Annex A9. A note about Spain in PISA 2018: Further analysis of Spain's data by testing date (updated on 23 July 2020)

Based on the additional information included in this note, the OECD has decided to release all available PISA 2018 data for Spain as follows:

- All plausible values for Spain are included in the international dataset, including plausible values for reading and for reading subscales. These were originally masked when the initial PISA 2018 data were released in December 2019.
- All raw data (including timing information and response patterns in fluency items) that were originally masked in the international dataset have been reinstated.
- A separate dataset for Spain, with student identifiers, information about the testing week, and a separate set of plausible values (not available for other countries) that do not use information from reading fluency items will be made available upon request, to enable replication of analyses included in this note and further comparisons within Spain.

While all data are released, Spain's performance results in PISA 2018 in all Volumes have been annotated, by drawing attention to a possible downward bias in performance results.

The following note regarding Spain's 2018 performance results has been added to PISA 2018 online tables:

> *In 2018, some regions in Spain conducted their high-stakes exams for tenth-grade students earlier in the year than in the past, which resulted in the testing period for these exams coinciding with the end of the PISA testing window. Because of this overlap, a number of students were negatively disposed towards the PISA test and did not try their best to demonstrate their proficiency. Although the data of only a minority of students show clear signs of lack of engagement (see Annex A9), the comparability of PISA 2018 data for Spain with those from earlier PISA assessments cannot be fully ensured.*

Further analysis of Spain's data by testing date

Two new variables received from the Spain National Centre were included in the analysis:

- WEEK_RND: Week of testing (mostly at the school level, but where schools ran two sessions, the week of testing is at the student level). This variable ranges from 1 to 10; but the few cases in weeks 1, 2 and 10 were combined with the nearest category (3 or 9) in order to avoid small cells. Weeks 7, 8, 9 correspond to the last three weeks of May.

EXT_JUN: Early high-stakes exams. This binary variable identifies the five regions that held their end-of-lower-secondary high-stakes exams in the second half of May 2018.[1] In all other regions[2] these exams were held in the first half of June 2018.

Table 1. Values and description of variable EXT_JUN

	EXT_JUN=0	EXT_JUN=1	Notes
Trimestral examinations	First half of June	Second half of May	
Final ordinary examinations	Second half of June	Early June	Only for students who did not pass trimestral examinations
Final extraordinary examinations	September	Second half of June	Only for student who did not pass initial ordinal examinations

Descriptive analysis

Most students were tested between week 3 and week 8, i.e. within a 6-week testing window (more than 100 schools in each of these weeks); a few schools conducted testing in week 9. Both types of regions conducted testing throughout the core period (weeks 3-9).

Table 2. Number of students assessed each week

Overall and by type of region

	Number of students	Percentage of students	In regions with non-early exams (EXT_JUN = 0)	In regions with early exams (EXT_JUN = 1)
Week 1 (2 April)	6	0.02	0	6
Week 2 (9 April)	5	0.01	0	5
Week 3 (16 April)	7 243	20.18	3 911	3 332
Week 4 (23 April)	7 263	20.23	4 229	3 034
Week 5 (30 April)	4 000	11.14	2 590	1 410
Week 6 (7 May)	6 633	18.48	4 480	2 153
Week 7 (14 May)	6 015	16.76	4 287	1 728
Week 8 (21 May)	4 228	11.78	2 525	1 703
Week 9 (28 May)	501	1.40	190	311
Week 10 (4 June)	2	0.01	0	2
Total	35 896	100.00	22 212	13 684

