RESEARCH REPORT

# PISA 2025

Montenegro

**199**

science items analysed

**0 of 149**

testable items flagged by programme in a uniform DIF screen

**15**

science items flagged by gender at the nominal 1% level

CENTER FOR PSYCHOLOGY

CENTER FOR PSYCHOLOGY RESEARCH REPORTS

# Where vocational schools fall behind: item-level evidence from PISA 2025

The OECD reports Montenegro's PISA 2025 results as averages and shares (OECD, 2026a). Computed from the public microdata, the distance between gymnasium and vocational students is 103 points in science. That number does not say whether the gap sits in particular kinds of task, or whether some items work differently for the two groups. This report examines both from the item responses themselves, using 199 science items and 168 mathematics items, the OECD's own classification of those items, and item response models fitted to Montenegro's data.

**Dr Milos Kankaras**

centerforpsychology.me

Center for Psychology | October 2026

DOI: 10.5281/zenodo.23389928

PISA 2025

## Key findings

**01** A uniform differential item functioning (DIF) screen detected no science item that functions differently for vocational students at the nominal 1% level: none of the 149 testable items, where one or two would be expected by chance. Fifty items could not be tested, and a null screen does not establish that every item functions identically.

**02** The gap differs between kinds of item, larger on procedural and epistemic knowledge (19.2 and 17.5 percentage points of available credit) than on content knowledge (13.7), but a model with one ability difference between the programmes and no category effects broadly reproduces these differences, to within 1.8 percentage points of available credit for every category shown.

**03** The apparent concentration of the gap on middle-difficulty items is largely what a single ability difference produces on a percent-correct scale, which compresses differences near its floor and ceiling.

**04** Mathematics shows the same: the screen flags none of 132 testable items by programme, and no content, context or process category carries a model-based item-specific component larger than 1.7 percentage points of available credit.

**05** Gender is different. An exploratory scan flags 15 of 161 testable science items for girls and boys of equal ability at the nominal 1% level, against 1.6 expected by chance: items about living systems tend to favour girls, items about physical and earth systems boys.

**06** Possible rapid guesses, flagged by a median-based timing rule, are four to five times more common in vocational schools (6.5% of timed responses against 1.4% in gymnasium), but excluding them changes the standardised gap only from 1.40 to 1.37.

**07** On items answered in both 2022 and 2025, the gap widened by 16.2 points in science and 25.7 in mathematics, with approximate standard errors that are conditional on the retained items and the score conversion.

01

## Data and the adaptive design

The analysis uses the PISA 2025 public cognitive item file: 6,436 Montenegrin students, of whom 1,159 attend gymnasiums and 3,578 vocational schools. Each item carries the OECD's classification from Annex A of the technical report (OECD, 2026b): competency, knowledge type, system, cognitive demand, format and proficiency level.

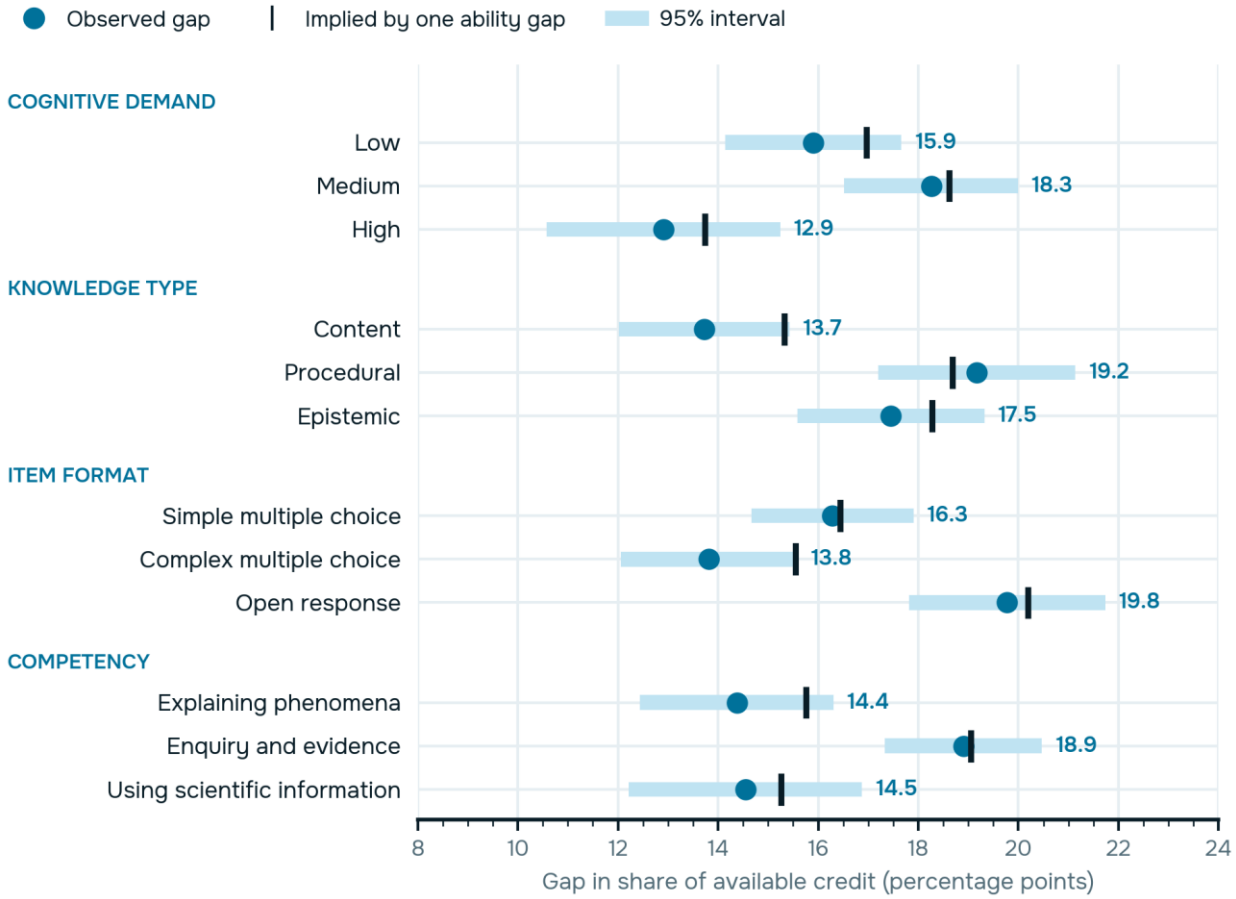
PISA 2025 science is a multistage adaptive test. After a randomly assigned core testlet, a student is routed to easier or harder testlets depending on earlier answers, and routing differed sharply by programme: 30% of vocational students were sent to the easiest stage-two testlets, against 10% in gymnasium. Under the assumed response model, routing based on observed responses can be handled by the likelihood, so the model-based analyses use all responses. A simple per-item statistic cannot handle it, because the students who see a hard item have already been selected on performance. The descriptive science comparisons below therefore use only responses that were not routed: all responses of students on the randomly assigned linear forms and the core-testlet responses of the rest, about half of all responses. The routing rules were reconstructed from the assignment variables and the annex testlet codes, and reproduce every one of the 100,822 adaptive responses.

