**Working Paper Department of Applied Economics (Statistics and Econometrics)**

**University of Malaga**

# How did Spain perform in PISA 2018? New estimates of children's PISA reading scores

John Jerrim[1]

Luis Alejandro Lopez-Agudo[2]

Oscar D. Marcenaro-Gutierrez[3],*

This draft January 2020

(Not to be quoted without authors' permission)

**No citar sin autorización expresa de Oscar D. Marcenaro-Gutiérrez**

## Abstract

International large-scale assessments have gained much attention since the beginning of the 21st century, influencing education legislation in many countries. This includes Spain, where they have been used by successive governments to justify education policy change. Unfortunately, there was a problem with the PISA 2018 reading scores for this country, meaning the OECD has thus far refused to release the results. This has caused much frustration amongst policymakers and other interest groups in Spain, particularly as reading was the subject of focus. Therefore, in this paper we attempt to estimate the likely PISA 2018 reading scores for Spain, and for each region within. We find that Spanish reading scores are likely to have fallen to between 475 and 483 test points in 2018, a decline of around 13 to 21 points (approximately 0.13-0.21 standard deviations) compared to 2015.

**Keywords:** PISA; multiple imputation; international large-scale assessments; reading; 2018.

**JEL codes:** I20, I21, I28.

**Acknowledgements:** This work has been partly supported by the *Ministerio de Economía, Industria y Competitividad* under Research Project ECO2017-88883-R, the FEDER funding under Research Project UMA18FEDERJA024 and the postdoctoral contract from the *Plan Propio* signed by the *Universidad de Málaga*.

[1] Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way, WC1H 0AL, London. E-mail: j.jerrim@ucl.ac.uk. Tel.: +44 02076126977. ORCID: 0000-0001-5705-7954

[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-mail: lopezagudo@uma.es. Tel.: +34 952137003. ORCID: 0000-0002-0906-3206

[3],* 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-mail: odmarcenaro@uma.es. Tel.: +34 952137003. ORCID: 0000-0003-0939-5064

## 1. Introduction

International large-scale assessments, such as the OECD's Programme for International Student Assessment (PISA) receive widespread attention. Since its implementation at the beginning of the new century, PISA has received attention from across the globe, even changing education legislation in many countries (Bieber & Martens, 2011; Hopfenbeck, Lenkeit, El Masri, Cantrell, Ryan, & Baird, 2018). This assessment also caused the so-called "PISA shock" in Germany, illustrating the weaknesses of 15-year-old students in key skills and the inequality in achievement between different groups (Waldow, 2009; Pons, 2012). It has also motivated several studies exploring the success of some high-achieving nations, such as Finland (Chung, 2008; Froese-Germain, 2010; Araujo, Saltelli, & Schnepf, 2017), whose education system has been used as a model of good education practices in the last two decades.

Nevertheless, in spite of their importance and ambitious objectives, international large-scale assessments such as PISA are not without controversy. For instance, The Guardian newspaper addressed an open letter to the OECD director (The Guardian, 2014, May 6) which highlighted five negative consequences of PISA, hence calling for its suspension: the narrower curriculum assessed by PISA, which has provoked the change in many countries to an education system based more on standardised tests, leaving other subjects as civic or sport unattended; the short-term fixes made by governments to solve some of the education problems remarked by PISA; its lack of transparency; the intervention of global-profit companies on the development of its testing instruments; and the excessive use of multi-choice testing, which has been increasingly applied by teachers at schools. However, the OECD answered this letter in OECD (2014c), refuting all the points raised. Authors such as Takayama (2015) have analysed this discussion and emphasised that the concerns raised against PISA may be the consequence of a narrow vision of its consequences, due to an excessive focus on a few countries.

Another issue is that the underlying procedures used to obtain and, thus, replicate the scores displayed in the PISA reports remains a mystery (Jerrim, Lopez-Agudo, Marcenaro-Gutierrez, & Shure, 2017) even with the information provided in the technical reports (OECD, 2012, 2014b, 2017, 2020a) and the official analysis manual (OECD, 2009). In fact, the difficulties in understanding the key messages from PISA was highlighted by authors such as Grey and Morris (2018), who found that governments tend to distort the messages, while the media adapt it to their own narrative. All these issues put into light the necessity of greater transparency and explanation on the procedures being performed (Araujo, Saltelli, & Schnepf, 2017; Jerrim, Lopez-Agudo, Marcenaro-Gutierrez, & Shure, 2017).

The complex nature of the PISA assessment has again caused issues with the release of the 2018 results, particularly in Spain. Controversially, the OECD have refused to release PISA 2018 reading scores for Spain, due to apparent anomalies with the data. In particular, the transition to a computer-adaptive assessment in the reading component of 2018 (where children of different abilities are assigned questions of different difficulty) led to some issues. On the PISA 2018 results day, the OECD stated:

*"A large number of Spanish students responded to a new section of the reading test (the reading-fluency section) in a manner that was obviously not representative of their true reading competency (…). In a number of instances, students rushed through the reading-fluency section, spending less than 25 seconds in total over more than 20 test items. In comparison, students who expended adequate effort on these tasks typically spent between 50 seconds and more than two minutes on this section, depending on how quickly*

*they could read. In addition, these students gave patterned responses (all yes or all no, etc.). This response behaviour was not uniform throughout the Spanish sample, but was observed predominantly in a small number of schools in some areas of Spain. The extent and concentration of rapid and patterned responses are unique to Spain, and affect the data on reading performance*" (OECD, 2019b).

This issue occurred for the reading-fluency section of PISA, influencing results for the entire (major) reading domain. On the other hand, Spain's mathematics and science results were deemed sufficiently reliable by the OECD to allow for their release.

The fact that PISA reading scores could not be released for Spain has obviously been the source of major embarrassment, both to the government, education policymakers and (most of all) to the OECD. Yet this is not the first time that such a major issue in the administration of PISA has occurred. For instance, in PISA 2009, a dispute between teacher unions and the education minister in Austria led to a boycott of PISA, meaning the Austrian data was deemed not comparable with previous cycles (OECD, 2010; Annex A4). Another example for PISA 2009 is Azerbaijan, where there were a number of anomalies with the data, including a suspicion that the test markers were sometimes too lenient (OECD, 2010; Annex A4). Nevertheless, in spite of these irregularities, the OECD included Azerbaijan in the PISA 2009 results (though without giving a clear reason why). Likewise, in PISA 2006 results for the United States were not reported due to a printing error. Clearly, the situation with the Spanish PISA reading data is not unique, with similar challenges occurring in other countries in the past, and will likely affect some other countries in the future[4].

In this context, the objective of our research will be: first, to predict the likely PISA reading scores for Spain in 2018. We do this via imputation procedures, based upon the correlation between reading scores and mathematics and science scores observed in other OECD countries. This will not only provide the first insight into Spain's 2018 PISA reading scores, but also potentially illustrate a methodology that can be used to estimate a country's PISA performance if similar problems arise in a future cycle. Second, we intend to describe as clearer as possible the procedure used to generate PISA scores across the different subject domains. This will allow to shed some light to interested readers about how PISA imputation process works.

The rest of the paper is structured as follows. First, we describe the PISA data. This is followed by an overview of the multiple imputation methodology we employ. The results are then presented, and conclusions then drawn.

## 2. Data

PISA is conducted by the OECD and intends to assess 15-year-old students' competences for many countries in reading, mathematics and science. Since 2015, most countries answered to PISA using computer-based assessments, instead of paper tests[5]. In order to participate, the countries which take computer-based assessment tests are required a sample of at least 42 15-year-old students in 150 schools, while those participating in paper-based assessment are required a minimum sample of 35 students in 150 schools. These students take a cognitive test in these domains and also answer a

[4] There are many other examples of problems with PISA data in specific countries; for example, in PISA 2012 Albania presented some serious irregularity (OECD, 2014a; Annex A4), in PISA 2015 Albania, Argentina, Kazakhstan and Malaysia (OECD, 2016; Annex A4) and, in PISA 2018, Viet Nam and Spain (OECD, 2019c; Annex A4).

[5] The list of countries participating on paper-based assessment in PISA 2018 can be found in OECD (2019c, Annex A5).

student questionnaire. In addition, head teachers answer a school background questionnaire[6].

