To weight or not to weight?: the case of PISA data

John Jerrim[1]; Luis Alejandro Lopez-Agudo[2]; Oscar D. Marcenaro-Gutierrez[2]; Nikki Shure[1]

[1]*Department of Social Science, UCL Institute of Education, University College London.*

[2]*Departamento de Economía Aplicada (Estadística y Econometría). Facultad de Ciencias Económicas y Empresariales. Universidad de Málaga.*

[1]20 Bedford Way London, WC1H 0AL. E-mails: J.Jerrim@ioe.ac.uk (John Jerrim); nikki.shure@ucl.ac.uk (Nikki Shure).

[2]Plaza de El Ejido s/n, 29013, Málaga (España). E-mails: odmarcenaro@uma.es (Oscar D. Marcenaro-Gutierrez); lopezagudo@uma.es (Luis Alejandro Lopez-Agudo).

International large-scale assessments (ILSA) –like e.g. PISA, TIMSS, PIRLS, etc.– have obtained a high worldwide popularity among researchers to study students' academic achievement. Nevertheless, in spite of their recently acquired relevance, there are few studies which really account for the complex survey and test designs that they present and follow the technical procedures suggested by their developers. The current study intends to provide researchers with a comprehensive explanation on how these databases should be dealt with and the drawbacks of not employing the recommended procedures. Furthermore, we also alert about the use of methodologies which would not be adequate in the context of these studies, as the common use of student fixed effects. Hence, the main goal of this research is to provide some kind of 'good practices guide' which can be a quick reference to researchers.

Keywords: survey design; test design; PISA; weights; replicate weights; plausible values.

Acknowledgements: This work has been partly supported by the Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía (PAI group SEJ-532 and Excellence research group SEJ-2727); by the Ministerio de Economía y Competitividad (Research Project ECO2014-56397-P); the FPU scholarship of the Ministerio de Educación, Cultura y Deporte (FPU2014 04518) and a short three-month stay funding (2016) at the Institute of Education (UCL), United Kingdom, by the Ministerio de Educación, Cultura y Deporte. Luis Alejandro Lopez-Agudo also acknowledges the training received from the University of Malaga PhD Programme in Economy and Business [Programa de Doctorado en Economía y Empresa de la Universidad de Malaga].

1 Introduction

International assessment programmes have received much attention over the last two decades, with academics, journalists and public policymakers all eagerly awaiting every set of updated results. Although the Programme for International Student Assessment (PISA) is perhaps the most well-known, a number of other studies fall into this group including the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS) and the Programme for International Assessment of Adult Competencies (PIAAC). These data are also increasingly being used by social scientists to investigate the correlates and consequences of young people's educational achievements. Given the widespread political and policy interest in studies such as PISA, such secondary analyses conducted by academics have the potential to generate hugely influential results.

Yet many of the aforementioned international assessment programmes also have ambitious objectives. PISA, for instance, attempts to benchmark 15-year-olds' achievement in three or four academic disciplines (e.g. reading, mathematics, science and collaborative problem-solving) across more than 70 countries.[1] This is despite PISA being a relatively short (two hour) low-stakes test. The way the survey organisers try to achieve this goal, through a complex survey and test design, is poorly understood by many applied researchers who often fail to treat the data as the survey organisers intended.

It is this misunderstanding of these data – particularly amongst economists – which has motivated the need for this paper. We appreciate this in how most studies using PISA published within five influential economics journals – Economics of Education Review, Education Economics, Empirical Economics, Journal of Human Resources and The Economic Journal –have failed to mention (or properly account for) at least one aspect of the survey or test design – this analysis will be provided from authors upon request –. Our aim is to provide a non-technical description of the major international large-scale assessment programmes (e.g. PISA), to clearly articulate what their designs imply for secondary analyses of these data and to provide a case study investigating whether ignoring these features has a substantive impact upon one particularly interesting set of empirical results.

In order to achieve these goals, we replicate a recent study published in *The Economic Journal* by Lavy (2015). This serves as a particularly interesting example, as fairly standard econometric approaches are applied to the PISA data, but with few adjustments made to account for the complex survey and test design. As noted above, we do not believe this to be unusual, with Lavy's (2015) methodology having recently been implemented by others using PISA data within the economics of education field (e.g. Cattaneo, Oggenfuss and Wolter 2016). However, as we will argue, the special features of these data mean that the econometric identification strategies used

[1] Particularly, in 2012, over half a million students participated, representing 28 million 15-year-olds in 65 countries and economies and in 2015, the figure was raised to 72 countries.

by Lavy (2015) and Cattaneo, Oggenfuss and Wolter (2016) should probably not have been applied. In doing so, we hope to generate a better understanding of how international assessments such as PISA are designed and what this subsequently means for secondary analyses of these data.

The paper now proceeds as follows. Section 2 provides a brief overview of Lavy (2015), which serves as our empirical case study. Section 3 then discusses the PISA survey design, including the purpose and use of the different sets of available weights. Section 4 follows with a description of the PISA test, and what this implies for the pupil fixed effects strategy employed by Lavy (2015) and Cattaneo, Oggenfuss and Wolter (2016). Conclusions then follow in section 5.

2 An overview of Lavy (2015)

Published in one of the leading economics journal (*The Economic Journal*), Lavy (2015) investigates whether spending more time learning a subject in school has a positive impact upon pupil's academic performance. Using PISA 2006 data, the author examines how the results compare between a set of developed, developing and Eastern European countries, with the aim of getting as close to a causal effect as possible.

