Does Paid Parental Leave Affect Children's Schooling Outcomes? Replicating Danzer and Lavy (2018)

Claudia Troccoli*

August 12, 2022

Abstract

Danzer and Lavy (2018) study how the duration of paid parental leave affects children's educational performance using data from PISA. An extension of the maximum duration from 12 to 24 months in Austria had no significant effect on average, but the authors highlight the existence of large and significant heterogenous effects that vary in sign depending on the education of mothers and children's gender. The policy increased the scores obtained by sons of highly educated mothers, as measured in standard deviations, by 0.33 in Reading and 0.40 in Science. On the contrary, sons of low educated mothers experienced a decrease of 0.27 in Reading and 0.23 in Science. In this article, I replicate their study following the recommended estimation procedure that takes into account that PISA relies on imputation to derive student scores. I show that the estimates of the effects of the parental leave extension become substantially smaller and non-significant.

*Aalto University, Department of Economics; email: claudia.troccoli@aalto.fi; phone: +358 405434583. The data and do files used in this replication are available at: `https://sites.google.com/view/claudiatroccoli`

1 Introduction

Danzer and Lavy (2018) (hereafter, DL) study the effects of a parental leave extension on children's schooling outcomes in Austria. While parents of children born before the 1st of July 1990 could take leave only until the child's first birthday, those giving birth afterwards were entitled to an additional year of paid and job-protected parental leave. DL compare the scores obtained at age 15 by children born before and after the 1st of July 1990, and account for potential month-of-birth effects by using information on a previous cohort that was not affected by the reform. They find no statistically significant overall effect of the additional year of leave. However, the authors emphasize the existence of strong heterogenous effects depending on children's gender and maternal education. They find that sons of highly educated mothers benefited greatly from the reform. Their scores are 0.33 standard deviations higher in reading (standard error=0.15) and 0.40 standard deviations higher in science (st. error=0.11). On the contrary, sons of lower educated mothers were harmed by the reform. They experience a decrease of 0.27 standard deviations in reading (st. error=0.13) and 0.23 in science (st. error=0.13). Given that the authors estimate the intention-to-treat effect of the policy in a context without full compliance, these estimates should be interpreted as the lower bound of the actual impact of parental leave. In contrast with DL findings, previous studies on the impact of paid parental leave tend to find considerably weaker effects for the children of highly educated mothers and non-negative impacts for the children of lower educated mothers.[1]

DL analyze data from PISA, a source which is widely used by economists to

[1] A number of previous papers have considered heterogeneous effects by maternal education. Liu and Skans (2010) find that an expansion of parental leave from 12 to 15 months in Sweden improved grades at age 16 of children of highly educated mothers by 0.05 standard deviations. They observe smaller and statistically insignificant effects for children of lower educated mothers. Carneiro et al. (2015) study the introduction of 4 months of paid maternity leave and 12 months of unpaid leave in Norway. The policy reduced high-school dropout rates and increased college attendance for children of both highly and lower educated mothers. Albagli and Rau (2019) find that a 3-month extension of maternity leave in Chile improved cognitive abilities of preschool-aged children by around 20% of a standard deviations, especially for children of less educated mothers.

study individual-level differences in schooling performance both within and across countries. As I explain in more detail below, the analysis of PISA data requires a specific procedure to address the complexity of the survey and the test design. As happens with other international large-scale assessments such as TIMSS, PIAAC and PIRLS, due to imputed values and stratified sampling, PISA provides five different plausible values for each score, each one representing a random draw from the posterior distribution, which need to be taken into account in the estimation (OECD, 2009). Studies that do not apply the appropriate statistical procedure are likely to underestimate standard errors and, when the sample size is relatively small, point estimates might also be greatly affected. Unfortunately, many studies in Economics, including DL, fail to implement the methodology recommended by PISA. As I show in the working paper version of this replication, in a sample of 56 papers using data from international large-scale assessments that were published between 2000 and 2019 in top economic journals, less than half clearly mention following the recommended procedure (Troccoli, 2020).[2]

DL do not take into account the imputed nature of the data and they consider in their regressions only one plausible value. In this replication, I show that, when the recommended method is applied, their main point estimates become significantly closer to zero, standard errors are larger, and none of the main estimates is anymore significant at standard levels. My analysis contributes to illustrate the relevance of following the appropriate procedure in the analysis of PISA data. Jerrim et al. (2017) offer an excellent discussion of the requirements for the analysis of data with imputed values. As an example, they analyze Lavy (2015), who fails to apply the recommended methodology and, as a result, underestimates standard errors. However, in this particular case, due to the large sample size, using the recommended

[2]I consider articles published in American Economic Review, Economics of Education Review, Journal of Labor Economics, Journal of Political Economy, Journal of Population Economics, Labour Economics and The Economic Journal. I presume that authors have not used the recommended procedure whenever they do not mention it explicitly in the paper.

procedure does not affect substantially the main findings of the paper.