The main data employed for the present research are those of PISA 2018, while PISA 2009, 2012 and 2015 are also used to check the robustness of our results. Although Spain also participated in PISA 2000 (only in reading), 2003 (only in reading and mathematics) and 2006, we focus on later cycles as they are more recent and reading was the main domain for PISA 2009, meaning it provides a reference for PISA 2018 (in which reading was the main domain again). Moreover, very few trend reading items from the 2000, 2003 and 2006 sweeps were used in later PISA cycles, meaning we believe that most recent waves (from 2009 onwards) are likely to be more comparable.

## 3. PISA score creation procedure

In the following we describe in the clearest possible way the procedures followed by PISA 2018 to create students' scores in the different domains[7,8]. In PISA, students do not answer all the cognitive test questions defined for each domain in order to reduce the burden and time taken by them, so the full set of items (that is, more than 600 computer-based and 250 paper-based items) is organised in different test forms and distributed in assessment designs, which contain a different proportion of questions in each cognitive domain (reading, mathematics and science) and with different difficulties. In particular, for PISA 2018, 92% of students who took the global competence assessment[9] (and 88% of those who did not take it) answered 1 hour of reading questions (as it is the major domain of 2018) and, after a short break, two 30-minute clusters of other domain. The remaining students (12% or 8%, respectively) took 1 hour of reading and, after a short break, two 30-minute clusters of two different domains. Hence, students do not answer all the test items for each domain and, in some cases, they do not answer any question about a particular domain. This is because this procedure is aimed at obtaining reliable population estimates, what comes at the cost of not being able to make valid inference on individuals' skills for a particular domain. In this sense, as the OECD has documents on many occasions (OECD, 2020a), any kind of inference made using the total number of correct responses to the administered items would not be valid. This is because the differences found in cognitive skills between individuals may be due to differences in the average difficulty of the test forms and not actually due to their ability.

In addition, for the first time, PISA 2018 used a multi-stage adaptive test only for the reading literacy domain. Concretely, students received a core reading test form, which diverged into easier or difficult test forms based on students' performance in the core

[6] Many more competences (such as financial literacy, problem-solving skills or the global competence) are assessed by PISA, together with other background questionnaires (parental, teacher, ICT, well-being, educational career questionnaires); nevertheless, their administration has been performed irregularly by PISA cycles and not all countries took them, so we focus on the competences and student information which remain fixed through PISA cycles.

[7] Official information on other previous PISA subjects such as sample design and weighting can be found at OECD (2009, 2012, 2014b, 2017, 2020a). A summary of this topic can be found in Jerrim, Lopez-Agudo, Marcenaro-Gutierrez, and Shure (2017).

[8] Due to the change from a paper- to a computer-based assessment since PISA 2015 some of these PISA procedures changed from one cycle to the following; hence, we focus here on the last cycle (2018), but more information on this subject for PISA 2009, 2012 and 2015 can be found at OECD (2012, 2014b, 2017).

[9] This global competence was new in PISA 2018 and it "examines students' ability to consider local, global and intercultural issues, understand and appreciate different perspectives and world views, interact respectfully with others, and take responsible action towards sustainability and collective well-being" (OECD, 2019c, p. 29).

test[10]. The change to computer-based assessment means it is possible to measure students' time on the test and to identify outliers in terms of time and response patterns (as happened with Spain). Hence, a total pool of 244 items for reading (30% were new items and 70% trend items) and 65 for reading fluency (all new items) were distributed to students using the adaptive test design. Note that, while 82 for mathematics (all trend items) and 115 for science (all trend items) were distributed to students in PISA 2018, the adaptive test design was not used for these domains.

In order to derive the final PISA scores, a multi-group item response theory (ITR) scaling model is used. Concretely, this is a two-parameter logistic model (2PL) for those items with binary response and a generalised partial credit model (GPCM) for those items with a polytomous configuration[11]. This methodology provides a comparable latent scale across countries and PISA cycles in each domain and let to place all students in a common proficiency scale, so that the performance of students in the population and the groups of this population can be obtained, in spite of answering different test forms and items. This is possible due to the regularities presented by the response patterns of students when answering same-skill level items, to the extent that these items show common characteristics. This IRT model, combined with a multivariate latent regression model that incorporates student background information, configure the population model.

Once students' answers to their administered cognitive test questions and their background characteristics[12] are collected, a proficiency distribution is defined. This proficiency distribution is created for all the cognitive domains at the same time (taking data from other domains) so that the accuracy of the estimates could be improved. In addition, the covariance among skill domains (i.e. reading, mathematics and science) is added to improve this estimation of proficiency distributions. Then, the plausible value methodology employs this distribution and, instead of using individual point estimates, it randomly draws multiple imputed proficiency values from it (commonly called plausible values) to account for error or uncertainty at the individual level. In PISA 2018, ten plausible values[13] for each domain were drawn from this distribution. Those plausible values from the domains that students did not take have a higher measurement error; because of that, the use of ten plausible values when estimating lets to account for this error. These plausible values are not unbiased for estimating each individual's skills, but they are consistent for the whole population. They are standardised to have mean 500 and standard deviation 100 and are used, together with final student weights and balanced repeated replication, to obtain estimations of the population skill level for each one of the domains under analysis.

## 4. Methodology

The method we use to predict Spanish reading scores for PISA 2018 centres around multiple imputation (Rubin, 1976; Rubin, 1987; Schafer, 1997). The intuition behind our approach is that mathematics and science scores for Spain were not influenced by the problem that affected the reading scores. Using data from the other 36 OECD

[10] More information on this procedure can be found in OECD (2019a).

[11] The software employed by the OECD to perform these IRT models is mdltm (von Davier, 2005).

[12] This background information was incorporated by, first, coding variables so that refused responses could be included (i.e. contrast coding); then, a principal component analysis was performed, so that background information can be summarised and information from students with missings can be kept, satisfying the linearity assumption for the model (OECD, 2020a).

[13] The OECD employed the software DGROUP (Rogers, Tang, Lin, & Kandathil, 2006) to estimate the multivariate latent regression model and obtain the plausible values to estimate this model, fixing the parameters of the cognitive items obtained from the multi-group IRT models.

countries participating in PISA 2018, we can estimate how performance in reading (which is unobserved for Spain) is related to performance in mathematics and science (which we can observe for Spain). We can then use this information on the relationship between reading, mathematics and science scores from other OECD countries to create a distribution of predicted average PISA reading scores in Spain. This distribution then provides us with a good idea of the likely average PISA score for Spain in 2018.

Specifically, for Spain and each other OECD country (which we call a "donor country"), we implement the following procedure:

- We begin by restricting the PISA data to only Spain and one other donor OECD country (e.g. Portugal).
- We then run a multiple imputation model for Spain, treating all the reading scores in this country as "missing data". Mathematics and science PISA scores (plausible values) for both countries are included in the imputation model. This means we essentially predict reading scores for Spanish pupils, based upon how they answered the PISA mathematics and science questions, under the assumption that the association between reading, mathematics and science is the same in Spain as in the other donor country (e.g. Portugal).
- We repeat the procedure above for every OECD country. In other words, we create estimated PISA reading scores for Spain, using the 36 other OECD countries as possible donors.
- This gives us 36 alternative estimates (plus 1 additional estimate using all the 36 countries at the same time) of the average PISA reading scores in Spain. These will vary due to the different relationship between reading, mathematics and science achievement across the OECD (plus the random component introduced by imputation).
- We repeat the above procedure using data from previous PISA cycles (2009, 2012 and 2015) where we can observe the "true" average reading scores for Spain. This is used to establish which of the 36 alternative estimates for Spain's PISA 2018 scores are likely to be the most plausible. It also allows us to check how well this procedure works in reproducing the "true" average reading score for Spain in previous cycles (indicating how confident we should be able our estimate of the PISA 2018 reading score for Spain).
- To test the robustness of our results, we run alternative versions of the imputation model used, most notably also including information from the background questionnaire (gender and economic, social and cultural status – ESCS – index[14] quartile)[15].

All imputation models have been estimated 10 times in order to reduce the imputation error to an acceptably small level. After performing the multiple imputation methodology, all PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009) to estimate the Spanish reading mean scores in PISA 2018.

[14] The ESCS index was created by the OECD using the highest level of education of parents, highest parental occupation, and home possessions by the use of principal component analysis (OECD, 2020a).
[15] These variables have been consistently found in the literature to be very relevant in the definition of the education production function (Hanushek, 1979; Hyde, Fennema, & Lamon, 1990; Sirin, 2005; Wößmann, 2005; Reilly, Neumann, & Andrews, 2015; Karadag, 2017).