The paper begins by presenting results from a set of basic OLS regression models, comparing how hours spent learning a subject per week in school is related to PISA test scores. These models are of the form:

$$P_{ij} = \alpha + \beta.X_{ij} + \gamma.H_{ij} + \varepsilon_{ij} \qquad \nabla k$$

Where:

P_{ij} = PISA scores of pupil i within school j.

X_{ij} = Basic set of pupils' demographic characteristics.

H_{ij} = Hours spent by pupil i learning a subject in school j per week.

ε_{ij} = The error term, with a Huber-White adjustment made to the estimated standard errors to take the clustering of pupils within schools into account.

i = Pupil i.

j = School j.

∇k = Indicating that separate models are estimated for each of the three PISA subjects.

Then, in a second set of models, the main identification strategy is employed. Pupil fixed effects are added, removing all the between-pupil variation.[2] The focus of these models is therefore pupil's *relative* performance across the different PISA subject areas. In other words, these pupil

[2] The data are now set up so that there are three observations per pupil (one for each of the three PISA subjects: reading, mathematics and science). The pupil fixed effects model includes a dummy variable for each pupil in the dataset, stripping away all the between-pupil information, and leaving only the within-pupil variation.

fixed effects models rely upon within-pupil variation only (e.g. how well pupil's perform in science relative to reading and mathematics) and how this relates to the time they spend learning science versus mathematics in school. Specifically, they are of the form:

$$P_{ik} = \alpha + \gamma.H_{ik} + \mu_i + \varepsilon_{ik}$$

Where:

P_{ij} = PISA scores of pupil i within subject k.

H_{ik} = Hours spent by pupil i learning subject k in school per week.

μ_i = Pupil fixed effects.

ε_{ik} = Random error for pupil i within subject k. A Huber-White adjustment is then made to the estimated standard errors to take the clustering of children within schools into account.

Both the OLS and pupil fixed effects models are estimated using large samples that have been pooled across several countries. This includes a sample of (a) 153,578 pupils from 22 OECD countries; (b) 59,005 pupils from 14 Eastern European countries and (c) 79,646 pupils from 13 developing countries.

Table 1 provides a summary of the key results. The OLS regression models suggest there is a substantial impact of study time upon pupils' PISA scores, with effect sizes ranging between approximately 0.2 (developed countries) and 0.4 standard deviations (developing and Eastern European countries) per additional study hour. However, these are vastly reduced once the pupil fixed effects strategy has been employed, particularly in developing countries, where the impact of an additional hour is only just above zero (0.03 standard deviations). This leads to a headline conclusion that although instruction time has a positive and statistically significant impact upon pupils' PISA academic achievement, the effect is much lower in the developing world.

<< Table 1 >>

Our decision to replicate this particular study is due purely to methodological considerations; we have little argument to make against the key substantive empirical results. Rather, the work of Lavy (2015) serves as an interesting case study as the empirical analysis largely ignores many of the subtle technical aspects of the PISA data; either wittingly or unwittingly, the paper does not follow the recommended practise in PISA data use. For instance, the final student and Balanced-Repeated-Replication (BRR) weights we shall discuss in section 3 have not been applied, while the implications of the complex test design have not been explored. Yet, as noted in the introduction, this empirical approach to the PISA data is not uncommon in the literature – and has been used by others working in this area (e.g. Cattaneo, Oggenfuss and Wolter 2016). Lavy (2015), therefore, provides an opportunity for us to consider what the complex PISA survey and test design implies for different statistical approaches to the PISA data, and how an interesting set of empirical results are affected once these issues have been taken into account.

3 The PISA survey design

PISA aims to draw a representative sample of in-school pupils in each country who are aged between 15 years and three months and 16 years and two months at the time of assessment. However, as with many school-based surveys, PISA is not a simple random sample from the population. Rather, a probabilistic, stratified and clustered survey design is used. In this section, we describe this design, and what it implies for analysis.

To begin, each participating country is divided into a set of mutually-exclusive groups – known as explicit strata. These explicit strata vary across countries, but typically include region and school-type. For instance, within England, the population of schools is divided into four regions (e.g. North, South, Midlands, London) and three different school types (e.g. comprehensive, selective, independent) in PISA 2009 and 2012. Within each of these explicit strata, schools are then ranked by a variable (or set of variables) that are likely to be strongly associated with PISA scores. This is known as implicit stratification, with historic GCSE performance of the school the most important variable used for this purpose in England.[3] Schools are then randomly selected, with probability proportional to size, from within each of these explicit strata. The OECD stipulates that a minimum of 150 schools from each country must participate.[4] Finally, from within each school, a random sample of pupils (usually around 30) is drawn.

Of course, as with any survey, the primary sampling unit (schools) may decline to participate. This has the potential to reduce the representativeness of the sample. PISA and other international surveys attempt to limit the impact of such non-response by allowing countries to approach 'replacement schools' to take the place of non-participating schools in the study. Specifically, for each initially sampled school, two potential replacements are assigned. These are drawn from within the same explicit stratum as the non-participating school, and are chosen so that they are as similar as possible to the school they have replaced in terms of the implicit stratification variables used[5]. The intuition is that these replacement schools will be similar to the originally sampled but non-participating school in terms of the stratifying characteristics. The hope is that, through the use of these replacement schools, any bias due to non-response will be minimised.

