What happens when econometrics and psychometrics collide? An example using the PISA data

John Jerrim[1]

Luis Alejandro Lopez-Agudo[2]

Oscar D. Marcenaro-Gutierrez[2]

Nikki Shure[3]

September 2017

http://johnjerrim.com/papers/

International large-scale assessments such as PISA are increasingly being used to benchmark the academic performance of young people across the world. Yet many of the technicalities underpinning these datasets are misunderstood by applied researchers, who sometimes fail to take their complex sample and test designs into account. The aim of this paper is to generate a better understanding amongst economists about how such databases are created, and what this implies for the empirical methodologies one should (or should not) apply. We explain how some of the modelling strategies preferred by economists seem to be at odds with the complex test design, and provide clear advice on the types of robustness tests that are therefore needed when analyzing these datasets. In doing so, we hope to generate a better understanding of international large-scale education databases, and promote better practice in their use.

Key Words: sample design; test design; PISA; weights; replicate weights; plausible values.

JEL Codes: I20, C18, C10, C55.

Acknowledgements: This work has been partly supported by the Andalusian Regional Ministry of Innovation, Science and Enterprise (PAI group SEJ-532 and Excellence research group SEJ-2727); by the Spanish Ministry of Economy and Competitiveness (Research Project ECO2014-56397-P) and scholarship FPU2014 04518 of the Ministry of Education, Culture and Sports [*Ministerio de Educación, Cultura y Deporte*]. This work was developed during a PhD visiting research internship at the Institute of Education of University College London (UCL) funded by Ministerio de Educación, Cultura y Deporte for FPU2014 04518 (2016). We also acknowledge the training received from the University of Malaga PhD Program in Economy and Business [*Programa de Doctorado en Economía y Empresa de la Universidad de Malaga*].

[1] Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way London, WC1H 0AL. E-mail: j.jerrim@ucl.ac.uk (John Jerrim).

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

[3] Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way London, WC1H 0AL and the Institute of Labor Economics (IZA). E-mail: nikki.shure@ucl.ac.uk (Nikki Shure)

1. Introduction

International assessment programs have received much attention over the last two decades, with academics, journalists and public policymakers all eagerly awaiting every set of new 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 these studies, such secondary analyses have the potential to generate hugely influential results.

Many of the aforementioned international assessment programs 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. This is despite PISA being a relatively short (two hour), low-stakes test. The way the survey organizers try to achieve this goal, through a complex sample and test design, is poorly understood by many applied researchers who often fail to treat the data as the survey organizers intended.

It is this misunderstanding of these data – particularly amongst economists – which has motivated the need for this paper. We highlight this point in Appendix A (available in the online materials), illustrating how most studies using PISA published in five influential economics journals have failed to mention (or properly account for) at least one aspect of the sample or test design. Our aim is to provide a non-technical description of the major international large-scale assessment programs (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).[4] This serves as a particularly interesting example, as fairly standard econometric approaches – OLS and pupil fixed-effects – are applied to the PISA data, but with few adjustments made to account for the complex sample and test design. As noted above and illustrated in Appendix A, we do not believe this to be unusual. Indeed, others in the economics of education field have used similar methods (e.g. Rivkin and Schiman 2015, Hanushek, Piopiunik and Wiederhold 2014, and Cattaneo, Oggenfuss and Wolter 2017). Although the substantive conclusions these papers reach may or may not be undermined, we nevertheless argue that the special features of the PISA data mean that the common econometric identification strategies used in these papers should have been through a series of important additional robustness tests (which we shall describe in section 5 of this paper). 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 an overview of the Lavy (2015) study. Section 3 then discusses the PISA sample 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). We then provide our recommendations for researchers who wish to apply fixed-effects within international achievement datasets such as PISA in section 5. Conclusions follow in section 6.

2. The Lavy (2015) study

We decided to replicate Lavy (2015), published in a leading economics journal (*The Economic Journal*), purely due to methodological considerations; we have little argument to make against the key substantive results. Rather, the work of Lavy (2015) serves as an interesting case study as the empirical analysis does not make any adjustment for many of the subtle technical aspects of the PISA data. For instance, the final student weights we 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 increasingly being used in the literature – and has been applied by others working in this area (e.g. Rivkin and Schiman 2015; Hanushek, Piopiunik and Wiederhold 2014; and Cattaneo, Oggenfuss and Wolter 2017).

[4] The syntax and data provided by *The Economic Journal* to replicate Lavy (2015) is publicly available in the "Supporting Information" section at http://onlinelibrary.wiley.com/doi/10.1111/ecoj.12233/abstract, which allows us to exactly reproduce Lavy's (2015) published results.

Additionally, to the extent that the syntax and data used by Lavy (2015) are publicly available, this paper provides an opportunity to consider what the complex PISA sample and test design implies for applying different estimation strategies to the PISA data, and how an interesting set of empirical results are affected once these issues have been taken into account.

Specifically, Lavy (2015) investigates whether spending more time learning a subject in school has a positive impact upon a 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:

$$P_{ij} = \alpha + \beta.X_{ij} + \gamma.H_{ij} + \varepsilon_{ij} \qquad \nabla k \tag{1}$$

Where:

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

X_{ij} = Basic set of pupil's 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. This means that the data are 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. The identification strategy relies on the assumption that β and γ are not indexed by k – see further discussion on this assumption in Lavy (2015, pp. F401-F402). The

focus of these models is therefore a pupil's *relative* performance across the different PISA subject areas. In other words, these pupil fixed-effects models rely upon within-pupil variation only (e.g. how well a pupil performs in science relative to reading and mathematics) and how this relates to the time they spend learning science versus reading (and mathematics) in school. Specifically, they are of the form:

$$P_{ik} = \alpha + \gamma . H_{ik} + \mu_i + \varepsilon_{ik} \quad (2)$$

Where:

P_{ik} = 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 pupils 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 estimations of the equation (2) model by OLS suggest there is a substantial impact of instruction 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 achievement, the effect is much lower in the developing world.