## 5. Results

### *5.1. Main imputation results*

Our main results are presented in Table 1. In particular, this first multiple imputation model uses only plausible values in mathematics and science as covariates within the prediction equation. We see that there is a great deal of variation in the estimated average PISA reading score for Spain in 2018. These range from a minimum of 457, when using Japan as the donor country, to a maximum of 503 when using Ireland as the donor country[16]. Yet most of the estimates do fall within a reasonably narrow range, i.e. the interquartile range spans from 475 (using Austria, Estonia or Lithuania) to 486 (Denmark, Norway or United Kingdom), with a score of 482 when using all OECD countries within the donor pool at the same time. Under the assumption that Spain is a fairly "typical" OECD country, in terms of the relationship between reading, mathematics and science scores, then one would anticipate Spain to be around the average across all the potential donors (i.e. 482 points) and, most likely, within a range of 475 to 486 (based upon the interquartile range).

**<< Table 1 >>**

Next, in Table 2, we present results having re-run our procedure using the PISA 2009, 2012 and 2015 data. Here the difference with PISA 2018 is that we actually know what the average PISA reading scores were for Spain. Hence we can observe where in the distribution of our possible values Spain actually fell. Thus, assuming that the relationship between reading, mathematics and science within Spain has not changed dramatically in 2018 compared to previously, this should help sharpen our prediction of where in the distribution of possible values Spain is likely to fall.

**<< Table 2 >>**

Importantly, Table 2 reveals that the actual observed PISA reading score for Spain in 2009, 2012 and 2015 was similar to that obtained when using all the OECD countries within the donor pool. For instance, the last time reading was the focus subject in PISA (2009), Spain achieved an average reading score of 481 points. Using the approach we have outlined in this paper, and taking all other OECD countries as donors, we would have estimated Spain's PISA 2009 score to have been 483 points. Similarly, again when using all OECD countries within the donor pool, differences between our predictions and observed average reading scores for Spain were small when testing our procedure using the 2012 (real 488 versus 491 predicted points) and 2015 (real 496 versus 492 predicted points) cycles. This makes us confident in our prediction that the average PISA 2018 reading score for Spain should sit around 482 test points.

In Appendix A (Tables A1 and A2) we present this same analysis but now also including sex and ESCS index quartile into the multiple imputation model. Again, when looking at the estimates using all the countries as donor countries in Table A1, we come to a prediction of 481 points for Spanish reading scores in PISA 2018. Consequently, the predicted score is almost identical to that obtained excluding students' sex and the ESCS index quartile from the imputation model.

[16] These results make sense. For PISA 2018, in Ireland reading scores are high compared to mathematics and science scores (518, 500, 496, respectively; OECD, 2019c), meaning we get the largest imputed value for Spain when using this nation as the donor country. On the other hand, in Japan reading scores are lower than mathematics and science scores (504, 527, 529, respectively; OECD, 2019c), meaning that our predicted score for Spain is very low.

We have also estimated our imputation model dividing the sample by gender. These results are presented in Table 3 and show that our multiple imputation model also makes good predictions of the real gender gap found in PISA reading scores in 2009, 2012 and 2015. Consequently, we also believe that our approach allows us to estimate the gender gap in PISA reading scores in Spain in 2018. Using our approach, we estimate that girls outperformed boys in reading by 25 points. This is of a similar magnitude to the gender gap in PISA reading scores observed for Spain in previous PISA cycles (OECD, 2010; OECD, 2014a; OECD, 2016)[17].

**<< Table 3 >>**

This same analysis has also been performed for each one of the Spanish regions in Table B1 (Appendix B). This uses all OECD countries as donor countries, which we also find provides a good approximation to the actual scores of Spain's different regions in our analysis of PISA 2009, 2012 and 2015. Thus, it seems that, following the trend for Spain as a whole, all regions have seen their reading scores decreased compared to PISA 2015. Those regions which presented the deepest decline were Madrid, Valencian Community and Castile and Leon (0.35, 0.23 and 0.22 standard deviations[18], respectively) while those with the lowest decline were Basque Country, Galicia and La Rioja (0.02, 0.02 and 0.03 standard deviations, respectively).

*5.2. Robustness checks. Alternative ways of choosing imputations*

Another approach for choosing imputations for Spanish reading scores in PISA 2018 could be checking which countries presented scores in reading, mathematics and science that were similar to those obtained by Spain during the PISA cycles under analysis. In order to do this, we have generated Table 4, which presents a test of mean differences between Spanish scores in reading, mathematics and science and those obtained by the rest of the OECD countries, for PISA 2009 to 2018. We can adopt here two different criteria. First, we focus attention on those countries which did not show large differences in reading scores compared to Spain for some PISA cycles. In this case, we can see that Portugal is a country that is similar to Spain in term of reading scores (it does not present significant differences in PISA 2012 and 2015 with Spanish reading scores). Hence, returning to Tables 1 and A1 (Appendix) and focusing on the Portugal row, we can see that imputed reading scores for Spain are again around 482 points. A second criterion could be using those countries with only small differences (relative to Spain) in mathematics and science in PISA 2018. This is the case of Hungary and Lithuania which, looking at Tables 1 and A1 (Appendix), suggests Spain's reading score will be around 475-478.

**<< Table 4 >>**

Alternatively, one could focus upon the results using a particular "donor" country that has a similar trend over time as Spain, i.e. a country where the trend in PISA scores

[17] In order to check the capacity of our model to predict gender differences in scores for PISA 2018, we have run a similar specification for Spain in mathematics, considering this domain as missing completely at random for PISA 2009 to 2018. This model has accurately estimated PISA 2009, 2012 and 2015 mathematics scores by gender and has also predicted a mathematics score for boys of 489 (being the real one 485) and 479 for girls (being the real one 478), so we can be quite confident on the results of our multiple imputation model also by gender. Results for mathematics scores in previous PISA cycles will be provided upon request to the authors.

[18] These differences in terms of standard deviations have been obtained by calculating the absolute difference between the previous and predicted reading scores for that region and dividing the result by 100 (which is plausible values' standard deviation).

over time in reading, mathematics and science scores has been similar to Spain. To do this, in the Appendix C, we have plotted all the PISA scores from 2000-2018 in reading, mathematics and science for all the OECD countries (Figures C1, C2 and C3, respectively). Canada is the country with a trend that resembles that of Spain across the three domains (i.e. the Canadian scores in the three domains fluctuate in a similar way across PISA cycles – although at a higher level – as Spain). Looking at our predictions in Table 1 and focusing upon Canada as the "donor" country, we predict average reading scores for Spain of 483. In addition, the imputations of Spanish reading scores using Canada as the donor country are quite similar to the actual PISA scores for Spain in PISA 2009-2015 (see Table 2 or A2 – Appendix A).

In the view of these results, we suggest that average PISA 2018 reading scores for Spain are likely to fall around 480 points, and likely within a range of around 475-483 points. This is around 13 to 21 points (0.13 to 0.21 standard deviations) lower than in PISA 2015.

## 6. Discussion and conclusions

In this paper we have explored the issue of the administration of PISA 2018 in Spain, focusing on a particular issue that affected scores in the reading domain. The fact that these results could not be reported with those of other countries in December 2019, despite the potential relevance of this information for policymakers, has received a great deal of media attention, and was a major source of embarrassment for both the OECD and the government in Spain.

In an attempt to resolve this situation, in this paper we have estimated the average reading scores that Spain would have likely achieved had the administration of the reading component not been problematic. Bearing in mind that we do not have any reading test answers to build the model, we based our estimates solely upon the link between reading, mathematics and science scores observed in other OECD countries. In doing so, we have provided the first insight into Spain's 2018 PISA reading scores. We have confidence in the approach we have used as our analysis shows that it worked well at recovering the "true" reading scores for Spain when the procedure has been applied to data from previous PISA cycles. The paper has therefore also presented a methodology that could be used to estimate a country's performance if similar problems arise in a future PISA cycle.

Our results show that Spain would have performed lower in PISA 2018 than in the previous cycle. Our headline estimate puts the PISA reading score for Spain at around 480 points, which is approximately 0.13-0.21 standard deviations lower than in PISA 2015. This striking result should make decision makers consider why PISA reading scores in Spain are likely to have declined, and to potentially change policy in response.

In spite of the efforts we have made to understand and summarise the complex procedures used by the OECD to generate PISA scores, there is still a great deal of uncertainty about its methodology. This is not helped by the fact that the OECD are not open and transparent about their statistical methodology, and refused to publish the code detailing how exactly PISA scores are generated. We hence strongly suggest that the OECD provides the syntax files used for the creation of the PISA "plausible values" (test scores). In our view, the only way for the OECD to regain their credibility in Spain after its embarrassing failure in the administration of PISA 2018 is for it to become much more transparent about its methodology, and a willingness to provide much more details about difficulties with administration when they arise.