The use of replacement schools is not without controversy (Sturgis, Smith and Hughes 2006). Nevertheless, the PISA technical report (e.g. OECD 2009a for PISA 2006) provides full details on response rates before and after replacement schools have been considered. If these response rates fall below a given threshold, then a country may be excluded from the study. The OECD

[3] School gender composition and local education authority area also play a role.
[4] In some very small countries such as Iceland, this effectively means that all schools take part.
[5] These replacement schools are "the schools immediately preceding and following it in the explicit stratum, which was ordered within by the implicit stratification".

has shown this to not be an idle threat; this was the fate that met the Netherlands in 2000, England in 2003 and Malaysia in 2015.

The final important feature of the PISA survey design is that some countries 'oversample' schools and/or pupils. This means that they recruit more schools and/or pupils to participate than is strictly required. These countries then have a much larger sample size; this is often done to facilitate comparisons within these countries at the state/provincial level. Consequently, in Canada, Spain, Italy and Mexico[6], more than 20,000 pupils participated in PISA 2012 (compared to an international median of around 5,000 pupils). In other countries, pupils with certain demographic characteristics may be oversampled. Australia is a prime example, where all Indigenous pupils within selected schools are asked to participate, so that reliable estimates of achievement can be produced for this important minority group.

This complex survey design of PISA, and other international large-scale assessments, has important implications for how the data are analysed by secondary users. We now discuss two particularly important issues: i) the use of sampling weights and ii) methods for adjusting the standard errors to account for how the sample was drawn.

What is the purpose of the PISA *respondent* weights, and what are the implications of not applying these in cross-country analyses?

In the official OECD reports, the PISA results are presented after applying a set a response weights. There are two possible ways to weight the data, known in the literature as:

a) *Final student (or sampling) weights*. These scale the sample up to the size of the population within each country. The contribution of each country to a cross-national analysis (e.g. a cross-country regression model) therefore depends upon its population size (i.e. bigger countries carry more weight).

b) *Senate weights*. These weights sum up to the same constant value within each country. Therefore, within a cross-country regression model, each country will contribute equally to the analysis (e.g. the results for Iceland will have the same impact upon estimates as results for the United States).

One of these sets of weights should almost always be applied when analysing international educational achievement data. If the research question is about the population of pupils living within a specific group of countries (e.g. the population of pupils living within Eastern Europe) then the final sampling weights should be applied. Senate weights are, on the other hand, more appropriate when countries form the unit of analysis; if, for instance, one wants to know the average of a statistic across a set of countries (e.g. the mean PISA science score across the OECD).

[6] For the whole set of countries, Canada represented a 4.5% of the total sample of students, 5.3% in the case of Spain, 6.5% for Italy and 7.0% for Mexico.

Details on the construction of these weights are available within the technical reports (e.g. OECD 2014: chapter 8). In summary, they essentially serve four functions:

- To account for the fact that schools are selected with probability proportional to size.
- To account for the different population sizes in different countries (as noted above).
- To adjust for the oversampling of schools/pupils in certain countries.
- To provide some correction for any remaining non-response.

If weights are not applied, then pupils/schools with particular characteristics may be either under or over represented within the analysis. This will, in turn, potentially lead to biased estimates. For instance, if weights are not applied when analysing the PISA data for Australia, Indigenous students will be overrepresented in the analysis and have an undue influence upon the results.[7] Indeed, it is only after applying these weights that point estimates (i.e. mean scores, regression coefficients) will be 'correct,' meaning that legitimate inferences can be made from the PISA sample about the population.

One feature of Lavy (2015) is that no weights are applied in any part of the analysis. Therefore, by not applying these weights in his pooled cross-country regression models, the statistical contribution of each country to the analysis is essentially arbitrary. Rather than being based upon population size (as with the final student weights) or treating each country equally (as with senate weights) the contribution is based solely upon the size of the sample each country has decided to draw. Our interpretation is that, as Lavy was attempting to make statements about the population of 15-year-olds living within a set of developed/developing/Eastern European nations, the final student weights should have been applied.

Table 2 drives this point home by illustrating the relative importance of each country to the Lavy analysis if (a) no weights; (b) final sampling weights; and (c) senate weights are applied.[8] By not applying weights, too much importance has been given to some countries, while not enough has been given to others. Amongst developed countries, Canada serves as a good example. This is a country which drew a particularly large sample in 2006 – over 22,000 pupils – so that results could be reported separately by province. Consequently, Canada accounts for 12 per cent of Lavy's developed country sample. However, when either the senate or student weights are applied, the contribution of Canada falls to around 5 per cent. Amongst developing countries, the figures for Mexico (another country that oversamples) are even more pronounced. Whereas this country drives around a third of Lavy's developing country estimates, it should only account for around 14 per cent based upon its population size. Finally, for Eastern Europe, the opposite holds

[7] This is due to the oversampling used in this country for this particular sub-group.
[8] Senate weights are simply a re-scaling of the final student weights. They are constructed so that the sum of the weights for each country equals the same constant (typically chosen to be 1,000). As Table 2 illustrates, when senate weights are applied, each country contributes equally to the analysis.

true for Russia. Despite accounting for more than half of Eastern Europe's 15-year-old population, by not applying the sampling weights, Russia's contribution to Lavy's analysis is less than 10 per cent.