<< Table 1 >>

3. The PISA sample design and the use of weights

This section summarises information contained in PISA technical documents (OECD 2009a, 2009b). PISA aims to draw a representative sample of in-school pupils in each country who are age 15 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 sample design is used.[5] One of the key features of this design is that in some countries schools and/or pupils are oversampled (this is often done to facilitate comparisons within these countries at the state/provincial level). These countries then have a much larger sample size; in Canada, Spain, Italy and Mexico 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.

Consequently, the PISA dataset comes with two sets of weights. These are:

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.[6] 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 usually be applied when analyzing international educational achievement data, particularly when reporting descriptive statistics or running regression models with a limited number of controls (an exception is that, if all factors used to create the weights are included as covariates within the regression model, then the weights no longer need

[5] We do not discuss here issues with regards the replication weights that are provided with the international achievement datasets, and how these should be used to calculate standard errors. Interested readers are directed to the working paper version of this publication, available from http://repec.ioe.ac.uk/REPEc/pdf/qsswp1704.pdf

[6] 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).

to be applied[7]). 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). If weights are not applied, then pupils/schools with particular characteristics may be either under or over represented within the analysis. 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 (including the descriptive statistics). 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.

Table 2 drives this point home by illustrating the relative importance of each country to the Lavy analysis if (a) no weights; (b) senate weights; and (c) final sampling 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 percent of Lavy's developed country sample. However, when either student weights or senate weights are applied, the contribution of Canada falls to around five-six percent. 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 percent based upon its population size. Finally, for Eastern Europe, the opposite holds true for Russia. Despite accounting for more than half

[7] As Solon, Haider and Wooldridge (2015, p. 310) note, a: "*practical example is where the survey organization provides sampling weights to adjust for differential nonresponse, including attrition from a panel survey, and these weights are based only on observable characteristics that are controlled for in the regression model (perhaps gender, race, age, location). In this situation, it is not clear that there is an advantage to using such weights*". See also Cook and Gelman (2006).

[8] As Table 2 illustrates, when senate weights are applied, each country contributes equally to the analysis.

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 percent.

<< 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 or senate weights 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 percent, 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 >>

Is this just an issue in cross-country analyses? Or does the decision to apply weights or not also have an impact upon within single country estimates? In online Appendix B we illustrate how Lavy's OLS regression estimates would change for three specific countries (Canada, Mexico and Russia) depending upon whether weights are applied. There are again some non-trivial differences, at least for Canada and Mexico, with the coefficient of interest (the impact of the number of hours studied) up to 26 percent lower once the final student weights have been applied.

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 and financial literacy)

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

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

Table 4 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.[10] 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 4. 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 B1 and B5), while others included questions in only science and reading (e.g. booklet B6) or only science and mathematics (e.g. booklets B3, B4, B8 and B10). Within each participating school, pupils were randomly assigned to one of these 13 booklets.

<< Table 4 >>

Based upon pupils' responses to the test questions, the survey organizers 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. 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, though an overview is provided in online Appendix C, with interested readers directed to von Davier and Sinharay (2014, p. 157 and p. 160 for further details). For a comprehensive overview of research on the measurement of student ability see Jacob and Rothstein (2016).

The result of this process is the creation of the international PISA database. Within the international database, there appears to be five separate test scores for each individual in each subject area. To illustrate this point, an extract from this database is presented in online

[10] 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.

Appendix D, referring to a set of pupils who completed test booklet B1 in PISA 2006 – a booklet that contains four clusters of science.

At this point, readers may be forgiven for suffering some confusion. Why are there *five* test scores per subject for each pupil rather than just one? And why do pupils who have not answered any reading or mathematics test questions seem to have a reading and mathematics test score (i.e. why do the pupils in Table 4 who all completed test booklet B1 – and therefore only answered science test questions – also have scores in reading and mathematics)?

The intuition is as follows. As illustrated in Table 4, 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 is used to create test scores for each pupil in each subject area regardless of whether they have answered questions in that particular cognitive domain or not.

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 subset of test questions they were randomly assigned (b) their responses to the background questionnaires and (c) school dummy variables. It is for this reason that the PISA database includes test scores ('plausible values') in, e.g. 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 is that one should follow a version of 'Rubin's rules' for handling multiple imputations (Rubin 1987). Further details are provided in OECD (2009a) and in online Appendix E.

Rather than using all five plausible values as recommended by the survey organizers (see online Appendix E), Lavy only uses the first imputed value throughout his analysis. Does this make a difference to his results? The answer may be found in Table 5. The impact appears to be minimal, with only trivial changes to the estimated effect sizes and associated standard errors. 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 one plausible value, or follows recommended practice in using all five, has no impact upon the results.[11]

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 fixed-effects. To see why, recall the PISA 2006 test design presented in Table 4, 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 her performance in reading and mathematics, relating this to the relative amount of time 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), the information they provided in the background questionnaire and school dummy variables.[12] In such a situation, the within-pupil variation in performance that exists across subjects is largely generated by the imputation procedure. Indeed, conceptually, it is not reliable to capture within-pupil variation in performance across different academic domains (e.g. relative performance in science compared to reading and mathematics) when some pupils have actually only answered questions in a single subject area (e.g. science). Moreover, because H_{ij} is included as one of the hundreds of regressors used to impute the outcome (the PISA plausible values) but is also the covariate of interest within the substantive model, endogeneity may potentially become a concern.

Alternatively, one could argue that the use of pupil fixed-effects violates the often cited principle within the multiple imputation literature for how imputation models should be built.

[11] Indeed, the survey organizers themselves recognize that the use of a single plausible value actually provides both unbiased point and sampling variance estimates, stating that '*using one plausible value or five plausible values does not really make a substantial difference on large samples*' (OECD 2009b, p. 46). The only aspect that using a single plausible value misses is the 'imputation error' – uncertainty that should be added to the standard error to reflect that multiple imputation is used to generate the science, reading and mathematics proficiency scores. Yet, in practice, 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.

