Insights into survey errors of large scale educational achievement surveys

Sylke V. Schnepf
European Commission, Joint Research Centre, Ispra, Italy

Paper prepared for the 49th Scientific meeting of the Italian Statistical Society,
20 to 22 June 2018

April 2018

Abstract

While educational achievement surveys revolutionised the research possibilities for answering policy relevant education questions in an international context, the surveys have been repeatedly subject of heated debate since first results were published. This paper discusses the soundness of survey methodology of large scale achievement surveys aiming to summarise results of the author's and other up to date research. Thereby, different components of the 'total survey error' will be discussed covering among others validity, non-response errors and the choice of item response models on results. Findings indicate that there are legitimate concerns about what educational achievement surveys measure and how. Survey organisers should be more transparent in the documentation of the methodological choices that underlie the creation of the data and more explicit about the impact of these choices on the results.

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.

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.

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 maths 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 measure school curriculum (TIMSS) or life skills (PISA and PIRLS) and collect in addition 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 in an unprecedented way leading also to PISA results having become highly influential for policy formulation (Schnepf and Volante, 2017).

Probably especially due to the importance of the surveys for policy design, they have been repeatedly the subject of heated debate since first results were published (i.e. Prais 2003, Brown at al 2007, Hopmann et all, 2009; 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, 20014) 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 let to some exchange with the OECD.

This paper was prepared as background for a guest lecture at 49th Scientific meeting of the Italian Statistical Society, 20 to 22 June 2018. It will not discuss conflicting values in the PISA controversy as displayed in *The Guardian* letter. Instead, it focuses on the soundness of survey methodology of large scale achievement surveys, a topic the author has examined from different angles in the past. While the discussion is framed around the most policy relevant survey PISA, it is mostly applicable to educational achievement surveys in general due to similar survey methodologies used.

2 Current problems of measuring educational achievement with cross-national educational achievement surveys

The quality of any survey including educational achievement surveys depends on two aspects: the quality of the measurement of the abstract construct it aims to capture and the representativeness of the sample for the population the survey aims to describe (Groves et al, 2009). Measurement (Section 2.1) and representativeness (Section 2.2) have both several quality components to be met. If these are not achieved they contribute to the total survey error (Section 2.3). In addition, researchers should avoid some caveats of the use of educational achievement surveys. Two examples are discussed in Section 2.4.

2.1 What does PISA measure?

A first problem of the existing PISA data relates to its comparability across countries for assessing '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' (OECD, 2004, p. 12). The OECD assumes the life skills needed to function in knowledge societies to be the same for all countries they cover. For example, 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 is therefore rather dubious and raises legitimate questions about how appropriate it is to rank them in a single table. The skills needed by a young adult are likely to depend on the characteristics of the society in which the person is living; hence, what it means to 'function' in a knowledge society will vary from country to country. While comparability of data across countries is desirable, a question raised by critics is whether comparability in PISA is only achieved by ignoring the great diversity of curricula across the participating countries – diversity which might in fact be a source of country-specific creativity and well-being.

Does PISA meet the requirements of validity?

Assuming that the construct of ‘the challenges of today’s knowledge societies’ is a reasonable one to explore, irrespective of cultural or developmental specificities, questions remain about how well the selection of items and the item response model used to summarize item answers into one overall score fit the abstract skill that PISA aims to measure. More specifically, the main criticism in the literature concerns the multidimensionality of the items being measured. Meyerhoefer (2007) discusses how the items used do not only measure, for example, ‘maths’ ability but also a student’s ability to comply with the test structure. According to Goldstein (2004, 2017), education skills are multidimensional. However, in PISA, in order to derive student scores, uni-dimensional item non-response models are chosen. This impacts on the choice of items. Items that demonstrate ‘poor psychometric characteristics in more than ten countries (‘dodgy’ items)’ can be deleted (OECD, 2012, p. 148). Such items are those most prone to multidimensionality and to reflect cultural bias (Goldstein, 2004). Their removal from the set therefore obscures differences between countries that might otherwise demonstrate greater heterogeneity on varying educational dimensions. For Goldstein, the aim of a cross-national study should not be to neglect, but rather to obtain, information on underlying differences between countries.