<< Table 2 >>

What impact does this have upon the reported OLS regression coefficients?[9] Table 3 reproduces Lavy's results once either the final sampling weights (weighting each country by its population size) or senate weights (weighting the contribution of each country equally) have been applied. Depending upon the choice of weight, there are some non-trivial differences from the reported results. Comparing figures across the first two rows, the estimated effect of an additional hour of instruction within developed countries increases by almost 50 per cent, up from 0.196 standard deviations when applying no weights to 0.276 standard deviations when applying the final sampling weights; moreover, the standard error has doubled (up to 0.014 from 0.007). In contrast, the effect size has almost halved for Eastern Europe, declining from 0.382 to 0.230 standard deviations. The developing country estimates have also fallen, but the change is less pronounced (fall from 0.366 to 0.325). When using senate weights, the effect size is similar to that of Lavy's, but with larger standard errors. Together, Table 3 highlights how important changes to parameter estimates and their standard errors can occur depending upon whether weights are applied within cross-country regressions or not.

<< Table 3 >>

<u>What are the purpose of the PISA *replication* weights, and what are the implications of not applying these in cross-country analyses?</u>

Although the importance of accounting for multi-level structures is widely appreciated across the social sciences, either via estimation of multi-level models or via Huber-White adjustments to estimated standard errors, others issues (such as accounting for stratification or for the use of replacement schools) are less widely understood. Moreover, although some statistics packages such as R and Stata include commands to adjust standard errors for stratification, this requires information identifying the strata within the dataset.[10] Unfortunately, for confidentiality reasons, this information is not typically available for all countries within international education databases such as PISA (e.g. information about strata are not available for China, Austria and Germany, amongst others, in PISA 2015).

The way PISA and other international studies resolve this issue is by providing a series of 'replicate weights'. These are based upon a re-sampling methodology, and work in a similar way

[9] We focus upon the OLS regression results here, as issues with the pupil fixed effects strategy will be covered in section 4 below.

[10] For instance, the '*svyset*' Stata command includes the 'strata' option where this element of the survey design can be taken into account.

to jack-knife and bootstrapping techniques. It is only through the application of these weights that secondary analysts can fully account for all elements of the complex PISA survey design within every participating country, and thus replicate the 'official' figures reported by the OECD. Although the major international surveys use slightly different variants of these replication procedures, most can be handled within standard statistical software packages, including Stata and R.[11] If these replication weights are ignored, then secondary analysts risk over or under estimating the amount of uncertainty (due to sampling error) in their results.

To account for the clustered nature of the PISA data (i.e. pupils nested within schools) Lavy (2015) followed standard practise in the economics literature, and applied a Huber-White adjustment to the estimated standard errors (i.e. standard errors were 'clustered' at the school level). However, as noted above, such adjustments do not take into account some features of the PISA survey design, such as the use of stratification and the use of 'replacement' schools. To what extent would doing so, via application of the replicate weights (here BRR weights), make an appreciable difference to the reported results?

<< Table 4 >>

Table 4 provides the answer, with our particular focus upon rows 2 and 3. There is, as expected, no change to the point estimates; the issue we are dealing with here only affects the standard errors. However, even the standard errors are quite similar across rows 2 and 3. For instance, the standard error for the developed country estimates declines from 0.14 (when applying final student weights with a Huber-White adjustment) to 0.11 (when applying final student weights and the BRR weights). This more generally reflects our experience in using international achievement datasets such as PISA. Taking the clustering of pupils in schools into account is clearly important when estimating the standard error; however, whether one simply applies a Huber-White adjustment (as per Lavy) or follows the recommended replication-weight procedure typically has relatively little impact upon the key substantive results.

4 The PISA test design

PISA is not a standard test; rather it has a complex psychometric design. A key feature is the use of 'multiple matrix sampling' (MMS), with the intuition behind this as follows. International assessments such as PISA attempt to measure pupils' skills in a number of different subject areas (reading, mathematics, science, problem solving, financial literacy) and within these a number of different sub-domains (e.g. 'explaining phenomena scientifically', 'identifying scientific issues' and 'using scientific evidence' in science). This results in a huge amount of test material to be

[11] For instance, Stata includes the option '*brrweight*' within the '*svy*' command. Moreover, a number of user written commands to handle this feature of international databases now exist, including the excellent Stata and R packages of Avvisati and Keslair (2014) and Caro (2016).

covered – up to 10 hours per subject – making it impossible to ask every pupil each test question. Consequently, in order to keep the length of the PISA test manageable (e.g. to two hours), participants are ***randomly assigned*** to complete one particular test booklet, each of which includes only a limited number of test questions.

Table 5 illustrates how this worked in practice in PISA 2006. In total, 108 science questions, 31 reading questions and 48 mathematics questions were included in the assessment framework.[12] These questions were then divided into seven science, four mathematics and two reading clusters (a cluster refers to a collection of test questions), each covering 30 minutes of test material. These clusters are labelled S1-S7, M1-M4 and R1-R2, respectively, in Table 5. Out of these clusters, a total of 13 test booklets were formed (labelled B1-B13). Note that some of these booklets included only science questions (e.g. booklets 1 and 5), while others included questions in only science and reading (e.g. booklet 6) or only science and mathematics (e.g. booklets 3, 4, 8 and 10). Within each participating school, pupils were randomly assigned to one of these 13 booklets.

