ANNEX A5
How comparable are the PISA 2018 computer- and paper-based tests?

In the vast majority of participating countries, PISA 2018 was a computer-based assessment. However, nine countries – Argentina, Jordan, Lebanon, the Republic of Moldova, the Republic of North Macedonia, Romania, Saudi Arabia, Ukraine and Viet Nam – assessed their students' knowledge and skills in PISA 2018 using paper-based instruments. These paper-based tests were offered to countries that were not ready, or did not have the resources, to transition to a computer-based assessment.[1] The paper-based tests comprise a subset of the tasks included in the computer-based version of the tests, all of which were developed in earlier cycles of PISA according to procedures similar to those described in Chapter 2. No task that was newly developed for PISA 2015 or PISA 2018 was included in the paper-based instruments; consequently, the new aspects of the science and reading frameworks were not reflected in the paper-based tests.

This annex describes the differences between paper- and computer-based instruments, and what they imply for the interpretation of results.

DIFFERENCES IN TEST ADMINISTRATION AND CONSTRUCT COVERAGE

Over the past decades, digital technologies have fundamentally transformed the ways we read and manage information. Digital technologies are also transforming teaching and learning, and how schools assess students. To reflect how students and societies now commonly access, use and communicate information, starting with the 2015 assessment cycle, the PISA test was delivered mainly on computers. Existing tasks were adapted for delivery on screen; new tasks (initially only in science, then, for PISA 2018, also in reading) were developed that made use of the affordances of computer-based testing and that reflected the new situations in which students apply their science or reading skills in real life.

Because pen-and-paper tests are composed only of items initially developed for cycles up to PISA 2012, the paper-based version of the PISA 2018 test does not reflect the updates made to the assessment frameworks and to the instruments for science and reading. In contrast, the paper-based instruments for mathematics and their corresponding computer-based versions have their roots in the same framework, originally developed for PISA 2012.

The changes introduced in the assessment of science, in 2015, and of reading, in 2018, have deep implications for the set of assessment tasks used. The new frameworks resulted in a larger amount of assessment tasks at all levels; extended coverage of the reading and science scales through tasks that assess basic reading processes and emerging science skills (proficiency Levels 1b in science and 1c in reading); an expanded range of skills measured by PISA; and the inclusion of new processes or new situations in which students' competence manifests itself. Table I.A5.1 summarises the differences between the paper- and computer-based tests of reading; Table I.A5.2 summarises the corresponding differences in science.[2]

In reading, newly developed tasks could include using hyperlinks or other navigation tools (e.g. menus, scroll bars) to move between text segments. At the beginning of the reading test, a section was added to measure reading fluency, using timed sentence-comprehension tasks (see Chapter 1, Annex A6 and Annex C). None of these tasks would be feasible in a large-scale paper-based assessment. In science, new "interactive" tasks were developed for the PISA 2015 assessment. These tasks used computer simulations to assess students' ability to conduct scientific enquiry and interpret the resulting evidence. In these tasks, the information that students see on the screen is determined, in part, by their own interactions (through mouse clicks, keyboard strokes, etc.) with the task.

There are other differences between the PISA paper- and computer-based tests in addition to the tasks included in the tests and the medium through which students interacted with those tasks.

While the total testing time for all students was two hours, students who sat the test using computers had to take a break before starting work on the second half of the test, and had to wait until the end of the first hour before doing so. Students who sat the paper-based test also had to take a break after one hour of testing, but they could start working on the second half of the test during that first hour.

Another difference in test administration was that students who sat the test using computers could not go back to questions in a previous test unit or revise their answers during the test or after reaching the end of the test sequence (neither at the end of the first hour, nor at the end of the second hour).[3] In contrast, students who sat the paper-based version could, if they finished earlier, return to their unsolved tasks or change the answers they had originally given to some of the questions.

In 2018, and on average across countries that delivered the test on computer, 50% of students completed the reading test within about 40 minutes, i.e. about 20 minutes before the end of the test hour (Table I.A8.15). For additional analyses on response-time data, see Annex A8 and in the *PISA 2018 Technical Report* (OECD, forthcoming[1]).

