-oecd-paris; www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics (accessed 12 April 2017).

Meyerhoefer, W. (2007) Testfähigkeit – Was ist das? [Does PISA keep what it promises?], in S. Hopmann, G. Brinek, and M. Retzl (Eds) *PISA according to PISA*, Vienna, University of Vienna Press.

Micklewright, J., Schnepf, S. V., and Silva, P. N. (2012) Peer effects and measurement error: the impact of sampling variation in school survey data (evidence from PISA), *Economics of Education Review*, *31*(6), 1136-1142. DOI: 10.1016/j.econedurev.2012.07.015

Micklewright, J., Schnepf, S. V., and Skinner, C. J. (2012) Non-response biases in surveys of school children: the case of the English PISA samples, *Journal of the Royal Statistical Society. Series A (General)*, 915-938.

OECD (2004) Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003, Paris, OECD Publishing.

OECD (2007) PISA 2006 – Science Competencies for Tomorrow's World. Volume 1: Analysis, OECD Publishing, Paris.

OECD (2012) *PISA 2012 Technical Report*, Paris, OECD Publishing, Paris. .

OECD (2016a), *PISA 2015 results (Volume 2)*, OECD Publishing, Paris.

OECD (2016b), PISA Technical Report, OECD Publishing, Paris.

Prais, S. J. (2003) *Cautions on OECD's recent educational survey* (PISA). Oxf. Rev. Educ., 29, 139–163.

Schneeweis, N. and Winter-Ebmer R (2007) Peer effects in Austrian schools, *Empirical Economics* 32: 387-409.

Schnepf, S. and Volante, L. (2017). PISA and the future of Global Educational Governance. In L. Volante (Ed.), *The PISA Effect on Global Educational Governance* (pp. 217-226). (Routledge Research in Education Policy and Politics). New York: Routledge.

Sjoberg, S. (2007) PISA and "real life challenges": Mission impossible? in Hopman, S (ed) *PISA according to PISA, Does PISA Keep What It Promises?* Wien: LIT Verlag

Spaul, N. (2017). Who makes it into PISA? Understanding the impact of PISA sample eligibility using Turkey as a case study, *OECD Education Working Papers* No. 154.

Wiseman, A. and Waluyo, B. (2017) 'The Dialectical Impact of PISA on International Eudcationl Discourse and National Education Reform, in Louis Volante (ed.), *The PISA Effect on Global Educational Governance*, Routledge.

Wuttke, J, (2007) in Hopman, Brinek and Retz (eds): *Pisa According to Pisa*, Wien: Lit_Verlag.

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*):

00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

More information on the European Union is available on the internet (http://europa.eu).

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);
- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or
  calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

KJ-AE-18-005-EN-N

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

Publications Office

doi:10.2760/219007

ISBN 978-92-79-85795-9