Engagement and performance, by week of testing

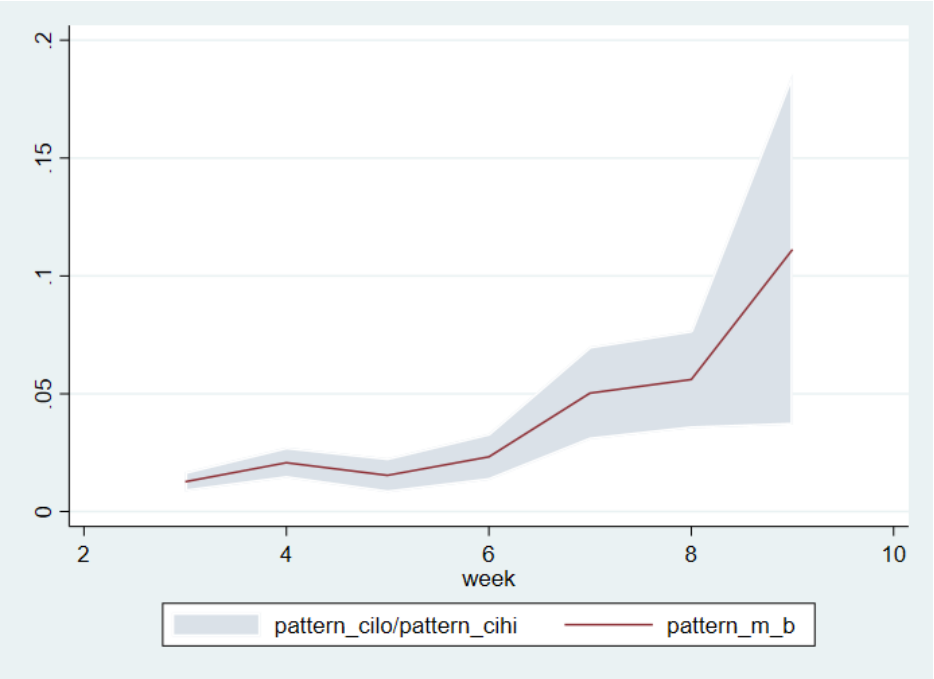
A clear association between the testing date and the proportion of students flagged with anomalies in their reading fluency items (all identical answers, with the whole fluency section completed in less than 20 seconds) could be established: this proportion increased to more than 5% by week 7 and continued to increase throughout the remaining weeks.

[1] Four of these regions changed examination dates between 2015 (the previous PISA assessment) and 2018. One region had already anticipated final examinations before 2015.

[2] In the traditional calendar, end-of-lower-secondary exams (ordinary session) are held in late June, and an extraordinary session is held in September. As a consequence, examinations for the last term are concentrated in the first half of June. In regions that hold the ordinary session in early June, a large number of last-term exams are held in the second half of May, potentially interfering with PISA administration.

Figure 1. Reading fluency anomalies, by week of testing

Proportion and confidence interval for the proportion



Similar patterns could be observed for other indicators of test engagement too: the proportion of students reporting, after completing the test, that they expended almost no effort (1, 2 or 3, on a 1-to-10 scale) on the PISA test increased towards the end of the testing window (Figure 2).

This lack of engagement – as reported by students themselves, and as observed in their behaviour on reading fluency items – is also reflected in a drop in students' performance in all three subjects[3] (Figure 3).

[3] Additional analyses were conducted to examine whether mathematics and science results were contaminated by problematic response behaviour in the reading test. Context effects could result from behavioural spill-overs (problematic response behaviour in mathematics and science being triggered by the testing experience in reading), or from dependencies in the analysis (the scaling of mathematics and science domains and the estimation of plausible values for performance in these domains being affected by problematic reading responses). No evidence could be found to support a significant role for context effects. In particular, performance patterns in mathematics and science are similar among students who took these domains *before* the reading test and among students who took these domains *after* the reading test; and plausible values in science (or mathematics) for students who did not take the science (or mathematics) test, and are therefore imputed based also on reading responses, are in line with the results for students who took these tests. The observed patterns in mathematics and science are therefore best explained by an overarching "student" effect, such as, student engagement, rather than by the test design. The nature and origin of this student effect is examined in detail in this note.

Figure 2. Percentage of students expending very little effort on the PISA test and on a test that would have counted towards their grades, by week of testing

Students reporting their effort on the PISA test as 1, 2 or 3 on a 1-to-10 effort scale (above, blue), and reporting that they would have expended similarly low effort if the test had counted towards their grades (below, red).