2 Estimation Procedure with International Large-Scale Assessments

As explained in the PISA Data Analysis Manual (OECD, 2009) and discussed in great detail by Jerrim et al. (2017), there are three crucial features of international large-scale assessments datasets that need to be taken into account for statistical analysis. First, the test scores reported in the datasets are not raw scores, but rather the result of imputation. In the case of PISA, the assessment aims at evaluating students on a vast set of skills in mathematics, science, reading and collaborative problem solving in many different countries. Test-takers do not answer the full set of questions, but only a small subsample. Each student is randomly allocated to one of thirteen test booklets, containing a subset of questions either on all the subjects, only one, or a combination of two or three. Due to the random assignment of booklets, the questions not answered by each student can be considered as Missing Completely At Random. Therefore, multiple imputation is applied to obtain five plausible values (PVs) for each subject. These PVs are random draws from a posterior distribution and are computed using information on students' performance and characteristics, including the school attended and the average scores obtained by the other pupils. PVs provide information on the final score that the student would have obtained, had they been tested on the full pool of questions.

Statistical analysis with PVs requires a specific procedure following a modified version of Rubin's rule for multiple imputation. The equation of interest is estimated five times, with each PV as outcome variable. The averages of the five parameters and the five sampling error estimates are, respectively, the final parameter (β_*) and the final sampling error (σ_*), and from the latter the standard error can be

calculated.[3] The plausible values for each individual are not perfectly correlated. For instance, in the sample used by DL, the correlation between the five plausible values is around 80%. If authors consider only one PV and additionally ignore the recommended procedure to adjust standard errors, these are underestimated, which may artificially inflate the statistical significance of the estimates.

Second, the procedure recommended by the survey organisers also requires the use of sampling weights. Within each country, participating individuals are selected based on stratified sampling, reflecting the population distribution, geographically and in terms of other characteristics. A student weight is assigned to each student, to rescale the sample to the size of the population within each country, taking into account the stratum each student is drawn from.[4] The inclusion of individual student weights allows to generalize to the overall population of potential test takers.

Finally, the statistical analysis of international large-scale assessments datasets also requires the use of Balanced Repeated Replication (BRR) weights. These weights adjust for uncertainty with regards to sampling by taking into account the two-stage stratification design, whereby schools are selected and students are randomly drawn from each school.

3 Replication

To replicate the analysis of DL, I use the original PISA data from 2003 and 2006 provided by OECD and, following DL, I restrict the sample to students born in Austria between May and August in 1987 and 1990, enrolled in "regular" academic and

[3]The magnitude of the imputation error δ_* can then be calculated with the formula $\delta_* = \frac{\sum_{pv=1}^{5}(\beta_{pv}-\beta_*)^2}{n_{pv}-1}$. The final standard error is equal to $\sqrt{\sigma_*^2 + (1+\frac{1}{PV})\cdot\delta_*^2}$.

[4]For PISA, around 30 students are randomly picked within each school. The choice of participating schools is also random - with probability proportional to each school's size - within explicit strata (schools in the same region and type). Additionally, in order to limit the bias caused by non-response, each selected school is assigned two substitute schools based on characteristics that are expected to be correlated with PISA scores (implicit stratification).

vocational schools, and for which mother's education is known. The resulting sample is practically identical to the one in DL. My sample includes 2860 observations, compared to 2840 in DL. All observations in DL are also in my sample.[5]

I consider exactly the same specification used by DL:

$$y_i = \alpha + \beta_1 Post\_June_i + \beta_2 bc1990_i + \beta_3 Post\_June_i \times bc1990_i + \mathbf{Birth\_month_i}\theta_\mathbf{m} + \mathbf{X_i}\mu + \epsilon_i \quad (1)$$