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

Namely that all variables included in the substantive model should also be included in the imputation model as well (see Carpenter and Kenward 2013). If this is not the case, the estimated effects in the substantive model could be biased. Of course, this principle then implies that individual fixed-effects should be included in the latent regression imputation model for PISA scores, as well as the substantive model linking hours of study to performance. However, this is not the case in the generation of the PISA plausible values, as such a model would be almost impossible to identify (with too little information available for each pupil).

To try and formalize our argument, consider pupils who were randomly assigned to complete science booklets 1 or 5. As illustrated by Table 4, these pupils only answered science test questions, and so have had their mathematics and reading scores imputed based upon (a) how they performed on the science test questions, (b) available background information and (c) the correlation between science and reading (and/or mathematics) scores of the pupils who answered both reading (and/or mathematics) and science test questions.

A simplified version of this process can be thought of as follows:

$$\widehat{R_i} = \alpha_1 + \beta_1.S_i + \gamma_1.X_i + \varepsilon_1 \quad (3)$$

$$\widehat{M_i} = \alpha_2 + \beta_2.S_i + \gamma_2.X_i + \varepsilon_2 \quad (4)$$

Where:

$\widehat{R_i}$ = Imputed reading test scores of pupil i.

$\widehat{M_i}$ = Imputed mathematics test scores of pupil i.

S_i = Performance of pupil i on the PISA science questions

X_i = A vector of background characteristics

ε = Imputation error.

With a pupil fixed-effects model, we are interested in the within-pupil variation only; the difference between these pupils imputed reading and mathematics scores (constraining the problem to $k = 2$ subjects for simplicity). Hence the difference, (3) – (4), becomes:

$$\widehat{R_i} - \widehat{M_i} = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2).S_i + (\gamma_1 - \gamma_2).X_i + (\varepsilon_1 - \varepsilon_2) \quad (5)$$

Particular challenges emerge in equation (5) when $\beta_1 \approx \beta_2$ (i.e. the association between science and reading scores is reasonably similar to the association between science and mathematics

scores) and $\gamma_1 \approx \gamma_2$ (i.e. the association between background factors and performance in reading and mathematics is similar). In such a situation, to the extent that ε_1 is only weakly correlated with ε_2, the final term of equation 5 $(\varepsilon_1 - \varepsilon_2)$ will dominate. Indeed, to the extent that $\beta_1 \approx \beta_2$ and $\gamma_1 \approx \gamma_2$ then (5) reduces to:

$$\widehat{R_\iota} - \widehat{M_\iota} \approx (\varepsilon_1 - \varepsilon_2) \quad (6)$$

In other words, the difference between these pupils' PISA reading and mathematics scores will simply be random noise. More generally, the signal-to-noise ratio in such a situation is likely to be extremely low, given the likely positive association between β_1 and β_2, and between γ_1 and γ_2, while ε_1 is only weakly correlated with ε_2[13][14]. Indeed, this has been implicitly recognized by the psychometricians who have designed the tests, who have warned that '*reliable individual proficiency estimates cannot be obtained*' (Oranje and Ye 2014, p. 204), that they '*are not intended to produce and disseminate individual results at the respondent or even the classroom or school level*' (von Davier and Sinharay 2014, p. 156) and that they '*lack accuracy on the individual test-taker'* (von Davier and Sinharay 2014, p. 156). In other words, the error component $(\varepsilon_1 - \varepsilon_2)$ is so large that test scores for individual pupils are unreliable.

Given the number of unknowns in equation (5), putting a sign or magnitude on the bias this may induce into one's analysis is not possible. Hence whether applying fixed-effects to such data produces reliable and robust estimates becomes an empirical question – which should be tested on a case-by-case basis. Therefore, in the following section, we provide direction to researchers applying fixed-effects to such databases as to how they can check the robustness of their results.

5. What should econometricians do if they want to apply individual fixed-effects using international assessment data?

[13] In PISA, multivariate imputation procedures are used which allow there to be some correlation between the error terms across different subjects (i.e. ε_1 is to some extent allowed to be correlated with ε_2). For instance, in PISA 2006 data the weighted correlation (for all countries) between the first plausible value for reading and the first plausible value for maths (which are jointly drawn) is 0.785. Whereas the correlation between the first plausible value for reading and the second plausible value for maths (which are not jointly drawn) is slightly weaker, standing at 0.759. Nevertheless, the difference between these correlations is small, suggesting the correlation allowed between the errors is relatively weak.

[14] For example, assume that a single X variable is used, and this is parental education. It is likely that there is a similar positive association between parental education and mathematics test scores and between parental education and reading scores. Hence γ_1 would be approximately equal to γ_2.

Although the previous section has outlined our concerns with the use of fixed-effects models applied to international databases, we also appreciate that robust yet pragmatic identification strategies are needed when using such resources to answer important and policy-relevant questions. In this section we therefore provide our advice to analysts who wish to use fixed-effect approaches when analyzing such data.

The intuition behind our recommendations is as follows. Section 4 set out the problem that some pupils do not answer any test questions in some subject areas, and hence have their scores imputed based largely upon how they performed in other domains. Thus, the difference between pupils' scores in these two subjects is likely to mainly be due to the imputation noise $(\varepsilon_1 - \varepsilon_2)$.

However, recall from Table 4 that some pupils do complete a reasonable amount of assessment material (approximately one hour of test questions) in *two* subject areas. For instance, those pupils who were randomly assigned to booklets 3, 4, 8 and 10 completed one hour of science test questions and one hour of mathematics questions. Hence for this sub-sample of pupils one should be less concerned that within-pupil variation in mathematics and science scores is being driven by random imputation error, and is actually likely to be due to genuine and observable differences in pupils' abilities. Consequently, our advice is that some robustness tests should be applied using this sub-sample of pupils only, with alternative test scores created for only those subjects where pupils have actually answered test questions.