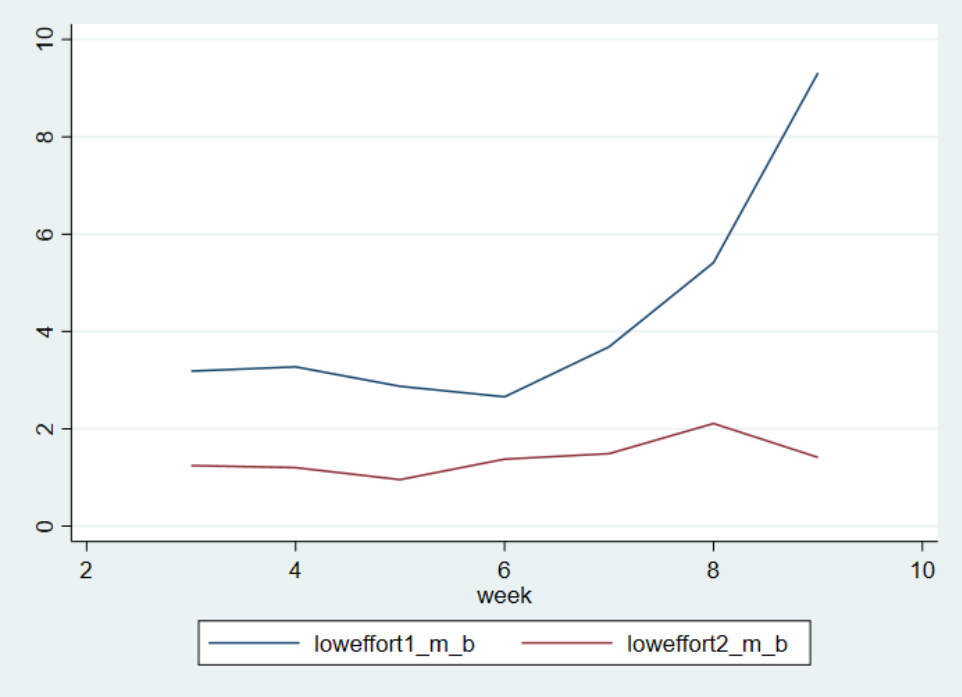
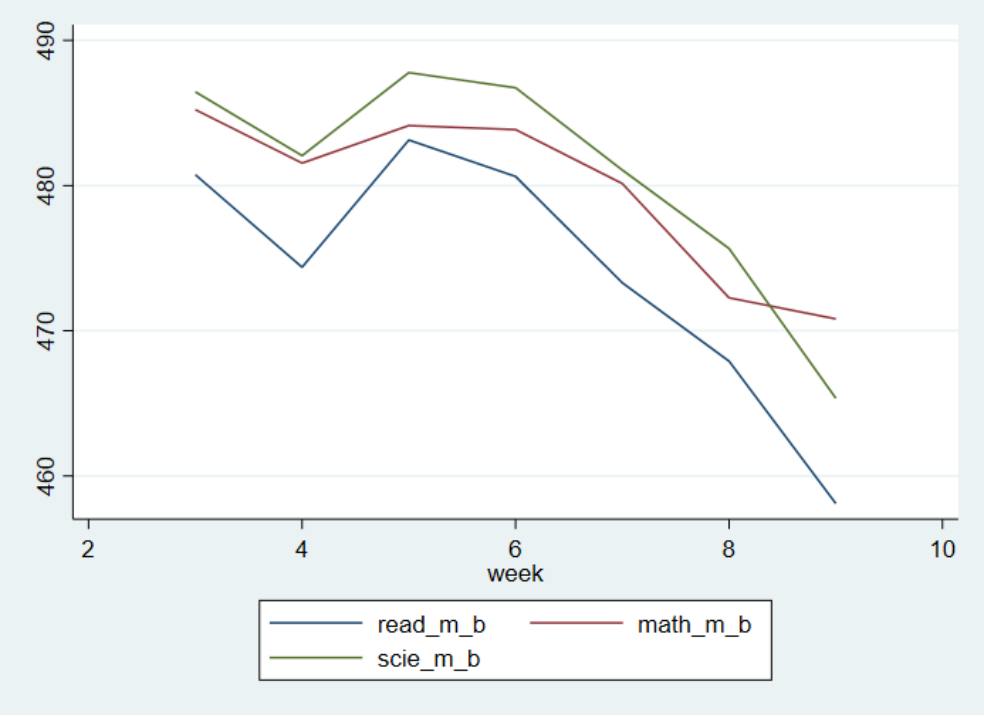