It is common for surveys to measure only one aspect of a specific construct; however, the specific aspect needs to be defined. It is less than clear in PISA which dimensions are being measured and which suppressed.

Furthermore, validity is very difficult to achieve 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. If items perform badly during these trials, they are withdrawn. 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.

The choice of item response models

Another problem regards the choice of item response models on league table rankings and variation within countries. Item response (IR) models estimate a person's 'proficiency' in the subject concerned (maths, science, reading, etc) from answers to a number of questions. The achievement scores are therefore derived data and the question arises as to whether the choices made over the method of derivation have an appreciable impact on the surveys' results. First, a choice needs to be made whether uni- or multi-dimensional IR models are used. Educational achievement surveys use generally uni-dimensional IR models, and therefore, as discussed above, are more likely to neglect differences between countries. Second, there is a choice between different uni-dimensional IR models. Normally, not much information can be gathered on organisers IT model choice. But even where secondary analysis is made of the microdata, the procedures involved in fitting the relevant IR models are sufficiently complex that it is impractical for most researchers to estimate variants for understanding the sensitivity of their results to alternative models. Nevertheless, if researchers would still like to embark on calculating the impact of the choice of IR model on rankings they cannot do so since not all items are made available to the research community.

To the knowledge of the author, the only opportunity to compare how educational achievement survey results on league tables change depending by IR model used derive from TIMSS 1995 data, for which achievement scores from two different IR models, the one parameter model and the three parameter model, were provided. TIMSS organisers used the one parameter model for TIMSS 1995 and the three parameter model for TIMSS 1999. In order to make results comparable over time, they retrospectively provided 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 IR models using exactly the same underlying 'raw data', which are the initial points each child scored on the test.

The one parameter model allows for differences in the degree of difficulty of each question. The three parameter model, now the preferred model for achievement surveys, allows in addition for the probability that the answer is simply guessed and for the ability of a question to discriminate between students of high and low ability. Formally, the models give the probability of a correct answer to question i by student j as:

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

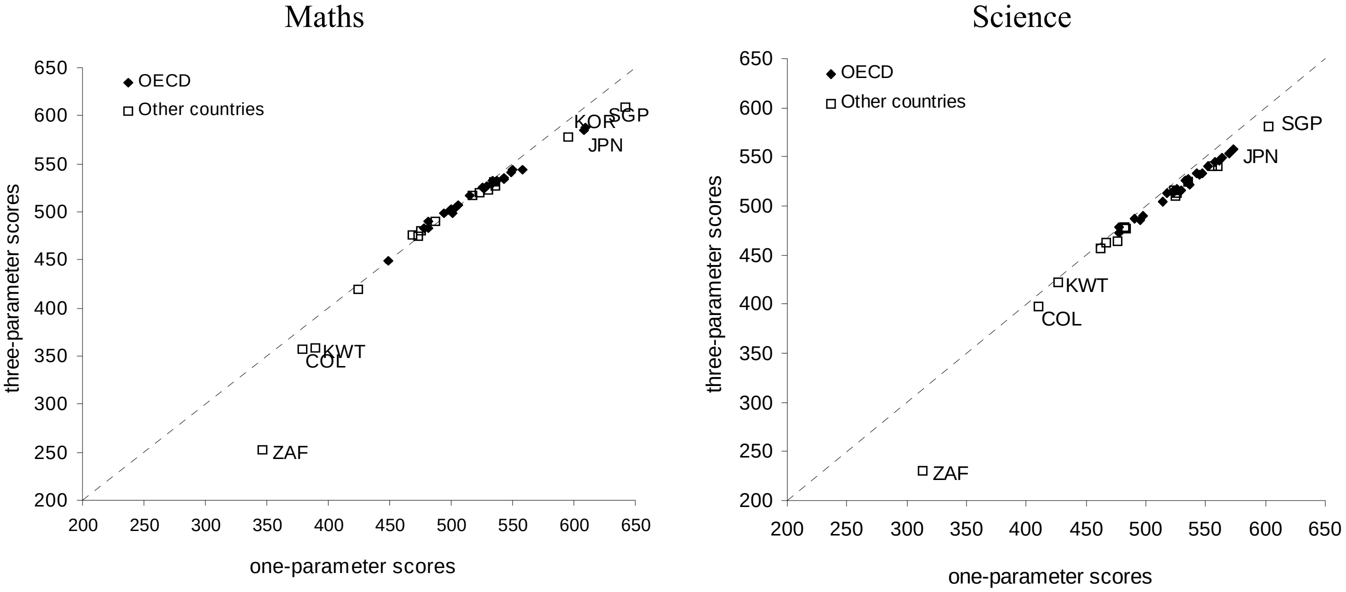
Three parameter model: $pij(\text{correct answer}) = \gamma i + (1- \gamma i) /[1+\exp(-\beta i(\theta j - \alpha i))]$