## References

Araujo, L., Saltelli, A., & Schnepf, S. V. (2017). Do PISA data justify PISA-based education policy? I*nternational Journal of Comparative Education and Development, 19*, 20–34. doi: 10.1108/IJCED-12-2016-0023

Bieber, T., & Martens, K. (2011). The OECD PISA Study as a Soft Power in Education? Lessons from Switzerland and the US. *European Journal of Education, 46*(1), 101–116. doi: 10.1111/j.1465-3435.2010.01462.x

Chung, J. (2008). An investigation of reasons for Finland's success in PISA (Doctoral dissertation). Retrieved from https://ora.ox.ac.uk/objects/uuid:62b7a22f-d930-4eb0-893d-d703fd9d182d. University of Oxford.

Froese-Germain, B. (2010). *The OECD, PISA and the Impacts on Educational Policy*. Canada: Canadian Teachers' Federation.

Grey, S., & Morris, P. (2018). PISA: multiple 'truths' and mediatised global governance. *Comparative Education, 54*(2), 109–131, doi: 10.1080/03050068.2018.1425243

Hanushek, E. A. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *The Journal of Human Resources, 14*(3), 351–388. doi: 10.2307/145575

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-O. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353. doi: 10.1080/00313831.2016.1258726

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155. doi: 10.1037/0033-2909.107.2.139

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review, 61*, 51–58. doi: 10.1016/j.econedurev.2017.09.007

Karadag, E. (2017). *The Factors Effecting Student Achievement: Meta-Analysis of Empirical Studies*. New York: Springer.

MECD (2010). *PISA 2009. Programa para la Evaluación Internacional de los Alumnos. Informe Español. Resultados y contexto. OECD*. Madrid: Ministerio de Educación, Cultura y Deporte.

MECD (2014). *PISA 2012 Programa Para La Evaluación Internacional de los Alumnos. Informe Español*. Madrid: Ministerio de Educación, Cultura y Deporte.

MECD (2016). *PISA 2015 Programa para la Evaluación Internacional de los Alumnos. Informe Español*. Madrid: Ministerio de Educación, Cultura y Deporte.

MECD (2019). *PISA 2018. Informe Español*. Madrid: Ministerio de Educación, Cultura y Deporte.

OECD (2009). *PISA Data Analysis Manual SPSS® Second Edition*. Paris: OECD Publishing.

OECD (2010). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)*. Paris: OECD Publishing. doi: 10.1787/9789264091450-en

OECD (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing. doi: 10.1787/9789264167872-en

OECD (2014a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. Paris: OECD Publishing. doi: 10.1787/9789264201118-en

OECD (2014b). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD (2014c). *Response to points raised in Heinz-Dieter Meyer "open letter"*. Retrieved from http://www.oecd.org/pisa/aboutpisa/OECD-response-to-Heinz-Dieter-Meyer-Open-Letter.pdf

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing. doi: 10.1787/9789264266490-en

OECD (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

OECD (2019a). *Introduction of multistage adaptive testing design in PISA 2018*. Paris: OECD Publishing. doi: 10.1787/b9435d4b-en

OECD (2019b). PISA 2018 in Spain. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/PISA2018Spain_final.pdf

OECD (2019c). *PISA 2018 Results (Volume I): What Students Know and Can Do*. Paris: OECD Publishing. doi: 10.1787/5f07c754-en

OECD (2020a). *PISA 2018 Technical Report*. Paris: OECD Publishing.

OECD (2020b). *PISA reading, mathematics and science performance*. Retrieved from: https://data.oecd.org/pisa/reading-performance-pisa.htm#indicator-chart

Pons, X. (2012). Going beyond the 'PISA Shock' Discourse: an analysis of the cognitive reception of PISA in six European countries, 2001-2008. *European Educational Research Journal*, *11*(2), 206–226. doi: 10.2304/eerj.2012.11.2.206

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology, 107*(3), 645–662. doi: 10.1037/edu0000012

Rogers, A., Tang, C., Lin, M.-J., & Kandathil, M. (2006). *DGROUP (computer software)*. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. doi: 10.2307/2335739

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

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research, 75*(3), 417–453. doi: 10.3102/00346543075003417

Takayama, K. (2015). *Has PISA helped or hindered?* Singapur: THF Lecture Series.

The Guardian (2014, May 6). *OECD and Pisa tests are damaging education worldwide – academics*. Retrieved from

https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics

von Davier, M. (2005). *A general diagnostic model applied to language testing data. (Research report No. RR-05-16)*. Princeton, NJ: Educational Testing Service.

Waldow, F. (2009). What PISA Did and Did Not Do: Germany after the 'PISA-shock'. *European Educational Research Journal, 8*(3), 476-483. doi: 10.2304/eerj.2009.8.3.476

Wößmann, L. (2005). Educational production in Europe. *Economic Policy, 20*(43), 445–504. doi: 10.1111/j.1468-0327.2005.00144.x

## Appendix A

**<< Table A1 >>**

**<< Table A2 >>**

## Appendix B

**<< Table B1 >>**

## Appendix C

**<< Figure C1 >>**

**<< Figure C2 >>**

**<< Figure C3 >>**

Table 1. Multiple imputations of Spanish reading scores for PISA 2018 using reading, mathematics, science plausible values (using data from different "donor" countries)

| **Donor country** | **Imputed reading score for Spain** |
|---|---|
| Ireland | 503 |
| Chile | 501 |
| Colombia | 495 |
| Israel | 493 |
| Mexico | 493 |
| United States | 493 |
| Greece | 492 |
| Sweden | 487 |
| Denmark (top quartile) | 486 |
| Norway (top quartile) | 486 |
| United Kingdom (top quartile) | 486 |
| Australia | 484 |
| Italy | 484 |
| Canada | 483 |
| Finland | 483 |
| New Zealand | 483 |
| Turkey | 483 |
| France (median) | 481 |
| Iceland (median) | 481 |
| Poland (median) | 481 |
| Portugal (median) | 481 |
| Germany | 478 |
| Hungary | 478 |
| Czech Republic | 477 |
| Korea | 477 |
| Belgium | 476 |
| Austria (bottom quartile) | 475 |
| Estonia (bottom quartile) | 475 |
| Lithuania (bottom quartile) | 475 |
| Luxembourg | 474 |
| Latvia | 471 |
| Slovak Republic | 471 |
| Slovenia | 470 |
| Switzerland | 466 |
| Netherlands | 460 |
| Japan | 457 |
| **All countries as donors** | **482** |

Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Imputation method: multiple imputation with 10 complete estimations has been employed. Spanish reading scores have been considered as missing completely at random.
Source: Authors' own calculations.

Table 2. Multiple imputations of Spanish reading scores for PISA 2009, 2012 and 2015 using reading, mathematics, science plausible values (using data from different "donor" countries)

| PISA 2009 | | PISA 2012 | | PISA 2015 | |
|---|---|---|---|---|---|
| Austria | a | Colombia | b | Lithuania | b |
| Colombia | b | Lithuania | b | Colombia | c |
| Lithuania | b | Latvia | c | Ireland | 510 |
| Latvia | c | Israel | 511 | Norway | 506 |
| Israel | 510 | Turkey | 507 | Greece | 505 |
| Greece | 501 | Greece | 506 | Chile | 503 |
| Chile | 496 | Ireland | 503 | Israel | 503 |
| Mexico | 496 | Norway | 503 | Mexico | 501 |
| Turkey | 496 | France | 500 | Sweden | 498 |
| Korea | 493 | United States | 498 | France (top quartile) | 497 |
| Norway | 491 | Mexico | 497 | Iceland (top quartile) | 497 |
| Iceland (top quartile) | 490 | New Zealand (top quartile) | 495 | Poland( top quartile) | 497 |
| Sweden (top quartile) | 490 | Hungary | 493 | Official Spain score | 496 |
| United States | 488 | Sweden | 493 | United States | 496 |
| Portugal | 486 | Belgium | 492 | Finland | 494 |
| Belgium | 485 | Canada | 492 | Germany | 494 |
| Denmark | 485 | Denmark | 491 | Korea | 494 |
| France | 485 | Italy | 491 | Canada | 493 |
| Poland | 485 | Portugal | 491 | Italy | 492 |
| Italy (median) | 484 | Chile (median) | 490 | Denmark (median) | 491 |
| Canada | 482 | Luxembourg (median) | 490 | Luxembourg (median) | 491 |
| Hungary | 482 | Poland (median) | 490 | New Zealand (median) | 491 |
| Ireland | 482 | Australia | 489 | Turkey (median) | 491 |
| Official Spain score | 481 | United Kingdom | 489 | Latvia | 490 |
| New Zealand | 481 | Japan | 488 | Portugal | 490 |
| Australia | 478 | Korea | 488 | Belgium | 488 |
| Finland | 477 | Official Spain score | 488 | Australia | 487 |
| Netherlands | 477 | Iceland | 486 | Czech Republic | 487 |
| United Kingdom (bottom quartile) | 476 | Czech Republic (bottom quartile) | 484 | United Kingdom (bottom quartile) | 486 |
| Luxembourg | 475 | Finland | 483 | Hungary | 484 |
| Slovak Republic | 473 | Germany | 482 | Slovenia | 484 |
| Switzerland | 473 | Netherlands | 482 | Estonia | 483 |
| Japan | 472 | Slovak Republic | 481 | Netherlands | 483 |
| Czech Republic | 470 | Estonia | 479 | Austria | 482 |
| Germany | 469 | Switzerland | 479 | Slovak Republic | 478 |
| Estonia | 468 | Austria | 478 | Switzerland | 477 |
| Slovenia | 460 | Slovenia | 462 | Japan | 475 |
| **All countries a donors** | **483** | **All countries as donors** | **491** | **All countries as donors** | **492** |