where y_i is the child i's score in mathematics, reading or science; $Post\_June$ is a dummy variable indicating that the child was born between 1st of July and 31st of August, and zero if the child was born between May 1st and June 30th; $bc1990$ is a dummy indicator for the birth cohort 1990; $Birth\_month$ is a set of dummy variables indicating the month of birth, and X includes the set of background controls considered by DL (mother's and father's educational attainment, school location and migration background). β_3 is the coefficient of interest measuring the intention-to-treat effect of the reform. Following DL, standard errors are clustered by school programme, school location and gender.[6] DL estimate this regression using as outcome variable the first plausible value (PV1) listed for each subject. They use student weights but not BRR weights. First, I replicate their analysis using the same estimation procedure and outcome variable. As shown in Table 1, my results are practically identical to DL in all five subsamples.[7] Similar to DL, the impact of parental leave is not significant for the overall sample (panel 1), for daughters of highly educated mothers (panel 4) or for daughters of lower educated mothers (panel 5), but there is a large and significant positive effect on the performance of sons of highly educated mothers in Science and Reading (panel 2, columns 2 and

[5]The two sample sizes might differ slightly because of the time at which the datesets were downloaded from the OECD website. It might be that some observations were missing altogether then and were added afterwards, or that they were previosuly excluded by DL due to missing information but were later updated.

[6]It is outside of the scope of this replication to investigate whether this is the appropriate level of clustering.

[7]These results are reported in Table 3 (p. 101) and 4 (p. 104) of DL.

3) and a large and significant negative impact for sons of lower educated mothers (panel 3, columns 2 and 3).

Next, I explore how results change when the other four plausible values reported by PISA are used as outcome variable (see rows 3-6 of each panel). In contrast with the initial estimates, point estimates tend to be closer to zero and none of these estimates is significant at standard levels. For instance, while according to PV1 parental leave increases the Math score of sons of highly educated mothers by 0.40 (st. error=0.11), using PV2-PV4 the estimate ranges between 0.12 and 0.22 and is never statistically significant.

In the following row of each panel, titled *PV1-PV5 (Rubin)*, I report the coefficients and standard errors calculated taking into account all plausible values, using Rubin's formula. Compared to the estimates reported in DL, the point estimates tend to be closer to zero and standard errors are larger. None of the estimates is significant.[8]

Finally, in the last row of each panel, titled *PV1-PV5 & BRR*, I report the coefficients and standard errors which are obtained following the recommended procedure, i.e. applying Rubin's rule for multiple imputation using the estimated coefficients and standard errors from the five estimations, and taking into account both individual student weights and the BRR weights. As expected, the coefficients are identical to the ones in the previous row and standard errors are slightly larger. Again, none of the estimates is statistically different from zero and, overall, they tend to be imprecise and uninformative. For instance, while DL conclude that an additional year of parental leave increases by 0.33 the reading scores of sons of highly educated mothers (CI: 0.04, 0.62), I cannot exclude negative effects of up to 0.33 or positive effects of up to 0.58. Similarly, while DL find an increase of 0.40 (CI: 0.18, 0.63) in science for sons of highly educated mothers, I cannot reject negative effects of up to

[8]For the anlysis, I use the Stata package PV (Macdonald (2019)).

0.21 or positive ones of up to 0.64.

In sum, when I use the recommended procedure there is a substantial increase in standard errors and a large decrease in the magnitude of point of estimates that were statistically significant in DL's analysis. While the increase in standard errors is unsurprising, the systematic decrease in the magnitude of point estimates is more intriguing. The standard errors increase because, by using a single plausible value and not taking into account the BRR weights, DL ignore the uncertainty generated by the imputation and sampling design. However, the systematic decrease in the magnitude of (statistically significant) point estimates is less obvious. In principle, using a single plausible value should provide unbiased point estimates. The problem may arise when journals (and authors) have a preference for papers with estimates that are statistically significant (Chopra et al., 2022). In this case, false positives would be more likely to be published and, when more accurate estimates become available, they will tend to be closer to zero. For instance, had DL written a paper using plausible values number 2, 3, 4 or 5, they would not have obtained any significant results in any of the subsamples and dimensions they consider (and arguably they might have struggled to publish their paper), but the point estimates obtained using the recommended procedure would have been generally of similar magnitude.

Both problems, the increase in standard errors and the decrease in the magnitude of significant point estimates, are likely to be more severe when the sample size is relatively small, as happens in some of the subsamples considered by DL (e.g. N=486 in panel 2 of Table 1). Conversely, as pointed out in the PISA manual, "using one plausible value or five plausible values does not really make a substantial difference on large samples" (OECD (2009), page 46, cited by Jerrim et al. (2017), footnote 8).

4 Conclusion

In this study, I replicate DL taking into account the procedure required for the analysis of datasets such as PISA which use imputed data and stratified sampling. When all plausible values are considered and the appropriate weights are taken into account, the main significant effects identified by DL disappear. The point estimates are reduced in size, standard errors are larger, and none of the main coefficients is statistically significant at standard levels.