In addition, the computer-based test in reading was a multi-stage adaptive test (see Chapter 1). In practice, the test forms consisted of three segments (stages): students were presented with a particular sequence of test tasks in the second and third stages based on a stochastic algorithm that took into account their performance on previous segments (OECD, forthcoming[1]; Yamamoto, Shin and Khorramdel, 2018[2]).[4] In science and mathematics (and also in reading for those countries that delivered the paper-based test), students were assigned test forms via a random draw, independent of the student's proficiency or behaviour on the test.

Table I.A5.1 **Differences between paper- and computer-based assessments of reading**

	Paper ("A" booklets)	Paper ("B" booklets)	Computer	Computer (excluding reading-fluency tasks)
Number of assessment tasks	88	87	309*	244*
Number of unique test booklets/forms	12	12	2304 possible paths through the assessment (12 reading fluency combinations x 192 adaptive paths)	192 possible paths through the assessment (128 unique combinations of items, of which 64 exist, in different disposition, as part of two paths)
Assignment of test booklets/forms to student	Random	Random	Random (reading fluency) + Adaptive	Adaptive
Assessment tasks, by PISA cycle in which they were first used				
PISA 2018	0	0	237	172
PISA 2009	49	59	44	44
PISA 2000	39	28	28	28
Range of task difficulty, on the PISA reading scale (RP62)**				
Min	224	224	67	224
10th percentile	377	373	213	378
Median	477	468	451	480
90th percentile	632	633	631	642
Max	1 045	1 045	1 045	1 045
Number of tasks, by proficiency level				
Level 6	4	4	10	10
Level 5	5	5	23	23
Level 4	18	14	34	34
Level 3	16	16	50	50
Level 2	22	23	71	71
Level 1a	20	18	47	46
Level 1b	2	5	31	9
Level 1c	1	2	12	1
Below Level 1c	0	0	31	0
Number of tasks, by sources required				
Single source	86	85	257	192
Multiple source	2	2	52	52
Number of tasks, by process				
Reading fluency	0	0	65	0
Locating information	17	18	49	49
Understanding	50	45	131	131
Evaluating and reflecting	21	24	64	64

Notes: "A" and "B" booklets have 72 test items in common; items unique to "A" booklets tend to be, on average, more difficult than the items unique to "B" booklets. Only items common to both "A" and "B" booklets are also used in the computer-based test. In the absence of adaptive testing, countries were invited to choose the booklet set that best matched the expected proficiency of their students. In 2018, only Ukraine used the "A" booklets; all other countries that delivered PISA 2018 on paper used the "B" booklets.

*Item CR563Q12, which was included in the computer-based test of reading but excluded from scaling, is not included in item counts in this table.

**All percentiles are unweighted. For the computer-adaptive test, the actual distribution of task difficulty, weighted by the proportion of students who responded to each task, is also a function of the distribution of student proficiency in the country.

Source: OECD, PISA 2018 Database; *PISA 2018 Technical Report* (OECD, forthcoming[1]).

Table I.A5.2 **Differences between paper- and computer-based assessments of science**

	Paper	Computer
Number of assessment tasks	85	115
Number of unique test booklets/forms	18	18
Assignment of test booklets/forms to student	Random	Random
Assessment tasks, by PISA cycle in which they were first used		
PISA 2018	0	76
PISA 2006	85	39
Range of task difficulty, on the PISA reading scale (RP62)		
Min	305	305
10th percentile	437	426
Median	539	535
90th percentile	649	659
Max	821	925
Number of tasks, by proficiency level		
Level 6	3	3
Level 5	8	16
Level 4	23	30
Level 3	31	37
Level 2	16	21
Level 1a	3	7
Level 1b	1	1
Below Level 1b	0	0
Number of tasks, by science competency		
Interpret data and evidence scientifically	28	36
Explain phenomena scientifically	41	49
Evaluate and design scientific enquiry	16	30
Number of tasks, by type of knowledge		
Content	51	49
Procedural	24	47
Epistemic	10	19
Number of tasks, by system		
Living	39	47
Earth and Space	18	30
Physical	28	38

Source: OECD, PISA 2018 Database; *PISA 2018 Technical Report* (OECD, forthcoming[1]).