Notes: [a] Austria does not have reading, mathematics and science scores due to a dispute between teacher unions and the education minister, which led to a boycott and a negative atmosphere which affected the conditions under which the assessment was administered, so they were not reliable (OECD, 2010; Annex A4). [b] The country did not participate in that cycle. [c] The country had scores but was not in the OECD in that cycle. PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Imputation method: multiple imputation with 10 complete estimations has been employed. Spanish reading scores have been considered as missing completely at random.
Source: Authors' own calculations.

Table 3. Multiple imputations of Spanish reading scores by gender, using as donors all OECD countries

| | Boys | | Girls | |
|---|---|---|---|---|
| **PISA cycle** | **Official reading scores for Spain** | **Imputation using plausible values in reading, mathematics and science, sex and ESCS index quartile** | **Official reading scores for Spain** | **Imputation using plausible values in reading, mathematics and science, sex and ESCS index quartile** |
| 2009 | 467 | 467 | 496 | 499 |
| 2012 | 474 | 474 | 503 | 506 |
| 2015 | 485 | 480 | 506 | 503 |
| 2018 | tbi | 469 | tbi | 494 |

Notes: "tbi" means "to be imputed". PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Imputation method: multiple imputation with 10 complete estimations has been employed. Spanish reading scores have been considered as missing completely at random.
Source: Authors' own calculations.

Table 4. Mean scores in reading, mathematics and science for PISA 2009-2018 and test of mean differences compared to Spain

| PISA cycle | Spain | Donor country | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reading scores** | | | | | | | | | | | | | | | | | | | |
| | | Australia | Austria | Belgium | Canada | Switzerland | Chile | Colombia | Czech Republic | Germany | Denmark | Estonia | Finland | France | United Kingdom | Greece | Hungary | Ireland | Iceland | Israel |
| 2009 | 481 | 515*** | a | 506*** | 524*** | 501*** | 449*** | b | 478* | 497*** | 495*** | 501*** | 536*** | 496*** | 494*** | 483 | 494*** | 496*** | 500*** | 474*** |
| 2012 | 488 | 512*** | 490 | 509*** | 523*** | 509*** | 441*** | b | 493*** | 508*** | 496*** | 516*** | 524*** | 505*** | 499*** | 477*** | 488 | 523*** | 483*** | 486 |
| 2015 | 496 | 503*** | 485*** | 499** | 527*** | 492* | 459*** | c | 487*** | 509*** | 500*** | 519*** | 526*** | 499** | 498 | 467*** | 470*** | 521*** | 482*** | 479*** |
| 2018 | tbi | 503[d] | 484[d] | 493[d] | 520[d] | 484[d] | 452[d] | 412[d] | 490[d] | 498[d] | 501[d] | 523[d] | 520[d] | 493[d] | 504[d] | 457[d] | 476[d] | 518[d] | 474[d] | 470[d] |
| **Mathematics scores** | | | | | | | | | | | | | | | | | | | |
| **PISA cycle** | **Spain** | **Donor country** | | | | | | | | | | | | | | | | | |
| | | **Australia** | **Austria** | **Belgium** | **Canada** | **Switzerland** | **Chile** | **Colombia** | **Czech Republic** | **Germany** | **Denmark** | **Estonia** | **Finland** | **France** | **United Kingdom** | **Greece** | **Hungary** | **Ireland** | **Iceland** | **Israel** |
| 2009 | 483 | 514*** | a | 515*** | 527*** | 534*** | 421*** | b | 493*** | 513*** | 503*** | 512*** | 541*** | 497*** | 492*** | 466*** | 490*** | 487** | 507*** | 447*** |
| 2012 | 484 | 504*** | 506*** | 515*** | 518*** | 531*** | 423*** | b | 499*** | 514*** | 500*** | 521*** | 519*** | 495*** | 494*** | 453*** | 477*** | 501*** | 493*** | 466*** |
| 2015 | 486 | 494*** | 497*** | 507*** | 516*** | 521*** | 423*** | c | 492*** | 506*** | 511*** | 520*** | 511*** | 493*** | 492*** | 454*** | 477*** | 504*** | 488 | 470*** |
| 2018 | 481 | 491*** | 499*** | 508*** | 512*** | 515*** | 417*** | 391*** | 499*** | 500*** | 509*** | 523*** | 507*** | 495*** | 502*** | 451*** | 481 | 500*** | 495*** | 463*** |
| **Science scores** | | | | | | | | | | | | | | | | | | | |
| **PISA cycle** | **Spain** | **Donor country** | | | | | | | | | | | | | | | | | |
| | | **Australia** | **Austria** | **Belgium** | **Canada** | **Switzerland** | **Chile** | **Colombia** | **Czech Republic** | **Germany** | **Denmark** | **Estonia** | **Finland** | **France** | **United Kingdom** | **Greece** | **Hungary** | **Ireland** | **Iceland** | **Israel** |
| 2009 | 488 | 527** | a | 507*** | 529*** | 517*** | 447*** | b | 500*** | 520*** | 499*** | 528*** | 554*** | 498*** | 514*** | 470*** | 503*** | 508*** | 496*** | 455*** |
| 2012 | 496 | 521*** | 506*** | 505*** | 525*** | 515*** | 445*** | b | 508*** | 524*** | 498 | 541*** | 545*** | 499 | 514*** | 467*** | 494 | 522*** | 478*** | 470*** |
| 2015 | 493 | 510*** | 495 | 502*** | 528*** | 506*** | 447*** | c | 493 | 509*** | 502*** | 534*** | 531*** | 495 | 509*** | 455*** | 477*** | 503*** | 473*** | 467*** |
| 2018 | 483 | 503*** | 490*** | 499*** | 518*** | 495*** | 444*** | 413*** | 497*** | 503*** | 493*** | 530*** | 522*** | 493*** | 505*** | 452*** | 481 | 496*** | 475*** | 462*** |