where θj is a student's proficiency, αi is a question's difficulty, γi is the probability that the answer to a question is guessed, and βi measures the power of a question to discriminate between individuals of high and low ability.

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 show that 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 1 and 2 compare the results between the two item response models which are based on identical raw data. Hence, 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. The rankings hardly change at all even though a few countries lie some way off the 45 degree line.

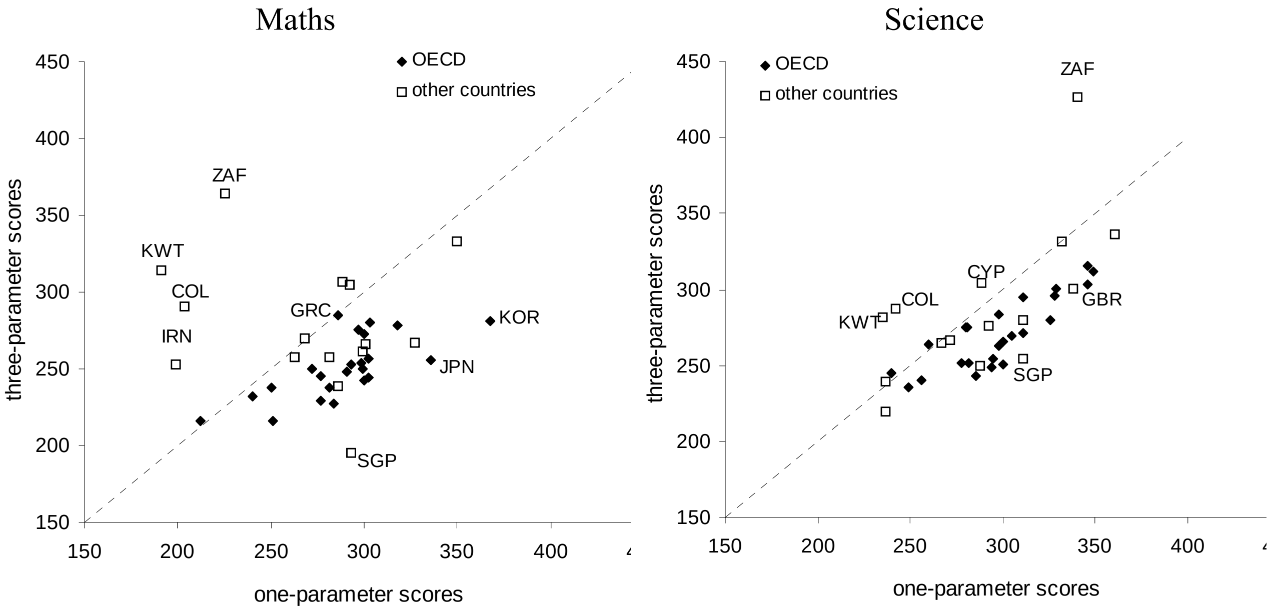
Figure 1: Comparison of medians of one-parameter and three-parameter values



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 et al 2007.

