Economics of Education Review 61 (2017) 51–58

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

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

John Jerrim[a,*], Luis Alejandro Lopez-Agudo[b], Oscar D. Marcenaro-Gutierrez[b], Nikki Shure[a,c]

[a] Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way London, WC1H 0AL, United Kingdom
[b] Departamento de Economía Aplicada (Estadística y Econometría), Facultad de Ciencias Económicas y Empresariales, Universidad de Málaga, Plaza de El Ejido s/n, 29013 Málaga, Spain
[c] Institute of Labor Economics (IZA), Schaumburg-Lippe-Straße 5-9, 53113 Bonn, Germany

ARTICLE INFO

Keywords:
Sample design
Test design
PISA
Weights
Replicate weights
Plausible values

JEL classification:
I20
C18
C10
C55

ABSTRACT

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 among 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 modeling 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.

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 among 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).[1] 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

* Corresponding author.
E-mail addresses: j.jerrim@ucl.ac.uk (J. Jerrim), lopezagudo@uma.es (L.A. Lopez-Agudo), odmarcenaro@uma.es (O.D. Marcenaro-Gutierrez), nikki.shure@ucl.ac.uk (N. Shure).

[1] 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.

http://dx.doi.org/10.1016/j.econedurev.2017.09.007
Received 3 February 2017; Received in revised form 10 August 2017; Accepted 24 September 2017
Available online 28 September 2017
0272-7757/ Crown Copyright © 2017 Published by Elsevier Ltd. All rights reserved.

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. Cattaneo, Oggenfuss, & Wolter, 2017; Hanushek, Piopiunik, & Wiederhold, 2014; Rivkin & Schiman, 2015). 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 Section 1, 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. Cattaneo et al., 2017; Hanushek et al., 2014; Rivkin & Schiman, 2015). 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:

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

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:

$$P_{ik} = \alpha + \gamma \cdot H_{ik} + \mu_i + \varepsilon_{ik} \tag{2}$$

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 Eq. (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.

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,

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 1 h 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: Tables 3 and 8).

stratified and clustered sample design is used.[2] 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 5000 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.[3] 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 to be applied[4]). 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;

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).

[2] We do not discuss here issues with regard 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.

[3] 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 1000).

[4] 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).

and (c) final sampling weights are applied.[5] By not applying weights, too much importance has been given to some countries, while not enough has been given to others. Among 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% of Lavy's developed country sample. However, when either student weights or senate weights are applied, the contribution of Canada falls to around 5–6%. Among developing countries, the figures for Mexico (another country that over-samples) are even more pronounced. Whereas this country drives around a third of Lavy's developing country estimates, it should only account for around 14% based upon its population size. Finally, for Eastern Europe, the opposite holds 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%.

What impact does this have upon the reported OLS regression coefficients?[6] 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%, 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.

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% lower once the final student weights have been applied.

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: Tables 3 and 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–S7 refers to the seven science clusters (italic clusters), M1–M4 the four mathematics clusters (bold clusters) and R1–R2 the two reading clusters (undeline clusters).
Source: OECD (2009a, p. 29) PISA 2006 technical report.

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) 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. 2 h), 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.[7] 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 min of test material. These clusters are labeled S1–S7, M1–M4 and R1–R2, respectively, in Table 4. Out of these clusters, a total of 13 test booklets were formed (labeled 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.

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 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.

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

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

[7] 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.

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). Coefficient: ***Significant at 1%.
Source: Authors' own calculations.

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.

4.1. 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.[8]

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.[9] 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. Namely that all variables included in the substantive model should also be included in the imputation model as well (see Carpenter & 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:

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

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

where:

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

[8] 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.

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

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) and (4), becomes:

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

Particular challenges emerge in Eq. (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 Eq. (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:

$$\widehat{R_i} - \widehat{M_i} \approx (\varepsilon_1 - \varepsilon_2) \quad (6)$$

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.[10,11] 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 & 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 & Sinharay, 2014, p. 156) and that they '*lack accuracy on the individual test-taker*' (von Davier & 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 Eq. (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?

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 1 h of test questions) in *two* subject areas. For instance, those pupils who were randomly assigned to booklets 3, 4, 8 and 10 completed 1 h of science test questions and 1 h 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).[12] The fixed-effect model presented in Eq. (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 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).

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

[10] 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.

[11] 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.

[12] 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.

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	9122	−0.023	0.018
Booklet 4	23,558	0.039***	0.011	12,332	0.004	0.019	9098	0.001	0.017
Booklet 8	23,614	0.047***	0.011	12,210	0.076***	0.021	9128	0.007	0.016
Booklet 10	23,676	0.045***	0.012	12,216	0.032	0.024	9014	−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: Tables 3 and 8) and authors' own calculations.

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 among 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 & 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.

Acknowledgments

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 Ph.D. 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 Ph.D. Program in Economy and Business [*Programa de Doctorado en Economía y Empresa de la Universidad de Malaga*].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2017.09.007.

References

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

Caro, D.. Package 'intsvy': International assessment data manager. (2016). From https://cran.r-project.org/web/packages/intsvy/intsvy.pdf (Accessed June 2017).

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

Cattaneo, M. A., Oggenfuss, C., & Wolter, S. C. (2017). The more, the better? The impact of instructional time on student performance. *Education Economics,* (5), 1–13. http://doi.org/10.1080/09645292.2017.1315055.

Cook, S. R., & Gelman, A. (2006). *Survey weighting and regression.* New York: Columbia University, Department of Statistics *Technical report*.

Hanushek, E. A., Piopiunik, M., & Wiederhold, S. (2014). *The value of smarter teachers: International evidence on teacher cognitive skills and student performance.* CESifo Working Paper, No. 5120.

Jacob, B., & 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. (2009). *PISA 2006 technical report.* Paris: OECD Publishing.

OECD. (2009). *PISA data analysis Manual: SPSS, second edition.* Paris: OECD Publishing.

Oranje, A., & 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., & 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.

Solon, G., Haider, S. J., & 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., & 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.