Table 4. Mean scores for PISA 2009-2018 and test of mean differences compared to Spain (continued)

| PISA cycle | Spain | Donor country | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reading scores** | | | | | | | | | | | | | | | | | | |
| | | Italy | Japan | Korea | Lithuania | Luxembourg | Latvia | Mexico | Netherlands | Norway | New Zealand | Poland | Portugal | Slovak Republic | Slovenia | Sweden | Turkey | United States |
| 2009 | 481 | 486*** | 520*** | 539*** | b | 472*** | c | 425*** | 508*** | 503*** | 521*** | 500*** | 489*** | 477** | 483 | 497*** | 464*** | 500*** |
| 2012 | 488 | 490 | 538*** | 536*** | b | 488 | c | 424*** | 511*** | 504*** | 512*** | 518*** | 488 | 463*** | 481*** | 483** | 475*** | 498*** |
| 2015 | 496 | 485*** | 516*** | 517*** | b | 481*** | 488*** | 423*** | 503*** | 513*** | 509*** | 506*** | 498 | 453*** | 505*** | 500*** | 428*** | 497 |
| 2018 | tbi | 476[d] | 504[d] | 514[d] | 476[d] | 470[d] | 479[d] | 420[d] | 485[d] | 499[d] | 506[d] | 512[d] | 492[d] | 458[d] | 495[d] | 506[d] | 466[d] | 505[d] |
| **Mathematics scores** | | | | | | | | | | | | | | | | | | |
| **PISA cycle** | **Spain** | **Donor country** | | | | | | | | | | | | | | | | |
| | | **Italy** | **Japan** | **Korea** | **Lithuania** | **Luxembourg** | **Latvia** | **Mexico** | **Netherlands** | **Norway** | **New Zealand** | **Poland** | **Portugal** | **Slovak Republic** | **Slovenia** | **Sweden** | **Turkey** | **United States** |
| 2009 | 483 | 483 | 529*** | 546*** | b | 489*** | c | 419*** | 526*** | 498*** | 519*** | 495*** | 487** | 497*** | 501*** | 494*** | 445*** | 487** |
| 2012 | 484 | 485 | 536*** | 554*** | b | 490*** | c | 413*** | 523*** | 489*** | 500*** | 518*** | 487 | 482 | 501*** | 478*** | 448*** | 481* |
| 2015 | 486 | 490** | 532*** | 524*** | b | 486 | 482** | 408*** | 512*** | 502*** | 495*** | 504*** | 492*** | 475*** | 510*** | 494*** | 420*** | 470*** |
| 2018 | 481 | 487*** | 527*** | 526*** | 481 | 483 | 496*** | 409*** | 519*** | 501*** | 494*** | 516*** | 492*** | 486*** | 509*** | 502*** | 454*** | 478** |
| **Science scores** | | | | | | | | | | | | | | | | | | |
| **PISA cycle** | **Spain** | **Donor country** | | | | | | | | | | | | | | | | |
| | | **Italy** | **Japan** | **Korea** | **Lithuania** | **Luxembourg** | **Latvia** | **Mexico** | **Netherlands** | **Norway** | **New Zealand** | **Poland** | **Portugal** | **Slovak Republic** | **Slovenia** | **Sweden** | **Turkey** | **United States** |
| 2009 | 488 | 489 | 539*** | 538*** | b | 484** | c | 416*** | 522*** | 500*** | 532*** | 508*** | 493*** | 490 | 512*** | 495*** | 454*** | 502*** |
| 2012 | 496 | 494** | 547*** | 538*** | b | 491*** | c | 415*** | 522*** | 495 | 516*** | 526*** | 489 | 471*** | 514*** | 485*** | 463*** | 497 |
| 2015 | 493 | 481*** | 538*** | 516*** | b | 483*** | 490 | 416*** | 509*** | 498*** | 513*** | 501*** | 501*** | 461*** | 513*** | 493 | 425*** | 496** |
| 2018 | 483 | 468*** | 529*** | 519*** | 482 | 477*** | 487*** | 419*** | 503*** | 490*** | 508*** | 511*** | 492*** | 464*** | 507*** | 499*** | 468*** | 502*** |

Notes: [a] Austria does not have reading, mathematics and science scores due to a dispute between teacher unions and the education minister, which led to a boycott and a negative atmosphere which affected the conditions under which the assessment was administered, so they were not reliable (OECD, 2010; Annex A4). [b] The country did not participate in that cycle. [c] The country had scores but was not in the OECD in that cycle. [d] Indicates that, although the mean score for reading has been reported, the test of mean differences has not been performed, as there is not any available information for the reading scores of Spain in PISA 2018. "tbi" means "to be imputed". PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Imputation method: multiple imputation with 10 complete estimations has been employed.
Significance of the test of mean differences: ***Significant differences at 1%, ** significant at 5%, * significant at 10%.
Source: Authors' own calculations.

Table A1. Multiple imputations of Spanish reading scores for PISA 2018 using reading, mathematics, science plausible values and sex and ESCS index quartile (using data from different "donor" countries)

| Donor country | Imputed reading score for Spain |
|---|---|
| Chile | 503 |
| Ireland | 502 |
| Colombia | 495 |
| United States | 495 |
| Mexico | 493 |
| Greece | 492 |
| Israel | 492 |
| Norway (top quartile) | 486 |
| Sweden (top quartile) | 486 |
| Australia | 485 |
| Denmark | 485 |
| United Kingdom | 485 |
| Finland | 484 |
| Canada | 483 |
| Italy | 483 |
| New Zealand | 483 |
| Turkey | 483 |
| Iceland (median) | 482 |
| Portugal (median) | 482 |
| France | 481 |
| Poland | 480 |
| Germany | 479 |
| Hungary | 477 |
| Korea | 477 |
| Czech Republic | 476 |
| Estonia (bottom quartile) | 475 |
| Lithuania (bottom quartile) | 475 |
| Austria | 474 |
| Belgium | 474 |
| Luxembourg | 474 |
| Latvia | 470 |
| Slovak Republic | 470 |
| Slovenia | 470 |
| Switzerland | 466 |
| Netherlands | 459 |
| Japan | 456 |
| **All countries as donors** | **481** |

Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Imputation method: multiple imputation with 10 complete estimations has been employed. Spanish reading scores have been considered as missing completely at random.
Source: Authors' own calculations.

Table A2. Multiple imputations of Spanish reading scores for PISA 2009, 2012 and 2015 using reading, mathematics, science plausible values and sex and ESCS index quartile (using data from different "donor" countries)

| PISA 2009 | | PISA 2012 | | PISA 2015 | |
|---|---|---|---|---|---|
| Austria | a | Colombia | b | Lithuania | b |
| Colombia | b | Lithuania | b | Colombia | c |
| Lithuania | b | Latvia | c | Ireland | 510 |
| Latvia | c | Israel | 509 | Norway | 507 |
| Israel | 508 | Greece | 506 | Chile | 504 |
| Greece | 500 | Turkey | 505 | Greece | 504 |
| Chile | 498 | Ireland | 503 | Israel | 501 |
| Turkey | 496 | Norway | 503 | Mexico | 500 |
| Mexico | 495 | France | 498 | Sweden | 498 |
| Korea | 493 | United States | 498 | Poland (top quartile) | 497 |
| Norway | 491 | Mexico | 496 | United States (top quartile) | 497 |
| Sweden (top quartile) | 490 | New Zealand (top quartile) | 495 | Finland | 496 |
| Iceland | 489 | Hungary | 492 | France | 496 |
| United States | 489 | Sweden | 492 | Iceland | 496 |
| Belgium | 484 | Belgium | 491 | Official Spain score | 496 |
| Italy | 484 | Canada | 491 | Korea | 495 |
| Poland | 484 | Chile (median) | 490 | Canada | 493 |
| Portugal | 484 | Denmark (median) | 490 | Germany | 493 |
| Denmark (median) | 483 | Italy (median) | 490 | Italy | 492 |
| France (median) | 483 | Portugal (median) | 490 | Denmark (median) | 491 |
| Ireland (median) | 483 | Australia | 489 | New Zealand (median) | 491 |
| Hungary | 482 | Luxembourg | 489 | Turkey (median) | 491 |
| New Zealand | 482 | Poland | 489 | Latvia | 490 |
| Canada | 481 | Japan | 488 | Luxembourg | 490 |
| Official Spain score | 481 | Official Spain score | 488 | Portugal | 490 |
| Australia | 478 | United Kingdom | 488 | Belgium | 488 |
| Finland | 477 | Iceland | 487 | Australia (bottom quartile) | 487 |
| United Kingdom | 477 | Korea | 487 | Czech Republic (bottom quartile) | 487 |
| Netherlands (bottom quartile) | 475 | Finland (bottom quartile) | 484 | United Kingdom (bottom quartile) | 487 |
| Luxembourg | 474 | Czech Republic | 483 | Slovenia | 485 |
| Japan | 472 | Germany | 482 | Hungary | 484 |
| Slovak Republic | 472 | Netherlands | 482 | Netherlands | 484 |
| Czech Republic | 470 | Slovak Republic | 482 | Estonia | 483 |
| Switzerland | 470 | Estonia | 479 | Austria | 481 |
| Estonia | 469 | Switzerland | 477 | Slovak Republic | 478 |
| Germany | 469 | Austria | 476 | Switzerland | 476 |
| Slovenia | 462 | Slovenia | 464 | Japan | 474 |
| **All countries as donors** | **482** | **All countries as donors** | **490** | **All countries as donors** | **492** |

Notes: [a] Austria does not have reading, mathematics and science scores due to a dispute between teacher unions and the education minister, which led to a boycott and a negative atmosphere which affected the conditions under which the assessment was administered, so they were not reliable (OECD, 2010; Annex A4). [b] The country did not participate in that cycle. [c] The country had scores but was not in the OECD in that cycle. PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).