Results however are very different once the focus is on the difference between 95th and 5th percentiles, a measure of inequality in educational achievement. For maths, the correlation between the two sets of values is essentially zero (0.03), for science it is much better with 0.67. 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.

Figure 2: 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 et al 2007.

While TIMSS 1995 is now more than 20 years old, it provides to the knowledge of the author the only opportunity for examining the impact of the choice of the item response model on educational achievement results used for policy design. Its results show that the choice of IR model has the potential to change ranking of countries in the league table. As such, the examination of the impact of IR choice on results is of great importance. However, the assessment of the impact is not straight forward. Even for those researchers who can estimate the more complicated IR models, the lack of provision of all items makes the reproducibility of country rankings from the raw data impossible. This implies that educational achievement organisers need to provide results of 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, Brown et al 2007).

Does educational achievement really follow a ratio scale?

Educational achievement scores derived from IR models are on a ratio scale. The meaning of the ratio scale used to measure ability must also be considered. Is a child who achieves a PISA score of 250 only half as well equipped for life as a child scoring 500? Atkinson (1975) notes that 'there is at present really no such thing as the distribution of ability: the distribution depends on the measuring rod used and cannot be defined independently of it. […] the fact that most IQ tests lead to a distribution of scores which follows the normal distribution does not necessarily tell us anything about the distribution of abilities: it may simply reflect the way in which the tests have been constructed' (1975, p. 89). This caveat should preface any educational achievement report.

2.2 Is PISA representative for 15 year-olds?

Besides the variety of problems associated with the process of translating a specific choice of an education construct into an overall student score, PISA has been recently criticised over whether the sample of students used for estimating the country average of 15 year-olds' achievement is representative for their population (Mickelwright et al, 2012; Araujo et al, 2017).

Survey coverage: exclusions from the target population

In particular, once sample selection is agreed upon, PISA organisers allow for the exclusion of students with special educational needs and newly arrived immigrants. This has raised some concern (i.e. Wuttke 2007), because some countries excluded more students than the five percent threshold set by PISA organisers.

Non-response bias

Another crucial issue is the 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 obvious possibility, 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 while the OECD weight provided to the research community aims to correct for non-response bias at the school level, it does not do so for possible non-response bias at the student level.

Micklewright et al (2012) however identify non-response patterns at the student level to be most important for non-response bias found in the England 2000 and 2003 PISA 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 (for the latter year England was excluded from the PISA report due not meeting response thresholds, for 2000 however it was assumed that non-response bias is negligible), 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 they are drawn.

Table 1 shows the results for PISA 2000, the year English data was deemed to have only a negligible non-response bias (in this year England had a 82% school 'after replacement' response rate and a 81 % pupil response rate, thereby being clearly considerably below the OECD average of the PISA sample, Micklewright et al 2012, Table 1).

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 et al 2012.

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.

The bias is considerable being two to three times bigger than the published standard error. Results are more moderate, if the impact of non-response on differences in achievement between countries is considered. England’s position shifts downward by three places for maths, two for science and none for reading if the non-response bias is taken into account in the 2003 league tables.

In later PISA rounds, response rates increased which might indicate that the data became more representative. While probably capturing more adequately low performing students after 2003, the English achievement scores seemed to reflect a trend of achievement decline relative to that of other countries. This perceived decline in achievement of England's pupils - resulting only from methodological problems - raised huge political debate within England. This is an example of how survey design errors may lead to unwanted side effects in the policy discourse. An in-depth description of how the problem of representativeness of data affected the English PISA data is provided in Jerrim (2011).