<< Table 5 >>

Based upon pupils' responses to the test questions they were randomly assigned, the survey organisers fit a complex item-response theory (IRT) model to the data. This involves estimating a set of random-effects logistic regression models, where test questions are nested within participating students (with their answers – mostly binary coded as 1 for correct and 0 for incorrect – as the dependent variable). Based upon this model, the difficulty of each test question is established and 'test scores' (or, more appropriately, proficiency estimates) for participants are produced. Describing the technical details behind this process is beyond the scope of this paper, with interested readers directed to von Davier and Sinharay (2014:157 and 161) for further details.

The result of this process is the creation of the international PISA database. This is published online, free for researchers to use from the OECD website.[13] Within the international database what appears to be five separate test scores for each individual in each subject area can be found. To illustrate this point, an extract from this database is presented in Table 6, referring to a set of pupils who completed test booklet 1 in PISA 2006.

< Table 6 >

At this point, readers may be forgiven for suffering some confusion. Why are there *five* mathematics test scores for each pupil rather than just one? And why do pupils who have not answered any reading test questions seem to have a reading test score? (i.e. why do the pupils in Table 6 who all completed test booklet 1 – and therefore only answered science test questions – also have scores in reading and mathematics)?

[12] One subject area is the focus in each cycle of PISA. In 2006, the focus was science, hence there were many more questions devoted to this subject than either reading or mathematics.
[13] See https://www.oecd.org/pisa/pisaproducts/ and http://www.oecd.org/pisa/data/

The answer is that international assessments such as PISA rely heavily upon multiple imputation. The intuition is as follows. As illustrated in Table 5, pupils answer only a limited number of questions from the total test item pool. Those questions they do not answer can be thought of as a form of 'missing data' (or item non-response). However, as pupils have been randomly assigned to test booklets, and thus to test questions, the missing data for the questions they have not been asked to answer can be considered to be Missing Completely At Random (MCAR). Consequently, multiple imputation can be used to 'fill-in' the missing information. The argument is that under an MCAR assumption the use of multiple imputation will raise efficiency (i.e. reduce standard errors) but not have any direct effect upon the estimate of pupil's proficiency scores.

The key take away message is therefore that the five PISA 'test scores' (known in the psychometric literature as 'plausible values') are essentially multiple imputations based upon (a) pupils' answers to the sub-set of test questions they were randomly assigned and (b) their responses to the background questionnaires. It is for this reason that the PISA database includes test scores ('plausible values') in reading even for pupils who did not actually answer any reading test questions.

What are the implications of this for secondary analyses of the PISA data?

How does one 'correctly' use these plausible values? The answer, according to the survey organisers, is that one should follow a version of 'Rubin's rules' for handling multiple imputations (see OECD 2009a; Rubin 1987). This procedure can be divided into four steps:

Step 1: Estimate the statistic/model of interest five times, once using each of the plausible values. This will generate five separate parameter estimates (β_{pv}) and five estimates of the sampling error (σ_{pv}).[14]

Step 2: To produce the final parameter and ***sampling*** error estimates, one simply takes the average of the five estimates produced in step 1:

$$\beta_* = \frac{\sum_{pv=1}^{5} \beta_{pv}}{n_{pv}}$$

$$\sigma_* = \frac{\sum_{pv=1}^{5} \sigma_{pv}}{n_{pv}}$$

Where: β_*= Final estimate of the statistic / parameter of interest

σ_* = Final estimate of the *sampling* error

n_{pv} = The number of plausible values (typically five)

Step 3: Estimate the magnitude of the ***imputation*** error, based upon the following formula:

[14] Note that the BRR weights described in the previous section should also be applied each of the five times the model is estimated.

$$\delta_* = \frac{\sum_{pv=1}^{5}(\beta_{pv} - \beta_*)^2}{n_{pv} - 1}$$

Where:

δ_* = The magnitude of the imputation error.

Step 4: Calculate the value of the final standard error by combining the sampling error (σ_*) and the imputation error (δ_*) via the following formula:

$$\text{Standard error} = \sqrt{\sigma_*^2 + \left(1 + \frac{1}{PV}\right).\delta_*^2}$$

One can then use the final parameter estimate (β_*) and its standard error to conduct hypothesis tests and construct confidence intervals following the usual methods.

Rather than following the steps outlined above, Lavy only uses the first imputed value throughout his analysis. Does this make a difference to his results? One can find the answer by returning to the bottom two rows of Table 4. The impact appears to be minimal, with only trivial changes to the estimated effect sizes and associated standard errors. Whether one uses just one plausible value, or follows recommended practise in using all five, has no substantive impact upon the results.

Although it can be dangerous to draw strong conclusions from a single analysis, this result again reflects our experience more broadly of using international achievement databases (and the PISA data in particular). Whether one uses just a single plausible value or closely follows the recommended procedure typically has a trivial impact upon substantive results. Indeed, the survey organisers themselves recognise that the use of a single plausible value actually provides both unbiased point and sampling variance estimates, stating how '*using one plausible value or five plausible values does not really make a substantial difference on large samples*' (OECD 2009b:46). The only aspect that using a single plausible misses is the 'imputation error' – uncertainty that should be added to the standard error to reflect the fact that multiple imputation is used to generate the science, reading and mathematics proficiency scores. Yet, in practise, this additional imputation error is almost always of negligible magnitude (as per the Lavy example), with key conclusions continuing to hold if it is simply ignored.