It would be unfair to single out DL for making this methodological mistake. An analysis of 56 articles that were published in top economic journals during the last two decades indicates that a majority of economists using this type of data fail to use the recommended procedure (Troccoli, 2020). In many of these papers, the statistical power of the analysis is relatively large and whether authors use or not the right procedure is unlikely to significantly affect their findings. The large impact that the correction has on DL's results probably reflects the lack of power of their study. Their standard errors are disproportionately large compared to the magnitude of the effects that have been detected in the literature, implying that the signal-to-noise ratio of the estimation is likely to be low. In this respect, my findings support the view held by some authors who have argued that social scientists should be more cautious in their interpretation of empirical evidence in contexts where statistical power is limited, an analysis plan has not been pre-specified, and standard errors are not adjusted for multiple testing (e.g. Ioannidis et al. (2017); Gelman and Loken (2013); Maniadis et al. (2014)). As these authors have pointed out, such estimates risk being uninformative, independently of whether they are statistically significant or not.

References

Albagli, P. and T. Rau (2019). The effects of a maternity leave reform on children's abilities and maternal outcomes in chile. *The Economic Journal 129*(619), 1015–1047.

Carneiro, P., K. V. Løken, and K. G. Salvanes (2015). A Flying Start? Maternity Leave Benefits and Long-Run Outcomes of Children. *Journal of Political Economy 123*(2), 365–412.

Chopra, F., I. Haaland, C. Roth, and A. Stegmann (2022). The null result penalty.

Danzer, N. and V. Lavy (2018). Paid Parental Leave and Children's Schooling Outcomes. *The Economic Journal 128*(608), 81–117.

Gelman, A. and E. Loken (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

Ioannidis, J. P., T. D. Stanley, and H. Doucouliagos (2017). The power of bias in economics research.

Jerrim, J., L. A. Lopez-Agudo, O. D. Marcenaro-Gutierrez, and N. Shure (2017). What Happens When Econometrics and Psychometrics Collide? An Example Using the PISA Data. *Economics of Education Review 61*, 51–58.

Lavy, V. (2015). Do Differences in Schools' Instruction Time Explain International Achievement Gaps? evidence from Developed and Developing Countries. *The Economic Journal 125*(588), F397–F424.

Liu, Q. and O. N. Skans (2010). The Duration of Paid Parental Leave and Children's Scholastic Performance. *The BE Journal of Economic Analysis & Policy 10*(1).

Macdonald, K. (2019). PV: Stata Module to Perform Estimation with Plausible Values.

Maniadis, Z., F. Tufano, and J. A. List (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review 104*(1), 277–90.

OECD (2009). PISA Data Analysis Manual: SPSS.

Troccoli, C. (2020). Comment on "Paid Parental Leave and Children's Schooling Outcomes". *Available at SSRN 3744650*.

Table 1: Estimated Impact of Parental Leave, % of a Standard Deviation.

	MATHEMATICS	READING	SCIENCE
	(1)	(2)	(3)
Panel 1, full sample, N=2860			
PV1 (DL, N=2840)	2.00	-4.14	2.10
	(6.70)	(8.09)	(7.41)
PV1	1.85	-3.81	2.01
	(6.62)	(7.99)	(7.31)
PV2	1.82	-5.85	0.52
	(6.83)	(8.11)	(7.66)
PV3	1.26	-4.90	1.26
	(7.73)	(8.66)	(7.82)
PV4	-1.30	-6.36	-0.95
	(7.14)	(7.06)	(6.62)
PV5	-3.08	-9.40	-1.15
	(6.63)	(7.71)	(7.03)
PV1-PV5 (Rubin)	0.11	-6.06	0.34
	(7.13)	(7.36)	(6.80)
PV1-PV5 & BRR	0.11	-6.06	0.34
	(9.30)	(9.23)	(8.91)
Panel 2, sons of high education mothers, N=484			
PV1 (DL, N=482)	15.83	33.12**	40.40***
	(12.28)	(14.99)	(11.45)
PV1	15.78	33.10**	40.38***
	(12.31)	(15.00)	(11.45)
PV2	15.46	5.67	12.47
	(13.25)	(14.69)	(13.61)
PV3	16.11	10.30	21.71
	(14.35)	(16.03)	(12.96)
PV4	6.35	13.24	19.99
	(13.26)	(16.83)	(15.92)
PV5	14.49	0.57	12.45
	(13.72)	(14.75)	(13.54)
PV1-PV5 (Rubin)	13.64	12.58	21.40
	(18.02)	(22.98)	(21.38)
PV1-PV5 & BRR	13.64	12.58	21.40
	(18.73)	(23.27)	(21.48)
Panel 3, sons of low education mothers: N=944			
PV1 (DL, N=944)	-9.03	-26.63**	-23.25*
	(11.77)	(12.87)	(13.38)
PV1	-9.03	-26.63**	-23.25*
	(11.77)	(12.87)	(13.38)
PV2	-2.30	-24.73*	-20.49
	(10.48)	(12.32)	(12.99)
PV3	-12.28	-18.72	-18.97
	(12.09)	(13.00)	(13.12)
PV4	-13.77	-15.31	-18.10
	(10.86)	(11.00)	(11.94)
PV5	-16.51*	-21.85	-25.30*
	(9.65)	(13.20)	(13.14)
PV1-PV5 (Rubin)	-10.78	-21.45	-21.22*
	(13.00)	(13.31)	(12.02)
PV1-PV5 & BRR	-10.78	-21.45	-21.22
	(16.41)	(16.45)	(15.61)