England had especially poor response to PISA 2000, data which by now are outdated. In addition, England's response has improved considerably since then. Nevertheless, to the knowledge of the author, a similar exercise of examining non-response bias for more recent educational achievement survey data is not available. The uncertainty about English data quality in 2000 and 2003 remains and higher response in subsequent survey rounds does not imply that any problems were absent. In addition, how important is non-response for other OECD and non-OECD countries? It would be useful to have similar analyses for all participating countries, and to examine in greater detail the response patterns across countries. Such an examination, which should not fall on data users, but data organisers could be useful in three ways.

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.

In sum, currently cross-national achievement surveys use rather simplified rules in order to ensure that data are representative for school pupils in countries' populations. While these rules are generally problematic, the lack of an investigation of these problems creates uncertainty on results and hampers opportunities to be grasped for any improvement of the survey design (i.e. selection of the sample and components to be considered for weights). The glamour of educational achievement surveys like PISA being constantly in the domain of policy makers should not blind survey organisers from their responsibilities to investigate the representativeness of their data and to strive to improve it.

2.3 The total error of the PISA score

Once questions about measurement issues or representativeness of data are concerned, each single problem discussed above can potentially contribute to the so called ‘total survey error’. The total survey error provides a measure of precision of a survey estimate taking all possible errors that can enter into the creation of the data into account. It is conventionally measured by the mean-squared error defined as the square of the bias plus the square of the standard error. The quadratic terms in the formula for the mean-squared error show that if errors increase, they can quickly inflate the total survey error.

The previous discussion of errors deriving from measurement and representation, shows that there is a considerable reason to assume that errors on the measurement side like validity (deriving from a questionable operationalisation of the construct to be measured) and measurement 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. To the knowledge of the author among those errors, the only other educational achievement survey error besides the standard error that could be quantified in the literature in the past was the non-response error for the English PISA sample in 2000 and 2003 (admittedly, besides the coverage error, which has also a considerable normative judgement component, these errors are difficult to quantify in general). As discussed above, the non-response error was about two to three times higher than the standard error. 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, nothing 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 is likely to be a rather 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 document better the choices made concerning measurement and representativeness of educational achievement surveys. This would increase the legitimacy of the information provided to the public.

2.4 Two examples of easy traps when using educational achievement data

As a note of caution for data users, educational achievement data are still limited in regard to specific research interests. In the following, the assumption of causality and the estimation of peer effects are discussed.

Causality

There is general agreement in academia that the main limitation of PISA is its reliance on cross-sectional data, i.e. data that refer in each round to different student cohorts. 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 because the various rounds do not follow the same students over the course of their school careers. Claims for causality based on cross-sectional data are problematic. Nevertheless, the predominant PISA-based policy narrative, which makes the economic case for education reforms, is based on an assumption of causality running from education to economic growth.