Imputation method: multiple imputation with 10 complete estimations has been employed. Spanish reading scores have been considered as missing completely at random.

Source: Authors' own calculations.

Table B1. Multiple imputations of Spanish reading scores by regions, using as donors all OECD countries

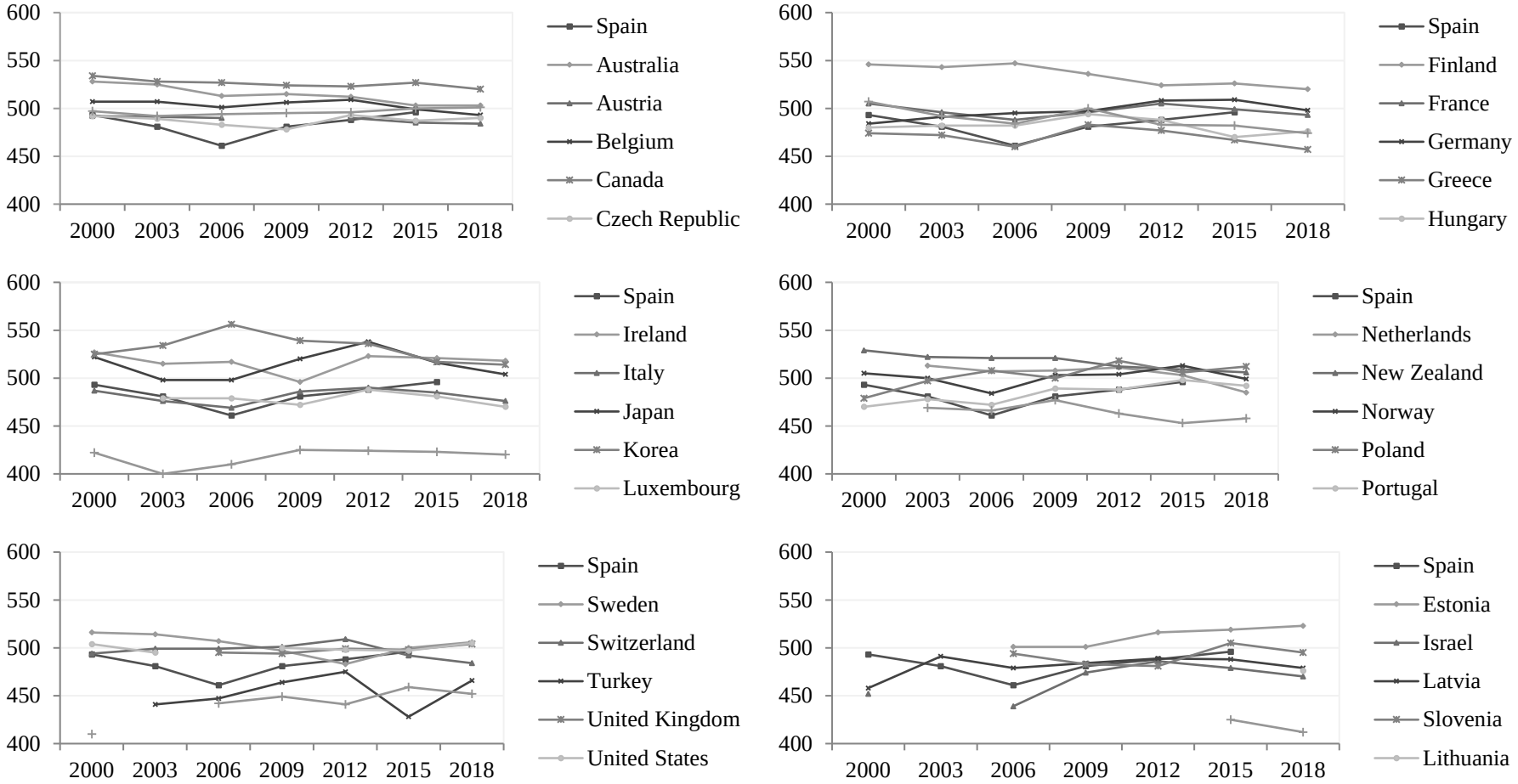
| Region | PISA cycle | Official reading scores for the region | Imputation using plausible values in reading, mathematics and science | Imputation using plausible values in reading, mathematics and science, sex and ESCS index quartile |
|---|---|---|---|---|
| **Andalusia** | 2009 | 461 | 465 | 463 |
| | 2012 | 477 | 481 | 479 |
| | 2015 | 479 | 473 | 473 |
| | 2018 | tbi | 470 | 468 |
| **Aragon** | 2009 | 495 | 500 | 500 |
| | 2012 | 493 | 499 | 499 |
| | 2015 | 506 | 505 | 504 |
| | 2018 | tbi | 493 | 494 |
| **Asturias** | 2009 | 490 | 495 | 493 |
| | 2012 | 504 | 508 | 508 |
| | 2015 | 498 | 499 | 499 |
| | 2018 | tbi | 494 | 494 |
| **Balearic Islands** | 2009 | 457 | 460 | 460 |
| | 2012 | 476 | 480 | 480 |
| | 2015 | 485 | 484 | 484 |
| | 2018 | tbi | 481 | 481 |
| **Basque Country** | 2009 | 494 | 493 | 495 |
| | 2012 | 498 | 503 | 504 |
| | 2015 | 491 | 484 | 485 |
| | 2018 | tbi | 489 | 489 |
| **Canary Islands** | 2009 | 448 | 447 | 444 |
| | 2012 | a | a | a |
| | 2015 | 483 | 473 | 472 |
| | 2018 | tbi | 466 | 465 |
| **Cantabria** | 2009 | 488 | 494 | 493 |
| | 2012 | 485 | 496 | 495 |
| | 2015 | 501 | 495 | 494 |
| | 2018 | tbi | 495 | 496 |
| **Castile and Leon** | 2009 | 503 | 509 | 510 |
| | 2012 | 505 | 512 | 512 |
| | 2015 | 522 | 515 | 515 |
| | 2018 | tbi | 501 | 500 |
| **Castile La Mancha** | 2009 | a | a | a |
| | 2012 | a | a | a |
| | 2015 | 499 | 495 | 495 |
| | 2018 | tbi | 482 | 481 |
| **Catalonia** | 2009 | 498 | 492 | 492 |
| | 2012 | 501 | 490 | 491 |
| | 2015 | 500 | 502 | 502 |
| | 2018 | tbi | 488 | 488 |
| **Ceuta** | 2009 | 423 | 419 | 418 |
| | 2012 | a | a | a |
| | 2015 | a | a | a |
| | 2018 | tbi | 412 | 411 |
| **Extremadura** | 2009 | a | a | a |
| | 2012 | 457 | 475 | 473 |
| | 2015 | 475 | 475 | 474 |
| | 2018 | tbi | 471 | 471 |
| **Galicia** | 2009 | 486 | 497 | 495 |
| | 2012 | 499 | 501 | 500 |
| | 2015 | 509 | 508 | 508 |
| | 2018 | tbi | 507 | 507 |
| **La Rioja** | 2009 | 498 | 502 | 501 |
| | 2012 | 490 | 504 | 506 |
| | 2015 | 491 | 498 | 498 |
| | 2018 | tbi | 488 | 488 |
| **Madrid** | 2009 | 503 | 499 | 499 |
| | 2012 | 511 | 509 | 510 |
| | 2015 | 520 | 512 | 512 |
| | 2018 | tbi | 485 | 485 |
| **Melilla** | 2009 | 399 | 419 | 418 |
| | 2012 | a | a | a |
| | 2015 | a | a | a |
| | 2018 | tbi | 435 | 435 |
| **Murcia** | 2009 | 480 | 479 | 478 |

| | | | | |
|---|---|---|---|---|
| | 2012 | 462 | 474 | 472 |
| | 2015 | 486 | 483 | 482 |
| | 2018 | tbi | 476 | 476 |
| **Navarra** | 2009 | 497 | 504 | 503 |
| | 2012 | 509 | 512 | 513 |
| | 2015 | 514 | 510 | 511 |
| | 2018 | tbi | 494 | 494 |
| **Valencian Community** | 2009 | a | a | a |
| | 2012 | a | a | a |
| | 2015 | 499 | 492 | 492 |
| | 2018 | tbi | 476 | 476 |

Notes: [a] The region does not have representative information for that cycle. "tbi" means "to be imputed". PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009). Official reading scores for the region were retrieved from MECD (2010, 2014, 2016, 2019).
Imputation method: multiple imputation with 10 complete estimations has been employed. Spanish reading scores have been considered as missing completely at random.
Source: Authors' own calculations.