We apply this suggestion to the analysis presented by Lavy. For pupils who have completed booklets 3, 4, 8 and 10 we have created new mathematics and science scores, calculated as simply the number of questions that they answered correctly (i.e. pupils are given one mark for each question they answered correctly, half mark for each question partially correct and zero when incorrectly answered; then, scores have been standardized by booklet, to compare with Lavy's estimates).[15] The fixed-effect model presented in equation (2) is then estimated using this sub-sample of pupils only, which capture the effect of the amount of time studying science versus mathematics upon pupils' science and mathematics test scores (i.e. the subjects, k, now include science and mathematics only and not reading). This model is estimated separately for each of the four booklets, as they each contain different sets of mathematics and science

[15] We note that more sophisticated methods could be used to create these scores, including IRT-based techniques. Summative scores have been used here for simplicity and transparency, which we believe to be important when explaining this general approach.

questions (testing different aspects of pupils' mathematics and science ability), and the results compared. Results from this analysis are presented in Table 6. We consider results to be 'robust' if the point estimates and substantive conclusions are consistent across each of the different rows in Table 6 (i.e. regardless of which test booklet is used).

<< Table 6 >>

The first two rows of Table 6 reproduce the results from Lavy (2015), with the estimates using only science and mathematics in the second row. For developed countries, the effect of hours on instruction on science and mathematics outcomes remains positive and statistically significant across the four booklets, with the magnitude of the effect ranging from 0.04 to 0.07 standard deviations. These particular results are clearly rather robust, and largely unaffected by the peculiarities of the PISA test design.

The results for Eastern European countries are somewhat different. The primary results of Lavy reported a positive and statistically significant effect of 0.06. However, this effect vanishes when the analysis is restricted to only those pupils who took science and mathematics test questions, and even turns negative and significant (-0.04) for pupils assigned to booklet 10. In this sense, we believe that simply relying upon the plausible values provided in the international database may lead one to reach the wrong conclusion – a small positive effect may be identified when one does not really exist.

The results for developing countries fall between these two extremes. There is a non-trivial difference in the point estimates when performing the estimates across the test booklets, though they all tend to be positive and small in terms of magnitude. However, we also believe that the additional robustness tests we have conducted in Table 6 bring into question one of the headline findings in Lavy's (2015) abstract, that the effect of instructional time is: "*much lower in developing countries*".

Taken together, Table 6 suggests that some results can change depending upon whether one uses the plausible values when implementing the pupil fixed-effects models, or when restricting the sample to only those pupils who have taken an adequate number of test questions within a given subject area. However, these changes seem to be relatively modest in terms of absolute magnitude. Our key conclusion is therefore that a pupil fixed-effects approach using international achievement databases such as PISA does seem to be a valid identification strategy, though one which should be subject to a series of additional robustness tests as we

have suggested, given the peculiar nature of the test design. This, we believe, is an important finding, and one which potentially opens up new opportunities to those analyzing these databases.

6. 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 practice across the world. Yet the complex survey and test designs used remain misunderstood by many consumers of these data. This not only includes politicians, policymakers and the general public who digest the results, but also the academics who analyze the data to produce secondary research. Resources such as PISA are consequently often being analyzed in a manner not intended by the survey organizers. 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 impact 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 argued how 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 analyze resources such as PISA 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 are needed from the survey organizers 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. 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 providing

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. Indeed, when evaluating the appropriateness of empirical strategies using these data, economists should be aware of how the imputation process is conducted, including the variables that are employed in the underlying imputation model. Various technical reports and user guides now exist, which include many of the key details (e.g. OECD 2009b). Applied researchers should also take more advantage of the many excellent software plugins for analyzing 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, highlighted the care that needs to be taken when analyzing 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 help people to better understand what can and cannot be done with the data, and place more nuanced interpretations 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 June 2017 from https://cran.r-project.org/web/packages/intsvy/intsvy.pdf

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Applications.* Chichester UK: John Wiley & Sons.

Cattaneo, M. A., Oggenfuss, C., and Wolter, S. C. (2017). *The More, the Better? The Impact of Instructional Time on Student Performance*. *Education Economics*, 1–13. http://doi.org/10.1080/09645292.2017.1315055.

Cook, S. R. and Gelman, A. (2006). *Survey Weighting and Regression. Technical Report*. New York: Columbia University, Department of Statistics.

Hanushek, E. A., Piopiunik, M., and Wiederhold, S. (2014). *The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance*. CESifo Working Paper, No. 5120.

Jacob, B. and Rothstein, J. (2016). The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives 30*(3), 85–108. http://doi.org/10.1257/jep.30.3.85

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.

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.

Rivkin, S. G. and Schiman, J. C. (2015). Instruction Time, Classroom Quality, and Academic Achievement. *The Economic Journal 125*, F425–F448. http://doi.org/10.1111/ecoj.12315

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

Rutkowski, L., von Davier, M., and Rutkowski, D. (2014). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*. Boca Raton: CRC Press.

Solon, G., Haider, S. J., and Wooldridge, J. M. (2015). What are we weighting for? *The Journal of Human Resources 50*(2), 301–316. http://doi.org/10.3368/jhr.50.2.301

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)

	OECD		Developing		Eastern Europe	
	Effect size	**SE**	**Effect size**	**SE**	**Effect size**	**SE**
OLS	0.196***	0.007	0.366***	0.012	0.382***	0.013
FE	0.058***	0.004	0.030***	0.008	0.061***	0.006
Observations	460,734		238,938		177,015	

Notes: Results refer to the estimated impact of a one hour increase in instructional time upon pupils' PISA test scores, reported as an effect size.
Coefficient: ***Significant at 1%, ** significant at 5%, * significant at 10%.
Source: Lavy (2015 Table 3 and Table 8).

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**%

Source: Lavy (2015: Table A1) and authors' own calculations from PISA (2006).

(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**%

Source: Lavy (2015: Table A3) and authors' own calculations from PISA (2006).

(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**%

Source: Lavy (2015: Table A2) and authors' own calculations from PISA (2006).