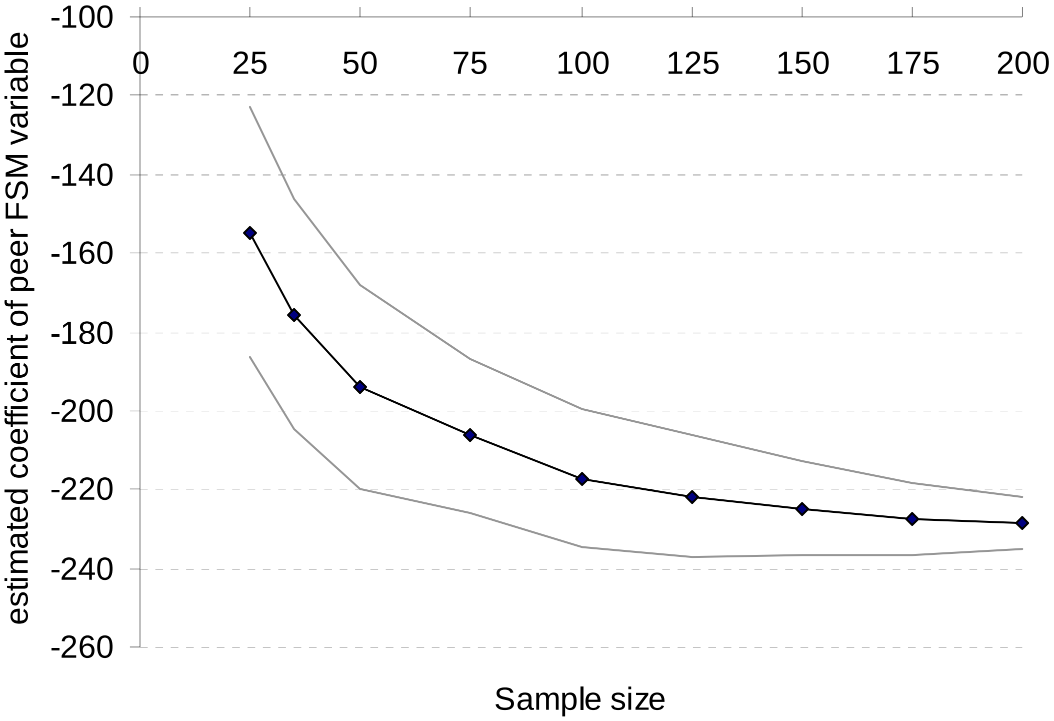
Measurement of peer effects

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, if the survey's design means that only a random sample of peers is observed for each individual, rather than all peers, then any summary statistic of peer attributes that is based on the survey data and used as an explanatory variable is subject to sampling variation. This generates measurement error. As a result, the estimated peer group coefficients in an OLS regression are subject to downwards attenuation bias. The problem has been recognised, for example by Ammermüller and Pischke (2009) for whom sampling variation is one source of error in peer group measurement.

Micklewright et al (2012) were able to quantify the extent of the bias in peer group estimates obtained. Having the National Pupil Database merged with English PISA 2003 school survey data, they have information on the population from which each sample of peers in the survey is drawn. The National Pupil Database includes information on whether a student receives Free School Meals (FSM), a state benefit for low income families. FSM is an indicator regularly used in the context of measuring students' socio-economic background in England's education literature using population data. Results show substantial attenuation bias when measuring peer receipt using just the peers present in the survey data. Figure 3 shows their results of a Monte Carlo simulation. 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.

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).

Figure 3. Monte Carlo simulation of the effect of changing within-school sample size on the estimate of the peer group FSM coefficient



Note: the series with symbols for each data point shows the mean value of the peer group FSM coefficient in regression models estimated for the same 3,459 individuals and with the same model specification as in Table 1. To generate each point, we estimate the model having randomly drawn a sample of the size indicated of peers (defined as 15 year olds in the same school) from the NPD for each individual, repeating the process 200 times and averaging the estimated peer FSM coefficient. The two series without symbols show the values of the mean +/− 2 standard deviations.
Source: Micklewright et al, 2012

3 Conclusions

Currently, there is a considerable controversy about the usefulness of educational achievement data, which are highly influential for policy design. This paper focused on possible errors deriving from the survey methodology implemented by organisers of educational achievement surveys, comprising the choice of educational achievement measure, 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, namely non-response error and measurement error.

As a consequence, survey designers should make a greater effort to document the choices they make in the generation of the data. 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 quality 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.

Atkinson, A. B. (1975) *The Economics of Inequality* (Oxford, Clarendon Press).

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.

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.

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

Jerrim, J. (2011): England's 'plummeting' PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline?, DoQSS Working Paper No. 11-09, http://www.ioe.ac.uk/Study_Departments/J_Jerrim_qsswp1109.pdf.

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. (2009) 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)*, 1-41.

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, Paris.

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

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.

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.