However, the fact that PISA scores are essentially imputations does raise other concerns regarding how these data should and should not be used. This includes the application of some fairly standard econometric procedures, such as the use of pupil fixed effects. To see why, recall the PISA 2006 test design presented in Table 5, and how pupils are randomly allocated to one of these 13 booklets. Moreover some pupils, like those assigned booklet 1, answer science test questions only, and none in reading or mathematics.

Now recall what a pupil fixed effects methodology is trying to achieve. It strips away all the between-pupil differences, so that only within-pupil variation in achievement is left to explain. For example, in Lavy (2015), the pupil fixed effects models essentially compare each pupil's own performance in science relative to their performance in reading and mathematics, relating this to the relative amount of time he/she spends attending classes in each subject per week. However, as noted above, pupils' 'test scores' (plausible values) are imputed, based upon how they answered a small number of test questions (sometimes just within a single subject area – e.g. just science questions in the case of pupils assigned booklet 1) and the information they provided in the background questionnaire.[15] In such a situation, any within-pupil variation in performance that exists across subjects is largely generated by the imputation procedure. Indeed, conceptually, it is impossible to accurately capture within-pupil variation in performance across different academic domains (e.g. relative performance in science compared to reading and mathematics), when many pupils have actually only answered questions in a single subject area (e.g. science).

This leads to a much more general point about international large-scale assessment data such as PISA. These tests have been designed to provide summary statistics about the population of interest within each country (OECD 2009a:156), and about simple correlations between key variables (e.g. between socio-economic status and pupil performance), but '*plausible values contain random error variance components and are not optimal as scores for individuals*' (OECD 2009a:156). In this sense, the psychometricians behind these tests warn how '*reliable individual proficiency estimates cannot be obtained'* (Oranje and Ye 2014:204), that they '*are not intended to produce and disseminate individual results at the respondent or even the classroom or school level'* (Oranje and Ye 2014:204) and that they '*lack accuracy on the individual test-taker*' (von Davier and Sinharay 2014:156). In other words, measurement error is so large at the individual level that test scores for individual pupils are unreliable. Consequently, even for those pupils who have actually taken test questions in all three of the PISA subjects (e.g. those allocated to booklet 13 in PISA 2006 – recall Table 5) the use of pupil fixed effects models is not advised.

Are there any wider implications of such test designs?

There are other methodological implications of such test designs beyond those we have discussed in relation to Lavy (2015). Although we are unable to cover all of these within this paper, we will highlight one important issue, related to the increasingly common practise of tracking PISA cohorts over time. This has been done in a number of countries by either re-surveying participating pupils at a later date (as is the case in Australia, Canada, Denmark and Switzerland[16])

[15] The information captured in the background questionnaires include demographic data and pupils' attitudes.

[16] See, for instance, http://www.lsay.edu.au/lsay-data/scope for Australia, http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4435 for Canada and

or by linking PISA to administrative data (as is the case in England). A major focus of research based upon these studies is to how performance in PISA correlates with other test score measures (e.g. GCSE grades in England) and outcomes in later life (e.g. access to university, labour market earnings).

If PISA had a 'standard' test design, this would not be an issue. However, the fact that PISA uses multiple imputation to generate the plausible values again potentially causes problems. This is because, within the multiple imputation literature, it is widely considered best practise to have a 'congenial' imputation model (Carpenter and Kenward 2013:70). This means that all variables included in the final substantive analysis should be included in the imputation model. If this is not the case, then estimated relationships between the imputed variable(s) and the variables *not* included in the imputation model will be biased (von Davier 2014). As the PISA plausible values are essentially multiple imputations, the same rules apply here as well. However, 'congenial' imputation models cannot currently be used for secondary analysis involving longitudinal follow-ups of the PISA data, since the models used to create PISA plausible values are not made publicly available. What this then implies is that any comparison made between PISA scores and national examination data (e.g. GCSE grades in England) or with young people's later lifetime outcomes (e.g. whether they enter university, their pay) are likely to suffer biases of unknown direction and magnitude.[17]

We believe this final fact highlights perhaps the critical point we have tried to make throughout this paper: the peculiar nature of PISA's survey and test design adds many additional complications to secondary analyses of these data. Some of these complications may be discussed by the survey organisers, and understood by some analysts, but others are not. Much more work is therefore needed to bring clarity and transparency to how these datasets are constructed, what this implies for how they should and should not be used, and the potential bias that could be introduced into secondary results.

5 Conclusions

International studies of educational achievement are becoming increasingly high-profile resources, with secondary analyses of these data having the potential to influence education policy and practise across the world. Yet the complex survey and test designs used remain poorly

http://forscenter.ch/en/data-and-research-information-services/2221-2/special-projects/tree/ for Switzerland.

[17] It has been suggested that the correlation between the imputed variable (e.g. PISA plausible values) and the variable(s) not included in the imputation model (e.g. future examination grades, university entry, future earnings) will typically be downward biased. However, we know of no research investigating this issue with respect to international large scale assessment test designs. Likewise, we believe the direction of the bias will be increasingly difficult to know when one moves beyond simple bivariate associations between the linked data and imputed plausible values.

understood by many consumers of these data. This not only includes politicians, policymakers and the general public who digest the results, but also by academics who analyse the data to produce secondary research. Resources such as PISA are consequently often being analysed in a manner not intended by the survey organisers, potentially leading to erroneous conclusions and biased results. The aim of this paper has therefore been to foster a better understanding of the complex features of international large-scale assessments, particularly amongst economists, who now frequently use these resources in their work.