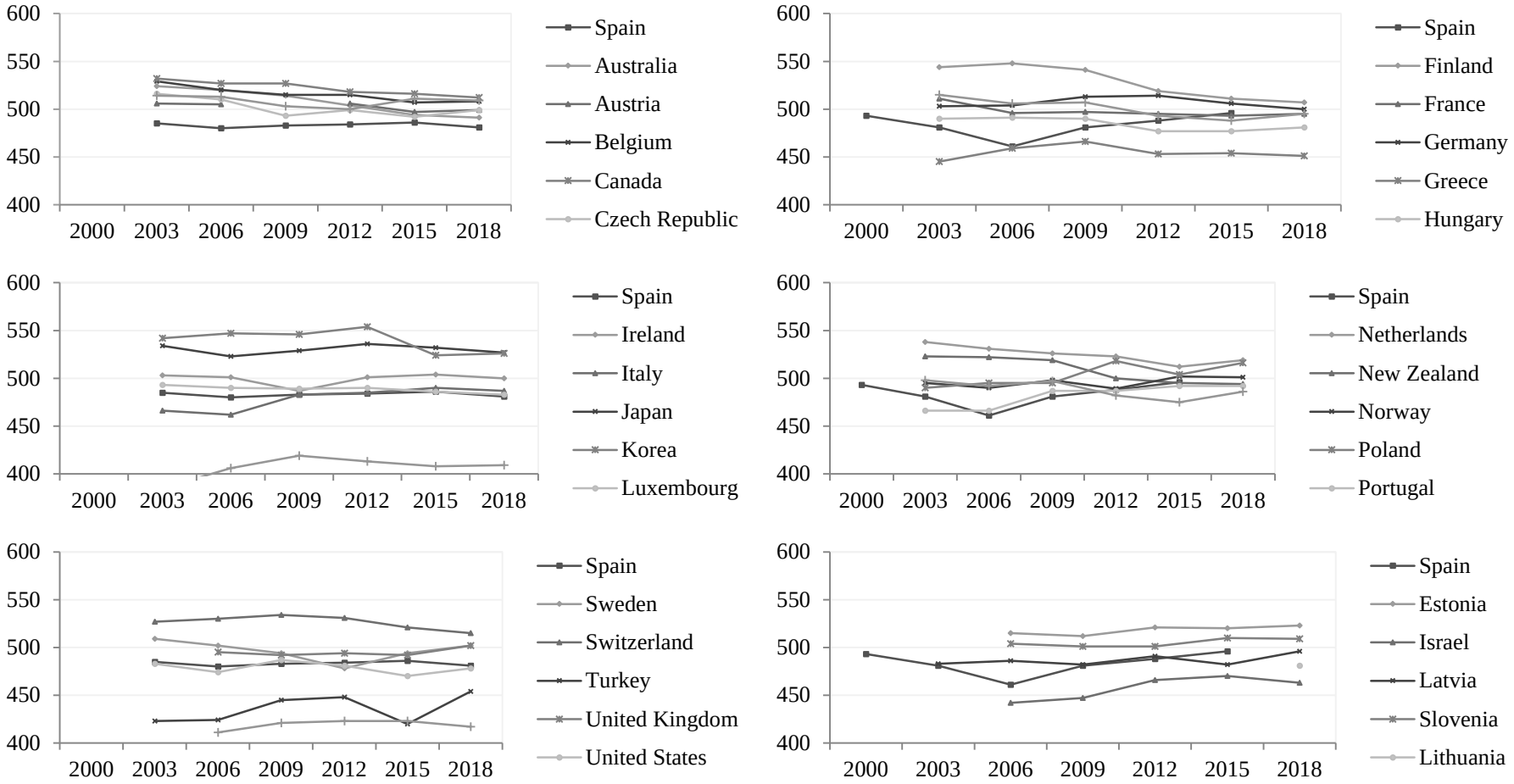
**Appendix C**

Figure C1. Trends in OECD mean reading scores in PISA from 2000 to 2018



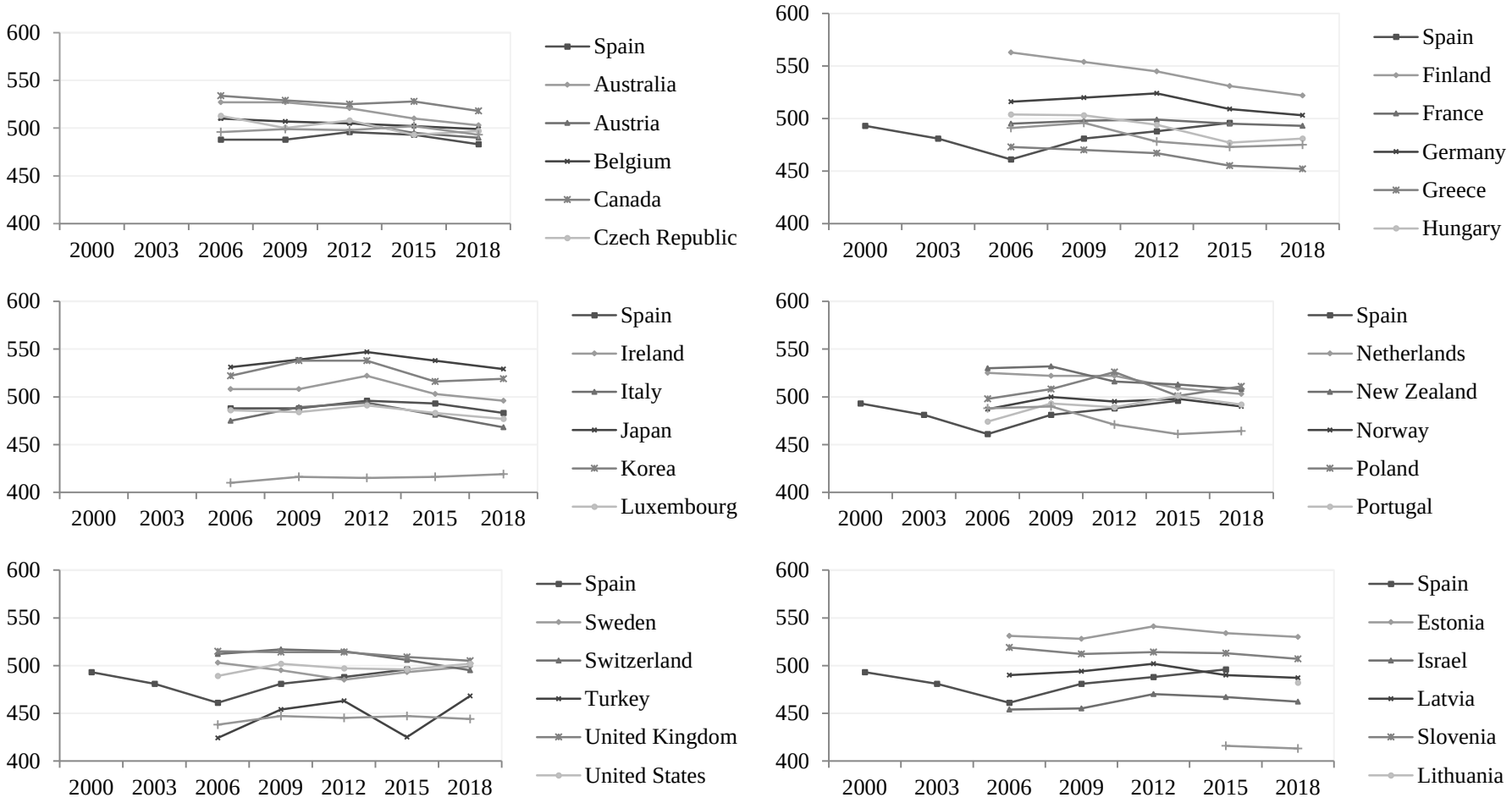

Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Source: Authors' own calculations from PISA data (OECD, 2020b).

Figure C2. Trends in OECD mean mathematics scores in PISA from 2000 to 2018




Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Source: Authors' own calculations from PISA data (OECD, 2020b).

Figure C3. Trends in OECD mean science scores in PISA from 2000 to 2018




Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2009).
Source: Authors' own calculations from PISA data (OECD, 2020b).