Table 3. Changes to Lavy's OLS estimates when the PISA weights 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*** (+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
Observations	460,734		238,938		177,015	

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.
Coefficient: ***Significant at 1%, ** significant at 5%, * significant at 10%.
Source: Lavy (2015 Table 3 and Table 8) and authors' own calculations.

Table 4. The PISA 2006 test design

Booklet	Clusters			
B1	S1	S2	S4	S7
B2	S2	S3	M3	R1
B3	S3	S4	M4	M1
B4	S4	M3	S5	M2
B5	S5	S6	S7	S3
B6	S6	R2	R1	S4
B7	S7	R1	M2	M4
B8	M1	M2	S2	S6
B9	M2	S1	S3	R2
B10	M3	M4	S6	S1
B11	M4	S5	R2	S2
B12	R1	M1	S1	S5
B13	R2	S7	M1	M3

Notes: 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).
Source: OECD (2009a, p. 29) PISA 2006 technical report.

Table 5. Changes to OLS estimates depending upon how the plausible values are used

	Developed		Developing		Eastern Europe	
	Effect size	SE	Effect size	SE	Effect size	SE
First plausible value only	0.276***	0.011	0.325***	0.016	0.230***	0.016
All five plausible values	0.277***	0.012	0.327***	0.017	0.230***	0.016
Observations	460,734		238,938		177,015	

Notes: Top row refers to the results after applying the final PISA response weights and Balanced-Repeated-Replication (BRR) weights, though only using the first plausible value. The bottom row provides the analogous estimates after following recommended practise in using all five plausible values (see online Appendix E).
Source: Authors' own calculations.

Table 6. Alternative estimates of the pupil fixed-effects model using information from booklets 3, 4, 8 and 10 only

	Developed			Developing			Eastern Europe		
	N	**Effect**	**SE**	**N**	**Effect**	**SE**	**N**	**Effect**	**SE**
Lavy estimates (a)	460,734	0.058***	0.004	238,938	0.030***	0.008	177,015	0.061***	0.006
Lavy estimates (b)	307,156	0.071***	0.006	159,292	0.032**	0.014	118,010	0.011	0.008
Booklet 3	23,554	0.073***	0.013	12,210	0.043*	0.026	9,122	-0.023	0.018
Booklet 4	23,558	0.039***	0.011	12,332	0.004	0.019	9,098	0.001	0.017
Booklet 8	23,614	0.047***	0.011	12,210	0.076***	0.021	9,128	0.007	0.016
Booklet 10	23,676	0.045***	0.012	12,216	0.032	0.024	9,014	-0.035**	0.017

Notes: Lavy estimates (a) stands for Lavy estimations using three subjects: reading, mathematics and science. Lavy estimates (b) stands for Lavy estimations using two subjects: mathematics and science.

Coefficient: ***Significant at 1%, ** significant at 5%, * significant at 10%.

Source: Lavy (2015: Table 3 and Table 8) and authors' own calculations.

Online Appendix A. A review of studies using PISA data in five economics journals

Author(s)	Year	Journal	Weights?	Replicate weights?	Plausible values?
Chevalier, Gibbons, Thorpe, Snell and Hoskins	2009	EER	No	No+	Mentioned*
Corak and Lauzon	2009	EER	Yes	Yes	No
Martins and Veiga	2010	EER	Yes	Yes	Yes
Nonoyama-Tarumi and Willms	2010	EER	Yes	Yes	Yes
Tramonte and Willms	2010	EER	Yes	Yes	Yes
Jensen and Rasmussen	2011	EER	No	No+	Mentioned*
Meunier	2011	EER	Yes	Yes	No
Song	2011	EER	No	No+	No
Woessmann	2011	EER	Yes	No+	No
Filippin and Paccagnella	2012	EER	Yes	Yes	No
Gamboa and Waltenberg	2012	EER	Yes	No	No
Jürges, Schneider, Senkbeil and Carstensen	2012	EER	Yes	No+	No
Micklewright, Schnepf and Silva	2012	EER	No	No+	Mentioned*
Schneeweis and Zweimüller	2012	EER	Yes	No+	No
Brunello and Rocco	2013	EER	Yes	No+	No
Deutsch, Dumas and Silber	2013	EER	Yes	No	No
Hanushek	2013	EER	No	No	No
Lounkaew	2013	EER	No	No+	No
Ryan	2013	EER	Yes	Yes	Yes
Herrero, Mendez and Villar	2014	EER	No	No	Mentioned*
Piopiunik	2014	EER	Yes	No+	No
Polidano and Tabasso	2014	EER	No	No	Mentioned*
Mendez	2015	EER	Yes	Yes	Yes
Vardardottir	2015	EER	Yes	No+	Mentioned*
Giannelli and Rapallini	2016	EER	Yes	No+	Yes
Ruhose and Schwerdt	2016	EER	Yes	No+	No
Hanushek and Wößmann	2006	Econ. J.	No	No+	No
West and Woessmann	2010	Econ. J.	Yes	No+	No
Ohinata and Ours	2013	Econ. J.	No	No+	Mentioned*
Brunello et al	2015	Econ. J.	No	No+	No
Rivkin and Schiman	2015	Econ. J.	No	No+	Mentioned*
Wößmann	2005	Ed. Ec.	Yes	No+	No
Ammermueller	2007	Ed. Ec.	Yes	No+	Yes
Rangvid	2007	Ed. Ec.	No	No+	No
Van Ours	2008	Ed. Ec.	No	No+	Mentioned*
Sprietsma	2010	Ed. Ec.	Yes	No+	Yes
Bratti, Checchi and Filippin	2011	Ed. Ec.	Yes	No+	Yes
Dardanoni, Modica and Pennisi	2011	Ed. Ec.	Yes	No+	No
Perelman and Santin	2011	Ed. Ec.	No	No	No
Agasisti	2013	Ed. Ec.	Yes	Yes	Mentioned*
Kiss	2013	Ed. Ec.	Yes	No	No
Patrinos	2013	Ed. Ec.	No	No	No
Polidano, Hanel and Buddelmeyer	2013	Ed. Ec.	Yes	No	No
Shafiq	2013	Ed. Ec.	Yes	Yes	Yes
Belot and Vandenberghe	2014	Ed. Ec.	Yes	No	No
Hof	2014	Ed. Ec.	No	No	No
Mahuteau and Mavromaras	2014	Ed. Ec.	Yes	Yes	Yes
Murat and Frederic	2015	Ed. Ec.	Yes	Yes	Yes
Oppedisano and Turati	2015	Ed. Ec.	No	No	No
Pritchett and Viarengo	2015	Ed. Ec.	No	No	No
Gramaṭki	2016	Ed. Ec.	Yes	No+	Yes