HOW THE EVIDENCE ABOUT MODE EFFECTS WAS USED TO LINK THE TWO DELIVERY FORMATS

In order to ensure comparability of results between the computer-delivered tasks and the paper-based tasks that were used in previous PISA assessments (and are still in use in countries that use paper instruments), for the test items common to the two administration modes, the invariance of item characteristics was investigated using statistical procedures. These included model-fit indices to identify measurement invariance (see Annex A6), and a randomised mode-effect study in the PISA 2015 field trial that compared students' responses to paper-based and computer-delivered versions of the same tasks across equivalent international samples (OECD, 2016[3]). For the majority of items, the results supported the use of common difficulty and discrimination parameters across the two modes of assessment. For some items, however, the computer-delivered version was found to have a different relationship with student proficiency from the corresponding, original paper version. Such tasks had different difficulty parameters (and sometimes different discrimination parameters) in countries that delivered the test on computer. In effect, this partial invariance approach both accounts for and corrects the potential effect of mode differences on test scores.

Table I.A5.3 shows the number of anchor items that support the reporting of results from the computer-based and paper-based assessments on a common scale. The large number of items with common difficulty and discrimination parameters indicates a strong link between the scales. This strong link corroborates the validity of mean comparisons across countries that delivered the test in different modes. At the same time, Table I.A5.3 also shows that a large number of items used in the PISA 2018 computer-based tests of reading and, to a lesser extent, science, were not delivered on paper. Caution is therefore required when drawing

conclusions about the meaning of scale scores from paper-based tests, when the evidence that supports these conclusions is based on the full set of items. For example, the proficiency of students who sat the PISA 2018 paper-based test of reading should be described in terms of the PISA 2009 proficiency levels, not the PISA 2018 proficiency levels, and similarly for science. This means, for example, that even though PISA 2018 developed a description of the skills of students who scored below Level 1b in reading, it remains unclear whether students who scored within the range of Level 1c on the paper-based tests have acquired these basic reading skills.

Table I.A5.3 **Anchor items across paper- and computer-based scales**

Scalar-invariant, metric-invariant and unique items in PISA 2018 paper and computer tests

	Reading	Mathematics	Science
Items with common difficulty and discrimination parameters across modes (scalar invariant)	40	50	29
Items with common discrimination parameter across modes, but distinct difficulty parameter (metric invariant)	32	31	10
Items with mode-specific parameters	0	1*	0
Items not delivered on computer (paper-based only)	15 ("A" booklets) 16 ("B" booklets)	1	46
Items not delivered on paper (computer-based only)	172+65 "fluency" items**	0	76

* In PISA 2015 and in the mode-effect study, Item M192Q01 was excluded from scaling in the computer-based version due to a technical issue. Its parameters were therefore freely estimated in 2018.

** In addition, item CR563Q12 was included in the computer-based test of reading but excluded from scaling due to a technical problem with the recording of students' answers.

Note: The table reports the number of scalar-invariant, metric-invariant and unique items based on international parameters. In any particular country, items that receive country-specific item parameters (see Annex A6) must also be considered.

Source: OECD, PISA 2018 Database; *PISA 2018 Technical Report* (OECD, forthcoming[1]).

Notes

1. Albania, Georgia, Indonesia, Kazakhstan, Kosovo, Malta, Panama and Serbia transitioned to the computer-based assessment in 2018. All other returning PISA 2018 participants, including all OECD countries, made the transition in 2015.
2. No subscales are estimated for students who sat the paper-based test of reading.
3. In the computer-based test, and with limited exceptions, students were still able to go back to a previous question within the same unit and revisit their answers. They were not allowed to go back to a previous unit.
4. Before the first segment of the adaptive test (also called "core" stage), all students also completed a 3-minute reading-fluency section, which consisted of 21 or 22 items per student, assembled, according to 12 possible combinations, from 65 available items. Performance on this reading-fluency section was not considered by the adaptive algorithm in the main section of the reading test.

References

OECD (2016), *The PISA 2015 Field Trial Mode-Effect Study*, OECD Publishing, Paris, www.oecd.org/pisa/data/PISA-2015-Vol1-Annex-A6-PISA-2015-Field-Trial-Mode-Effect-Analysis.pdf (accessed on 1 July 2019). [3]

OECD (forthcoming), *PISA 2018 Technical Report*, OECD Publishing, Paris. [1]

Yamamoto, K., H. Shin and **L. Khorramdel** (2018), "Multistage Adaptive Testing Design in International Large-Scale Assessments", *Educational Measurement: Issues and Practice*, Vol. 37/4, pp. 16-27, http://dx.doi.org/10.1111/emip.12226. [2]