Using Lavy (2015) as a case study, we have provided an overview of the survey methodology underpinning studies such as PISA, highlighting the importance of applying the survey weights when conducting cross-country analyses using pooled international samples. Likewise, several unusual features of the PISA test design have been explored, including the use of multiple matrix sampling and the resulting imputations of pupils' proficiency scores ('plausible values'). In doing so, we have highlighted how some fairly standard econometric approaches (such as the use of pupil fixed effects) should not be applied to these data, and that the statistical techniques required to robustly analyse these resources are perhaps more complicated than first meets the eye.

What do these findings then imply for the users, producers and consumers of these data? We offer two suggestions. First, more clarity and greater transparency is needed from the survey organisers about the test design, and exactly how the proficiency values (i.e. the 'PISA scores') are produced. Indeed, the imputation models used to generate the so-called plausible values remain a black-box. Most people (including many highly-skilled academics) do not understand the fact that PISA scores are actually imputations and, consequently, what statistical methodologies are and are not appropriate to apply. Although some of the relevant details are available in the depths of the technical reports, we believe a more open, transparent and widespread discussion of the methodologies underpinning these studies would be hugely beneficial. This, we believe, is key to getting a broader cross-section of researchers to understand what these data can and cannot reveal, and how much faith should be placed upon the results. Our suggestion is that the code to reproduce the imputation models, allowing independent researchers to see how the plausible values are derived from the underlying data, represents a first critical step in this direction.

Second, at the same time, it is also the responsibility of users of these resources to develop a better understanding of the properties of the data. Various technical reports and user guides now exist, with many of the key details included within (e.g. OECD 2009b). Applied researchers should also take more advantage of the many excellent software plugins for analysing these datasets now available for standard statistical packages such as R and Stata (Avvisati and Keslair 2014; Caro 2016), which greatly reduce the computational burden. Moreover, despite the limitations and complications we have highlighted with these data, we continue to believe they are a useful and valuable source of secondary data.

In highlighting these points, we hope to have improved the transparency of the methodology behind international large-scale education achievement surveys, the care that needs to be taken when analysing these data and the caveats that are required when interpreting the results. Although we continue to see the value in international studies of educational achievement such as PISA, and their potential to influence education policy for the better, we also feel that far more scrutiny needs to be given to the unusual features of their design. This, we believe, will only help people to better understand what can and cannot be done with the data, and for more nuanced interpretations to be placed upon the PISA results.

References

Avvisati, F. and Keslair, F. (2014). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values*. Statistical Software Components S457918, Boston College Department of Economics.

Caro, D. (2016). *Package 'intsvy': International Assessment Data Manager*. Accessed 18/01/2017 from https://cran.r-project.org/web/packages/intsvy/intsvy.pdf

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Application*. Chichester: Wiley.

Cattaneo, M. A., Oggenfuss, C., and Wolter, S. C. (2016). *The More, the Better? The Impact of Instructional Time on Student Performance*. IZA DP No. 9797.

Lavy, V. (2015). Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries. *The Economic Journal 125*, F397–F424. http://doi.org/10.1111/ecoj.12233

OECD (2009a). *PISA 2006 Technical Report*. OECD Publishing.

OECD (2009b). *PISA Data Analysis Manual: SPSS, Second Edition*. OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*. OECD Publishing.

Oranje, A. and Ye, L. (2014). Population Model Size, Bias, and Variance in Educational Survey Assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 203–228). Boca Raton: CRC Press.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rutkowski, L., Gonzalez, E., Joncas, M., and von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher 39*(2), 142–151. http://doi.org/10.3102/0013189X10363170

Sturgis, P., Smith, P., and Hughes, G. (2006). *A study of suitable methods for raising response rates in school surveys*. Research Report No 721. Department for Education and Skills, London.

von Davier, M. (2014). Imputing Proficiency Data under Planned Missingness in Population Models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 175–201). Boca Raton: CRC Press.

von Davier, M. and Sinharay, S. (2014). Analytics in International Large-Scale Assessments: Item Response Theory and Population Models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 155–174). Boca Raton: CRC Press.

Table 1. An overview of key results from Lavy (2015).

	OLS		**Pupil fixed effects**	
	Effect size	**SE**	**Effect size**	**SE**
OECD sample	0.196*	0.007	0.058*	0.004
Developing country sample	0.366*	0.012	0.030*	0.008
Eastern European sample	0.382*	0.013	0.061*	0.006

Notes: Source is Lavy (2015) Table 3 and Table 8. Results refer to the estimated impact of a one hour increase in instructional time upon pupils' PISA test scores, reported as an effect size. * indicates significantly different from zero at the one per cent level.

Table 2. The role of weights in determining countries' importance in pooled cross-country analyses.