Jakubowski, Patrinos, Porta and Wiśniewski	2016	Ed. Ec.	Yes	Yes	Yes
Ost, Gangopadhyaya and Schiman	2016	Ed. Ec.	No	No	No
Yang	2016	Ed. Ec.	Yes	No^{+}	No
Ammermueller	2007	Em. Ec.	Yes	No^{+}	Mentioned*
Fuchs and Wößmann	2007	Em. Ec.	Yes	No^{+}	No
Rangvid	2007	Em. Ec.	Yes	No^{+}	No
Schneeweis and Winter-Ebmer	2007	Em. Ec.	Yes	No^{+}	Mentioned*
Mueller and Wolter	2014	Em. Ec.	No	No	No
Jakubowski	2015	Em. Ec.	No	No	No
Foley, Gallipoli and Green	2014	JHR	Yes	No^{+}	No

Notes: All articles using international educational assessment PISA data published in five economics journals since 2005: *Economics of Education Review (EER), The Economic Journal (Econ. J.), Education Economics (Ed. Ec.), Empirical Economics (Em. Ec.)* and *The Journal of Human Resources (JHR).*
*= Mentioned plausible value methodology, but did not average across the five.

+ = A Huber-White Adjustment/clustering was made to the standard errors.

EER search based upon http://www.sciencedirect.com/science/journal/02727757?sdc=1

Econ. J. search based upon http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1468-0297

Ed. Ec. search based upon http://www.tandfonline.com/toc/CEDE20/current

Em. Ec. search based upon http://link.springer.com/journal/181

JHR search based upon http://jhr.uwpress.org/search

References

Agasisti, T. (2013). The efficiency of Italian secondary schools and the potential role of competition: a data envelopment analysis using OECD-PISA2006 data. *Education Economics 21*(5), 520–544. DOI: 10.1080/09645292.2010.511840

Ammermueller, A. (2007). PISA: What makes the difference? Explaining the gap in test scores between Finland and Germany. *Empirical Economics 33*(2), 263–287. DOI: 10.1007/s00181-006-0102-5

Ammermueller, A. (2007). Poor Background or Low Returns? Why Immigrant Students in Germany Perform so Poorly in the Programme for International Student Assessment. *Education Economics 15*(2), 215–230. DOI: 10.1080/09645290701263161

Belot, M. and Vandenberghe, V. (2014). Evaluating the 'threat' effects of grade repetition: exploiting the 2001 reform by the French-Speaking Community of Belgium. *Education Economics 22*(1), 73–89. DOI: 10.1080/09645292.2011.607266

Bratti, M., Cecchi, D. and Filippin, A. (2011). Should you compete or cooperate with your schoolmates? *Education Economics 19*(3), 275–289. DOI: 10.1080/09645292.2011.585021

Brunello, G. and Rocco, L. (2013). The effect of immigration on the school performance of natives: Cross country evidence using PISA test scores. *Economics of Education Review 32*, 234–246. DOI: 10.1016/j.econedurev.2012.10.006

Brunello, G., Weber, G. and Weiss, C. T. 2016. Books Are Forever: Early Life Conditions, Education and Lifetime Earnings in Europe. *The Economic Journal* doi:10.1111/ecoj.12307

Chevalier, A., Gibbons, S., Thorpe, A., Snell, M., and Hoskins, S. (2009). Students' academic self-perception. *Economics of Education Review 28*(6), 716–727. DOI: 10.1016/j.econedurev.2009.06.007

Corak, M. and Lauzon, D. (2009). Differences in the distribution of high school achievement: The role of class-size and time-in-term. *Economics of Education Review 28*(2), 189–198. DOI: 10.1016/j.econedurev.2008.01.004

Dardanoni, V., Modica, S., and Pennisi, A. (2011). School grading and institutional contexts. *Education Economics 19*(5), 475–486. DOI: 10.1080/09645292.2010.488482

Deutsch, J., Dumas, A., and Silber, J. (2013). Estimating an educational production function for five countries of Latin America on the basis of the PISA data. *Economics of Education Review 36*, 245–262. DOI: 10.1016/j.econedurev.2013.07.005

Filippin, A. and Paccagnella, M. (2012). Family background, self-confidence and economic outcomes. *Economics of Education Review 31*(5), 824–834. DOI: 10.1016/j.econedurev.2012.06.002

Foley, K., Gallipoli, G., and Green, D. A. (2014). Ability, Parental Valuation of Education, and the High School Dropout Decision. *The Journal of Human Resources 49*(4), 906–944. DOI: 10.3368/jhr.49.4.906

Fuchs, T. and Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics 32*(2), 433–464. DOI: 10.1007/s00181-006-0087-0

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

Giannelli, G. C. and Rapallini, C. (2016). Immigrant student performance in Math: Does it matter where you come from? *Economics of Education Review 52*, 291–304. DOI: 10.1016/j.econedurev.2016.03.006

Gramatki, L. (2016). A comparison of financial literacy between native and immigrant school students. *Education Economics*, in press. DOI: 10.1080/09645292.2016.1266301

Hanushek, E. A. (2013). Economic growth in developing countries: The role of human capital. *Economics of Education Review 37*, 204–212. DOI: 10.1016/j.econedurev.2013.04.005

Hanushek, E. and Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differencesin-differences evidence across countries. *The Economic Journal* 116: C63-C76.