Continued on next page...

Panel 4, daughters of high education mothers, N=468			
PV1 (DL, N=461)	16.00	13.91	6.33
	(15.18)	(19.08)	(15.82)
PV1	16.03	14.38	6.16
	(14.90)	(18.76)	(15.60)
PV2	4.68	22.38	18.11
	(14.65)	(18.41)	(16.57)
PV3	6.83	-2.80	-8.28
	(16.39)	(18.96)	(15.53)
PV4	11.88	-4.98	-2.22
	(13.74)	(14.59)	(13.28)
PV5	5.54	-1.00	-1.60
	(18.04)	(13.80)	(16.21)
PV1-PV5 (Rubin)	8.99	5.60	2.43
	(18.30)	(22.04)	(20.36)
PV1-PV5 & BRR	8.99	5.60	2.43
	(23.83)	(25.25)	(23.01)
Panel 5, daughters of low education mothers, N=964			
PV1 (DL, N=953)	-2.02	-8.91	5.82
	(13.28)	(13.98)	(13.13)
PV1	-2.91	-9.39	5.27
	(13.17)	(13.82)	(12.92)
PV2	-4.18	-8.90	5.37
	(14.45)	(15.67)	(16.68)
PV3	3.40	-1.40	13.46
	(14.83)	(15.63)	(15.84)
PV4	-0.94	-9.72	5.06
	(15.48)	(14.20)	(13.79)
PV5	-4.46	-9.48	12.98
	(13.06)	(14.86)	(13.00)
PV1-PV5 (Rubin)	-1.82	-7.78	8.43
	(11.68)	(11.99)	(11.89)
PV1-PV5 & BRR	-1.82	-7.78	8.43
	(13.60)	(13.96)	(13.05)

Notes: The table reports estimates for β_3 of Equation 1. The outcome variable standardised such that the standard deviation is 100, hence the coefficients can be interpreted as "percent of a standard deviation" (e.g. 2.00 is 2% of a st.dev. or 0.02 st.dev.). Each cell corresponds to a separate regression. The first column reports results using PISA scores in Mathematics, column 2 in Reading and column 3 in Science. Each panel presents information for a different sample of children. The first row of each panel, titled *PV1 (DL)*, presents the results reported by Danzer and Lavy (2018) in Table 3 (p. 101) and 4 (p. 104), where they estimate equation 1 using as outcome variable the first Plausible Value and accounting for individual inverse probability weights. In the second row, titled *PV1*, I report the results that I obtain when I replicate their analysis using the same outcome variable and specification, but with a slightly different sample size. Next, in rows 3-6 I report similar estimates using as outcome variables Plausible Values 2-5. Row 7, titled *PV1-PV5 (Rubin)*, provides the coefficients for the estimation that uses all five Plausible Values and the Student Weights. Finally, row 8, titled *PV1-PV5 & BRR*, reports the results using the procedure recommended by the survey organisers, e.g. using all five Plausible Values and the Student Weights, as well as the eighty Balance Replication Weights. Following DL 2018, all regressions include controls for gender (only in panel 1), month of birth, mother's and father's educational attainment, school location (urban or rural) and migration background (whether the student's home language is not German). Robust standard errors in parentheses (following DL 2018: clustered by school programme (academic or vocational), school location and gender). *** p<0.01, ** p<0.05, * p<0.1