Figure 3. Performance in reading, mathematics and science, by week of testing



Engagement and performance, by week of testing and region

The descriptive analyses were also repeated by separating the sample into two groups of regions. The first group (EXT_JUN=1) comprises five regions that anticipated the testing dates for high-stakes end-of-lower-secondary schooling exams in the second half of May (with catch-up exams in early June and in the second half of June). The second group (EXT_JUN=0) includes all remaining regions, where these tests were in the first half of June with catch-up exams organised in the second half of June and in September.

The proportion of reading fluency anomalies clearly displays an increasing pattern in the first group of regions only; it remains stable, and below 4% of the sample, in the remaining regions (Figure 4).

The difference between the two regions is less stark regarding the proportion of students who expended very little effort and regarding performance (noting that there are few cases in the last testing week, where differences appear greatest but could easily be confounded by school characteristics) (Figure 5, Figure 6 (next page)).

Figure 4. Reading fluency anomalies, by week of testing and region

Proportion of students with patterned responses on reading fluency items in regions that anticipated high-stakes tests (above, red) and regions that kept the traditional timeline for high-stakes tests (below, blue)

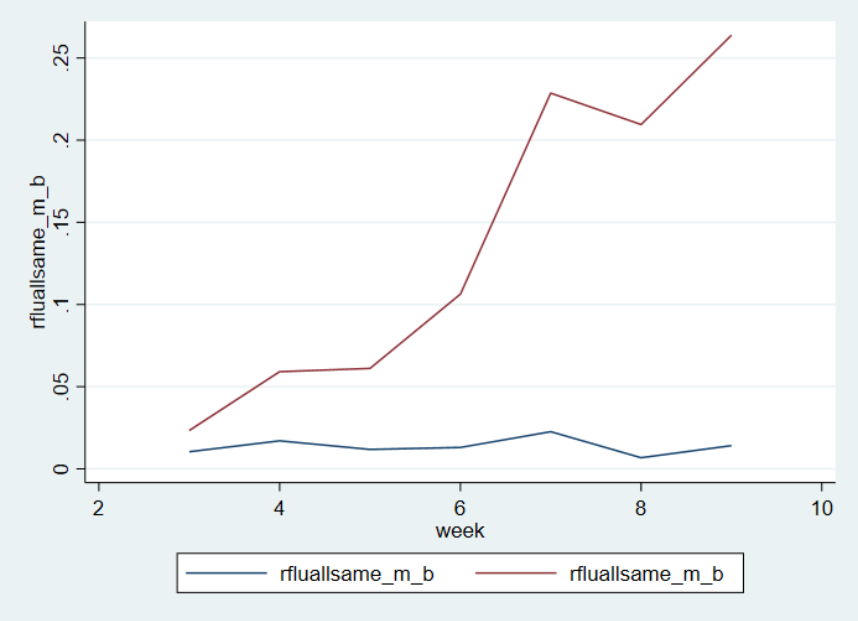


Figure 5. Percentage of students expending very little effort on the PISA test and on a test that would have counted towards their grades, by week of testing and region

Students reporting their effort on the PISA test as 1, 2 or 3 on a 1-to-10 effort scale (above, blue), and reporting that they would have expended similarly low effort if the test had counted towards their grades (below, red); left panel: regions that anticipated high-stakes test; right panel: regions that kept traditional dates for high-stakes tests