Herrero, C., Mendez, I., and Villar, A. (2014). Analysis of group performance with categorical data when agents are heterogeneous: The evaluation of scholastic performance in the OECD through PISA. *Economics of Education Review 40*, 140–151. DOI: 10.1016/j.econedurev.2014.02.001

Hof, S. (2014). Does private tutoring work? The effectiveness of private tutoring: a nonparametric bounds analysis. *Education Economics 22*(4), 347–366. DOI: 10.1080/09645292.2014.908165

Jakubowski, M. (2015). Latent variables and propensity score matching: a simulation study with application to data from the Programme for International Student Assessment in Poland. *Empirical Economics 48*(3), 1287–1325. DOI: 10.1007/s00181-014-0814-x

Jakubowski, M., Patrinos, H. A., Porta, E. E., and Wiśniewski, J. (2016). The effects of delaying tracking in secondary school: evidence from the 1999 education reform in Poland. *Education Economics 24*(6), 557–572. DOI: 10.1080/09645292.2016.1149548

Jensen, P. and Rasmussen, A. W. (2011). The effect of immigrant concentration in schools on native and immigrant children's reading and math skills. *Economics of Education Review 30*(6), 1503–1515. DOI: 10.1016/j.econedurev.2011.08.002

Jürges, H., Schneider, K., Senkbeil, M., and Carstensen, C. H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review 31*(1), 56–65. DOI: 10.1016/j.econedurev.2011.08.007

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics 21*(5), 447–463. DOI: 10.1080/09645292.2011.585019

Lounkaew, K. (2013). Explaining urban–rural differences in educational achievement in Thailand: Evidence from PISA literacy data. *Economics of Education Review 37*, 213–225. DOI: 10.1016/j.econedurev.2013.09.003

Mahuteau, S. and Mavromaras, K. (2014). An analysis of the impact of socio-economic disadvantage and school quality on the probability of school dropout. *Education Economics 22*(4), 389–411. DOI: 10.1080/09645292.2014.918586

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

Mendez, I. (2015). The effect of the intergenerational transmission of non-cognitive skills on student performance. *Economics of Education Review 46*, 78–97. DOI: 10.1016/j.econedurev.2015.03.001

Meunier, M. (2011). Immigration and student achievement: Evidence from Switzerland. *Economics of Education Review 30*(1), 16–38. DOI: 10.1016/j.econedurev.2010.06.017

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

Mueller, B. and Wolter, S. C. (2014). The role of hard-to-obtain information on ability for the school-to-work transition. *Empirical Economics 46*(4), 1447–1471. DOI: 10.1007/s00181-013-0709-2

Murat, M. and Frederic, P. (2015). Institutions, culture and background: the school performance of immigrant students. *Education Economics 23*(5), 612–630. DOI: 10.1080/09645292.2014.894497

Nonoyama-Tarumi, Y. and Willms, J. D. (2010). The relative and absolute risks of disadvantaged family background and low levels of school resources on student literacy. *Economics of Education Review 29*(2), 214–224. DOI: 10.1016/j.econedurev.2009.07.007

Ohinata, A. and van Ours, J. C. (2013). How Immigrant Children Affect the Academic Achievement of Native Dutch Children. *The Economic Journal* 123: F308–F331. doi:10.1111/ecoj.12052

Oppedisano, V. and Turati, G. (2015). What are the causes of educational inequality and of its evolution over time in Europe? Evidence from PISA. *Education Economics 23*(1), 3–24. DOI: 10.1080/09645292.2012.736475

Ost, B., Gangopadhyaya, A., and Schiman, J. C. (2016). Comparing standard deviation effects across contexts. *Education Economics*, in press. DOI: 10.1080/09645292.2016.1203868

Patrinos, H. A. (2013). Private education provision and public finance: the Netherlands. *Education Economics 21*(4), 392–414. DOI: 10.1080/09645292.2011.568696

Perelman, S. and Santin, D. (2011). Measuring educational efficiency at student level with parametric stochastic distance functions: an application to Spanish PISA results. *Education Economics 19*(1), 29–49. DOI: 10.1080/09645290802470475

Piopiunik, M. (2014). The effects of early tracking on student performance: Evidence from a school reform in Bavaria. *Economics of Education Review 42*, 12–33. DOI: 10.1016/j.econedurev.2014.06.002

Polidano, C., Hanel, B., and Buddelmeyer, H. (2013). Explaining the socio-economic status school completion gap. *Education Economics 21*(3), 230–247. DOI: 10.1080/09645292.2013.789482

Polidano, C. and Tabasso, D. (2014). Making it real: The benefits of workplace learning in upper-secondary vocational education and training courses. *Economics of Education Review 42*, 130–146. DOI: 10.1016/j.econedurev.2014.06.003

Pritchett, L. and Viarengo, M. (2015). Does public sector control reduce variance in school quality? *Education Economics 23*(5), 557–576. DOI: 10.1080/09645292.2015.1012152

Rangvid, B. S. (2007). School composition effects in Denmark: quantile regression evidence from PISA 2000. *Empirical Economics 33*(2), 359–388. DOI: 10.1007/s00181-007-0133-6

Rangvid, B. S. (2007). Sources of Immigrants' Underachievement: Results from PISA-Copenhagen. *Education Economics 15*(3), 293–326. DOI: 10.1080/09645290701273558

Rivkin, S. G. and Schiman, J. C. 2015. Instruction Time, Classroom Quality, and Academic Achievement. *The Economic Journal* 125: F425–F448. doi:10.1111/ecoj.12315

Ruhose, J. and Schwerdt, G. (2016). Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. *Economics of Education Review 52*, 134–154. DOI: 10.1016/j.econedurev.2016.02.004

