How does the PISA multi-stage adaptive test of reading work?

To provide the most comprehensive measure of reading literacy, PISA would have to present each student with the complete set of test items. Asking students to answer all such items would be the best way to eliminate any gaps or biases in the assessment. However, this would result in a test that would take more than six hours to complete.

To make it feasible to measure student proficiency in all domains, the test material in all PISA cycles, up to and including PISA 2015, was divided into several 30-minute clusters or test booklets. These clusters were linked across domains and organized into test forms, which were then randomly allocated to students. Students received two 30-minute clusters of test material in the major domain along with two booklets in one or two of the other domains.

Each student saw only a small subset of the test material and was thus assessed on only a selection of the skills and competencies that comprise each domain. Nonetheless, students in an education system, when taken as a group, were examined on the complete set of skills. Despite this randomization procedure, one source of inaccuracy remains. Most students in OECD countries score near the middle of the score, or at around 500 points. Most of the test material is also targeted to middle-performing students, which allows for more refined differentiation of student ability at this level. However, this means that there is a relative lack of test material at the higher and lower ends of student ability, and that the scores of both high- and low-performing students are determined with less accuracy than the scores of middle-performing students.

In order to increase the accuracy of such measurements, PISA 2018 introduced **adaptive testing** in its reading assessment. Instead of using fixed, predetermined test booklets as was done through PISA 2015, the reading assessment given to each student was dynamically determined, based on how the student performed in prior stages.

There were three stages to the PISA 2018 reading assessment: Core, Stage 1 and Stage 2.[1] Students first saw a short Core stage, which consisted of between 7 and 10 items.[2] The vast majority of these items (at least 80 percent and always at least 7 items) were automatically scored. Students' performance in this stage was provisionally classified as low, medium or high, depending on the number of correct answers to these automatically scored items.[3]

The various Core Blocks of material delivered to students did not differ in any meaningful way in their difficulty. Stage 1 and 2, however, both existed in two different forms: comparatively easy and comparatively difficult.[4] Students who displayed medium performance in the Core stage were equally likely to be assigned an easy or a difficult Stage 1. Students who displayed low performance in the Core stage had a 90 percent chance of being assigned to an easy Stage 1 and a 10 percent chance of being assigned to a difficult Stage 1. Students who displayed high performance in the Core stage had a 90 percent chance of being assigned to a difficult Stage 1 and a 10 percent chance of being assigned to an easy Stage 1.

Students were assigned to easy and difficult Stage 2 blocks of material in much the same way. In order to classify student performance as precisely as possible, however, responses to automatically scored items from both the Core stage and Stage 1 were used.

As with many of the new features in the PISA 2018 reading framework, adaptive testing was made possible through the use of computers. One potential drawback of an adaptive design is that students are unable to return to a question after it has been answered or skipped. This was already the case in the PISA 2015

computer-based assessment. However, with adaptive testing, students' responses in the Core stage and in Stage 1 affected not only their performance but also the questions that they saw later in the assessment. The *PISA 2018 Technical Report* (OECD forthcoming) presents further indicators of the impact of adaptive testing on students' test-taking behavior.

---

[1] For a more detailed description of the adaptive testing design, and a discussion of the considerations that guided its development, see: *PISA 2018 Technical Report* (OECD forthcoming); and Yamamoto, K., Shin, H.J., and Khorramdel, L. (2018). Multistage Adaptive Testing Design in International Large-scale Assessments. *Educational Measurement*, 37 (4), 16-27.

[2] The Core stage was delivered as one of eight possible Core blocks of material, analogous to how reading was delivered as one of six possible test booklets in PISA 2015 (and how mathematics and science were delivered in PISA 2018). Each Core block was composed of two units, and each unit comprised a set of items developed around shared stimulus material.

[3] To select an adequate test form while students sat the assessment, their performance was classified using only the automatically scored items. However, all items, including those that required human coding, were used to evaluate overall performance and report students' proficiency in reading.

[4] More specifically, both Stages 1 and 2 were delivered as one of 16 possible blocks of material, 8 of which were comparatively easy and 8 of which were comparatively difficult. Each Stage 1 block was built from 3 units that, in total, amounted to 12 and 15 items, of which between 8 and 11 were automatically scored. Similarly, each Stage 2 block was built from 2 units that, in total, amounted to between 12 and 15 items, of which between 6 and 12 were automatically scored.