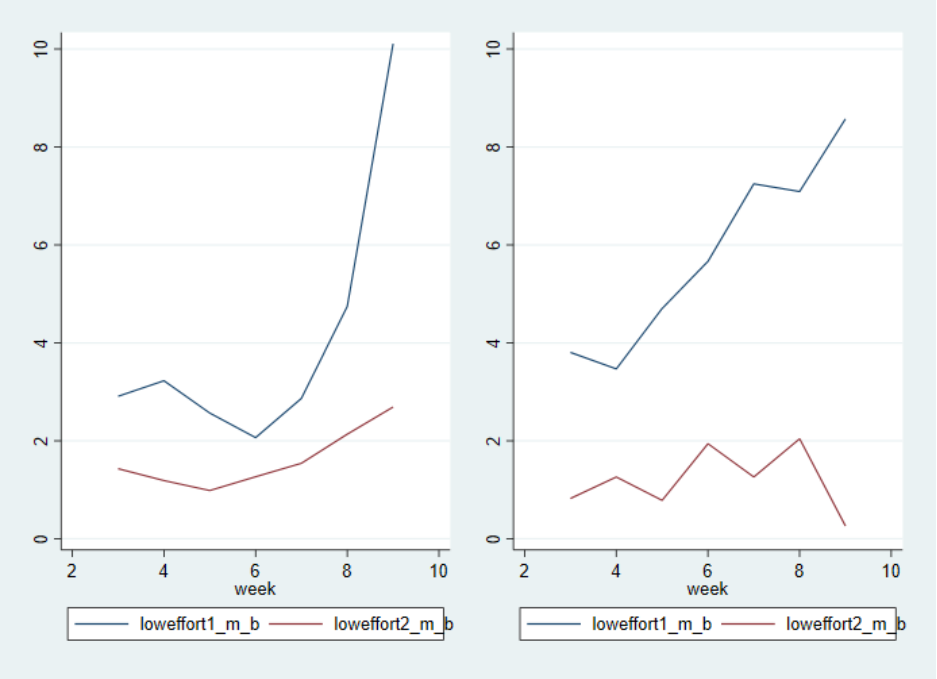
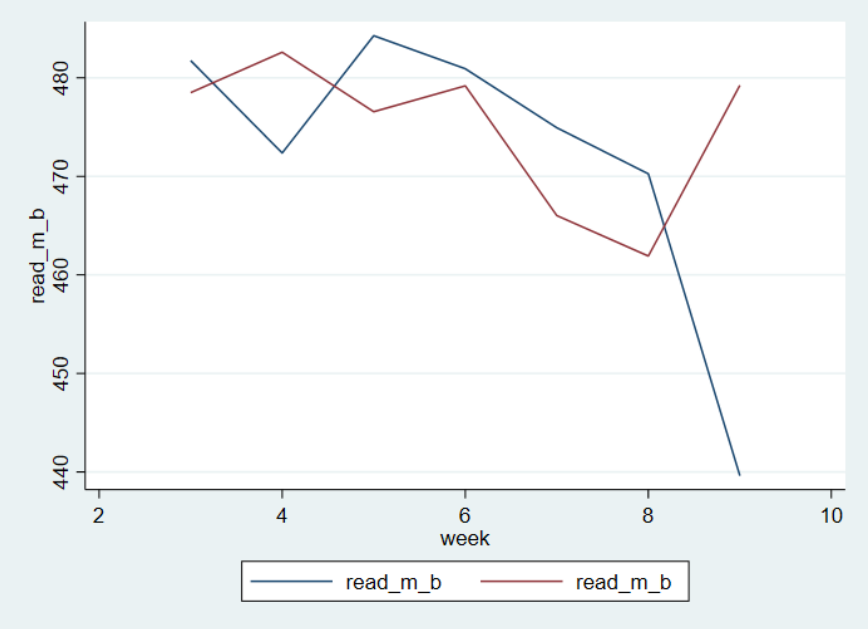


Figure 6. Performance in reading, by week of testing and region

Regions that anticipated high-stakes tests in the second half of May (blue); regions that kept traditional dates for high-stakes tests (red)



Regression analysis

In order to assess the significance of the differences and control for potential confounding factors (in particular, differences in socio-economic status), a regression analysis was conducted. In these analyses (Table 3) a "late test" indicator was created, identifying students who sat the test in week 7 or later. The regression analysis takes the form of a typical differences-in-differences estimation, with the two indicators for "Late testing" and "Early exam", and their interaction. The coefficient on the "Late testing" dummy indicates the "baseline" effect of testing late in the testing window, for regions where the high-stakes exams were not held until June with the last follow-up exams in September. The sum of this coefficient and the interaction term indicates the effect of late testing for schools that moved the high-stakes exams to the second half of May with the last follow-up exams in June.