Ryan, C. (2013). What is behind the decline in student achievement in Australia? *Economics of Education Review 37*, 226–239. DOI: 10.1016/j.econedurev.2013.08.008

Schneeweis, N. and Winter-Ebmer, R. (2007). Peer effects in Austrian schools. *Empirical Economics 32*(2), 387–409. DOI: 10.1007/s00181-006-0091-4

Schneeweis, N. and Zweimüller, M. (2012). Girls, girls, girls: Gender composition and female school choice. *Economics of Education Review 31*(4), 482–500. DOI: 10.1016/j.econedurev.2011.11.002

Shafiq, M. N. (2013). Gender gaps in mathematics, science and reading achievements in Muslim countries: a quantile regression approach. *Education Economics 21*(4), 343–359. DOI: 10.1080/09645292.2011.568694

Song, S. (2011). Second-generation Turkish youth in Europe: Explaining the academic disadvantage in Austria, Germany, and Switzerland. *Economics of Education Review 30*(5), 938–949. DOI: 10.1016/j.econedurev.2011.03.010

Sprietsma, M. (2010). Effect of relative age in the first grade of primary school on long-term scholastic results: international comparative evidence using PISA 2003. *Education Economics 18*(1), 1–32. DOI: 10.1080/09645290802201961

Tramonte, L. and Willms, J. D. (2010). Cultural capital and its effects on education outcomes. *Economics of Education Review 29*(2), 200–213. DOI: 10.1016/j.econedurev.2009.06.003

Van Ours, J. C. (2008). When do children read books? *Education Economics 16*(4), 313–328. DOI: 10.1080/09645290801976902

Vardardottir, A. (2015). The impact of classroom peers in a streaming system. *Economics of Education Review 49*, 110–128. DOI: 10.1016/j.econedurev.2015.09.002

West, M. R. and Woessmann, L. (2010), 'Every Catholic Child in a Catholic School': Historical Resistance to State Schooling, Contemporary Private Competition and Student Achievement across Countries*. *The Economic Journal* 120: F229–F255. doi:10.1111/j.1468-0297.2010.02375.x

Wößmann, L. (2005). The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics 13*(2), 143–169. DOI: 10.1080/09645290500031165

Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review 30*(3), 404–418. DOI: 10.1016/j.econedurev.2010.12.008

Yang, W. (2016). Does 'compulsory volunteering' affect subsequent behavior? Evidence from a natural experiment in Canada. *Education Economics*, in press. DOI: 10.1080/09645292.2016.1182622

Online Appendix B. Changes to Lavy's OLS estimates if weights are applied

	Canada		Mexico		Russian Federation	
	Effect size	**SE**	**Effect size**	**SE**	**Effect size**	**SE**
Lavy (2015)	0.172***	0.023	0.432***	0.022	0.198***	0.033
+ final student weights	0.127*** (-26%)	0.026	0.377*** (-13%)	0.042	0.193*** (-3%)	0.032
Observations	55,281		70,836		15,051	

Notes: Top row refers to the results obtained following Lavy (2015) procedures, 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 procedures, but now applying the final student weights.
Coefficient: ***Significant at 1% significance level, ** significant at 5%, significance level * significant at 10% significance level.
Source: Authors' own calculations.

Online Appendix C. An overview of how PISA scores ('plausible values') are generated

This appendix provides a simplified overview of the process used to generate the PISA plausible values. Readers interested in further technical details are directed to Rutkowski, von Davier and Rutkowski (2014) and the PISA technical reports (e.g. OECD 2009a) for further discussion.

In order to describe the process used, we have broken it down into the following four steps:

Step 1. Estimation of the international item parameters. In this first stage, a data file that pools together information from multiple countries is used. An Item-Response Theory (IRT) model is estimated using these pooled data to estimate the "item-difficulty" parameter for each question (one can think of this as the IRT equivalent of a percentage correct statistic for each question).[16] This model only incorporates information from pupils who have actually answered the test question to estimate the item-parameters, and occurs before the imputation step. These item international IRT parameters are estimated and considered fixed at this point.

Step 2. A giant principal components analysis (PCA) is conducted upon all the variables included in the background questionnaires. In reference to our empirical study of interest, the information pupils report on instruction time in each subject (H_{ij}, following the notation in the main text) is included within this PCA. The PCA is then estimated separately for each country, with the number of components retained set so that they jointly explain around 80 to 95 percent of the variation in all the background characteristics.[17]

Step 3. A "latent regression" model is estimated, separately for each country, using (a) pupils' responses to the test questions that they answered; (b) the international IRT parameters estimated in step 1; (c) the retained principal components derived in step 2 and (d) school dummy variables. This results in a distribution of possible PISA scores for each pupil in each subject (regardless of whether they have actually answered questions in that subject or not). For those pupils who have not answered a question in a given subject, this distribution is based upon (i) how they responded to the test questions in the other subject areas (ii) the responses they gave to the background questionnaire and (iii) school dummy variables.

Step 4. Five or ten 'plausible values' are drawn from this distribution for each PISA domain.[18] These random values are jointly drawn so to maintain the covariance structure between the scores in the different subject areas. These are the final PISA scale scores.

[16] In PISA 2000 to 2012, a Rasch model was used, and hence only the item-difficulty parameter was estimated from the IRT model (with the discrimination parameter constrained to 1). This changed in PISA 2015, when item-discrimination was only estimated for some items.

[17] The percentage of variance explained by the components has changed between cycles. Although it was 95 percent in 2006, it is a smaller proportion (80 percent) in 2015.

[18] In PISA 2000 to 2012 five plausible values were drawn. In 2015, this increased to 10.

Online Appendix D. 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, referring to a set of pupils who completed test booklet B1. 'PV' stands for plausible value.

Online Appendix E. Rubin's rules for handling multiple imputations (and plausible values)

The correct procedure for handling the plausible values provided in the international achievement databases can be divided into four steps, based upon the original work of Rubin (1987) for handling multiple imputations:

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}).[19]

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:

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

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:

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

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

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.