(a) Developed countries			
	No weight	**Senate weight**	**Student weight**
Canada	12%	5%	6%
Italy	12%	5%	9%
Spain	11%	5%	6%
Australia	8%	5%	4%
UK	7%	5%	12%
Switzerland	7%	5%	1%
Belgium	5%	5%	2%
Japan	3%	5%	18%
Portugal	3%	5%	1%
Austria	3%	5%	1%
Germany	3%	5%	15%
Greece	3%	5%	2%
Netherlands	3%	5%	3%
New Zealand	3%	5%	1%
Finland	3%	5%	1%
France	3%	5%	12%
Norway	3%	5%	1%
Ireland	2%	5%	1%
Luxembourg	2%	5%	0%
Denmark	2%	5%	1%
Sweden	2%	5%	2%
Iceland	2%	5%	0%
Total	**100**%	**100**%	**100**%

(b) Developing countries			
	No weight	**Senate weight**	**Student weight**
Mexico	30%	8%	14%
Indonesia	10%	8%	27%
Brazil	9%	8%	22%
Jordan	6%	8%	1%
Thailand	6%	8%	8%
Kyrgyzstan	6%	8%	1%
Chile	5%	8%	3%
Azerbaijan	5%	8%	1%
Turkey	5%	8%	8%
Uruguay	5%	8%	0%
Tunisia	5%	8%	2%
Columbia	4%	8%	6%
Argentina	4%	8%	6%
Total	**100**%	**100**%	**100**%

(c) Eastern European countries			
	No weight	**Senate weight**	**Student weight**
Slovenia	9%	7%	1%
Czech Republic	8%	7%	4%
Russian Federation	8%	7%	57%
Poland	8%	7%	16%
Croatia	7%	7%	1%
Romania	7%	7%	7%
Estonia	7%	7%	1%
Serbia	7%	7%	2%
Lithuania	7%	7%	2%
Slovak Republic	7%	7%	2%
Latvia	7%	7%	1%
Bulgaria	6%	7%	2%
Hungary	6%	7%	3%
Montenegro	6%	7%	0%
Total	**100**%	**100**%	**100**%

Table 3. Changes to Lavy's OLS estimates when the PISA weights are applied.

	OECD		Developing		Eastern Europe	
	Effect size	**SE**	**Effect size**	**SE**	**Effect size**	**SE**
No weights (Lavy 2015)	0.196*	0.007	0.366*	0.012	0.382*	0.013
Final student weights	0.276* (+41%)	0.014	0.325* (-11%)	0.019	0.230* (-40%)	0.014
Senate weights	0.188* (-4%)	0.010	0.340* (-7%)	0.018	0.362* (-5%)	0.015

Notes: 'Final student weights' equivalent to weighting by the population size of the country, while 'senate weights' give equal weights to all countries, regardless of size. * indicates significantly different from zero at the one per cent level.

Table 4. Changes to Lavy's OLS estimates if weights and plausible values are applied.

	Developed		Developing		Eastern Europe	
	Effect size	**SE**	**Effect size**	**SE**	**Effect size**	**SE**
Lavy (2015)	0.196*	0.007	0.366*	0.012	0.382*	0.013
+ final student weights	0.276*	0.014	0.325*	0.019	0.230*	0.014
+ BRR weights	0.276*	0.011	0.325*	0.016	0.230*	0.016
+ plausible values	0.277*	0.012	0.327*	0.017	0.230*	0.016

Notes: Top row refers to the results presented by Lavy (2015) where no weights are applied, a Huber-White adjustment has been made to the estimated standard errors and only the first plausible value is used. Results in the second row replicate the Lavy analysis, but now applying the final student weights. The third row uses the BRR weights to account for the complex PISA survey design, rather than making a Huber-White adjustment. In the final row, all five plausible values have been used, following recommended practise by the OECD.

* indicates significantly different from zero at the one per cent level.

Table 5. The PISA 2006 test design.

Booklet	**Clusters**			
1	S1	S2	S4	S7
2	S2	S3	M3	R1
3	S3	S4	M4	M1
4	S4	M3	S5	M2
5	S5	S6	S7	S3
6	S6	R2	R1	S4
7	S7	R1	M2	M4
8	M1	M2	S2	S6
9	M2	S1	S3	R2
10	M3	M4	S6	S1
11	M4	S5	R2	S2
12	R1	M1	S1	S5
13	R2	S7	M1	M3

Notes: OECD (2009a:29) PISA 2006 technical report. S1 to S7 refers to the seven science clusters (white shading), M1 to M4 the four mathematics clusters (light grey shading) and R1 to R2 the two reading clusters (dark grey shading).

Table 6. An extract illustrating the 'plausible values' within the PISA database.

			Reading					Mathematics					Science				
Country	**School id**	**Student id**	**PV1**	**PV2**	**PV3**	**PV4**	**PV5**	**PV1**	**PV2**	**PV3**	**PV4**	**PV5**	**PV1**	**PV2**	**PV3**	**PV4**	**PV5**
Argentina	1	10	410	329	394	348	371	349	309	359	394	389	330	279	326	310	362
Australia	1	4	444	448	439	490	448	454	477	460	489	513	483	473	472	456	526
Austria	1	26	604	668	664	664	669	623	729	697	697	655	647	705	692	692	699
Azerbaijan	1	2	455	520	370	445	436	535	540	526	514	521	509	541	491	514	486
Belgium	1	13	427	380	386	363	351	448	366	456	458	451	434	379	416	434	420
Bulgaria	2	5	572	572	484	460	484	408	408	403	491	403	433	433	374	417	374
Brazil	2	12	386	372	325	342	299	324	337	358	357	341	370	379	377	333	352
Canada	1	5	492	478	469	535	551	489	486	520	506	573	473	477	485	484	499
Switzerland	1	6	442	501	469	408	448	478	439	453	432	475	471	508	473	456	515
Chile	1	3	591	613	498	613	478	454	475	457	475	434	554	553	533	553	548

Notes: Extract from the PISA (2006) database. 'PV' stands for plausible value.