The proportion of students with all identical (all yes, or all no) answers to reading fluency items increased significantly (+16 percentage points) in the final weeks, but only in those regions that moved the high-stakes examinations to an earlier date. Similar stark contrasts are found for all indicators based on reading fluency.

The proportion of students reporting that they spent very little effort (1, 2 or 3) on the PISA test also increased significantly in the final weeks, and more strongly so in those regions that moved high-stakes examinations earlier. The increase was not significant for the remaining regions (p-value 0.056), but was significant in the five regions with early high-stakes exams (the difference between the two effects, however, has a p-value of 0.054).

Finally, there was a significant drop in performance in reading (-14 score points) for late-testing schools in regions with early high-stakes exams, while the drop was not significant (p-value 0.471) in the remaining regions. The difference between the two types of regions was significant.[4] A similar drop – limited to regions with early high-stakes exams – was also found in mathematics (-11 points)[5], but was smaller, and not significant, in science and global competence. Some drop in performance was also found in the smaller (and separate) financial literacy sample; but in the case of financial literacy, the drop was not significantly different across regions.

Table 3. Regression results

Dependent variable	Patterned fluency responses		Low self-report effort in PISA		Reading performance		Mathematics performance	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
latetest	0.0030	(0.0056)	0.0103	(0.0054)	-2.5	(3.53)	-0.4	(2.96)
EXT_JUN	0.0462*	(0.0080)	0.0160*	(0.0056)	-6.4*	(2.77)	-7.1*	(2.31)
latetest*EXT_JUN	0.1603*	(0.0263)	0.0197	(0.0102)	-11.9*	(5.42)	-10.6*	(4.67)
female	-0.0089*	(0.0025)	-0.0174*	(0.0034)	24.7*	(1.69)	-8.0*	(2.10)
escs_nonmiss	-0.0047*	(0.0019)	0.0008	(0.0018)	27.6*	(0.85)	29.7*	(0.90)
escs_miss	0.0332*	(0.0160)	0.0651*	(0.0198)	-65.2*	(6.99)	-51.7*	(8.26)
constant	0.0166*	(0.0025)	0.0341*	(0.0028)	472.0*	(2.34)	489.3*	(2.08)
N. of observations	35847		29948		35896		35896	
Sum of latetest and latetest*EXT_JUN	0.1634*	(0.0258)	0.0300*	(0.0090)	-14.5*	(4.54)	-11.0*	(4.11)

Note: *: significant at the 5% level.

[4] There was no significant drop in performance on the main part of the reading test (as shown by an alternative set of PVs, ignoring information from the reading fluency items). This indicates that the reading fluency section was particularly sensitive to the motivation of students to try their best.

[5] In mathematics, the drop in performance remained significant (-8 points) even if the reading fluency information is ignored in imputation. This implies that the drop in performance was not *due* to reading fluency items, but to a more general reason that affected both mathematics and reading (fluency) performance.

Conclusion

- There is a good explanation why “Rapid and patterned responses were not uniformly present in the Spanish sample, but observed predominantly in a small number of schools in some areas of Spain” (the anomalies mentioned in Annex A9 in the Initial Report). These schools are all late-testing schools, in regions where the high-stakes tests overlapped with the end of the PISA testing window.
- Such patterns reflect more general negative dispositions towards the PISA test among a minority of students, due to the particular circumstances of the PISA 2018 test, as shown by the large number of students who admitted having spent very little effort on the PISA test they just completed (self-reported effort equal to 1, 2 or 3 on a 1-to-10 effort scale).
- Student performance in regions with early high-stakes exams showed marked declines during the last weeks of testing, not only in reading fluency, but also in mathematics.
- While this negative student disposition had a negative impact on these students’ performance in PISA, the overall impact on the country’s mean performance did not exceed a handful of PISA points. The impact is larger on the results of the five subnational entities with early high-stakes exams.
- The analysis of Spain’s data also reveals how the inclusion of reading fluency items may have strengthened the relationship between test performance and student effort in PISA more generally. The OECD is therefore exploring changes to the administration and scoring of reading fluency items to limit the occurrence of disengaged response behaviour and mitigate its consequences.