02

## The gap is general

Averaged over items, gymnasium students earn 51% of the available credit and vocational students 35%. The difference varies across the OECD's categories: it is larger on procedural (19.2 percentage points of available credit) and epistemic (17.5) than on content knowledge (13.7), larger on constructing and evaluating designs for scientific enquiry (18.9) than on explaining phenomena (14.4), and larger on open-response items (19.8) than on complex multiple choice (13.8).

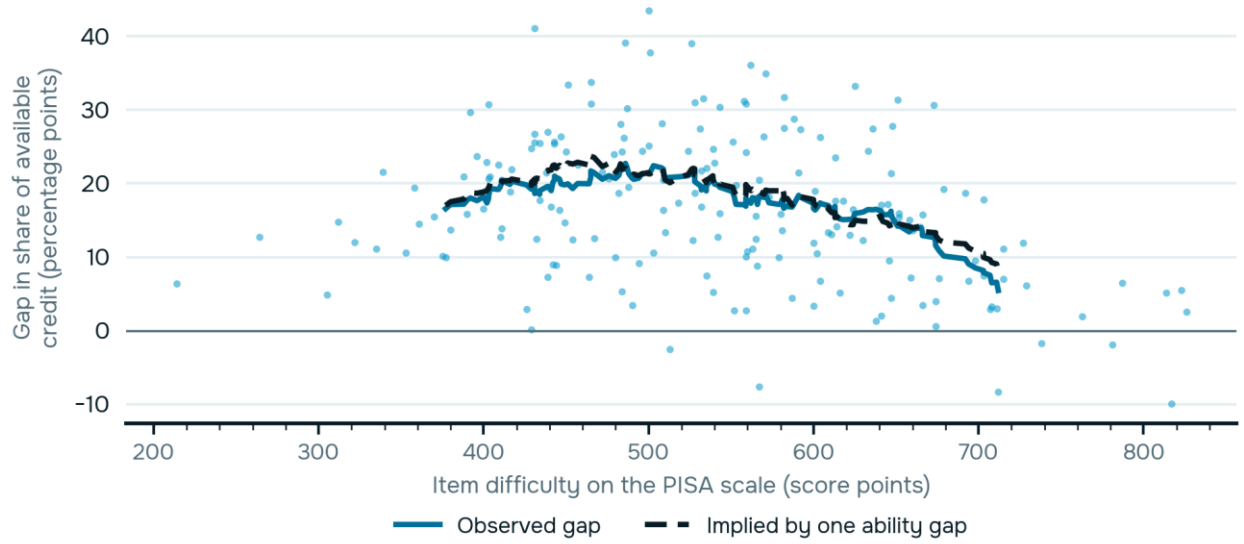
**Figure 1. Gap in share of available credit by item category, observed and implied by a single ability difference.**




Source: OECD PISA 2025 public use file (cognitive items) and Technical Report Annex A; author's calculations. Non-routed responses only. Bars: observed gap with 95% interval. Ticks: gap predicted by a two-group model with no category effects.

Read alone, that pattern suggests that vocational students struggle specifically with scientific procedure. Item response modelling points elsewhere. A two-group model with one ability distribution per programme and no category effects broadly follows the category pattern: observed and predicted category gaps differ by at most 1.8 percentage points of available credit among the categories in Figure 1, and by 2.9 for items at Level 2 difficulty; individual-item residuals are larger. Items about procedure and evidence are largely the items on which a difference in overall ability shows most on a percent-correct scale. The same holds for item difficulty: the per-item gap peaks on items around Levels 2 and 3 and falls to about 5 points on the hardest items, and the single-difference model follows that curve. Floor and ceiling compression on a percentage scale can produce such a hump; this does not prove that the scale is the only explanation.

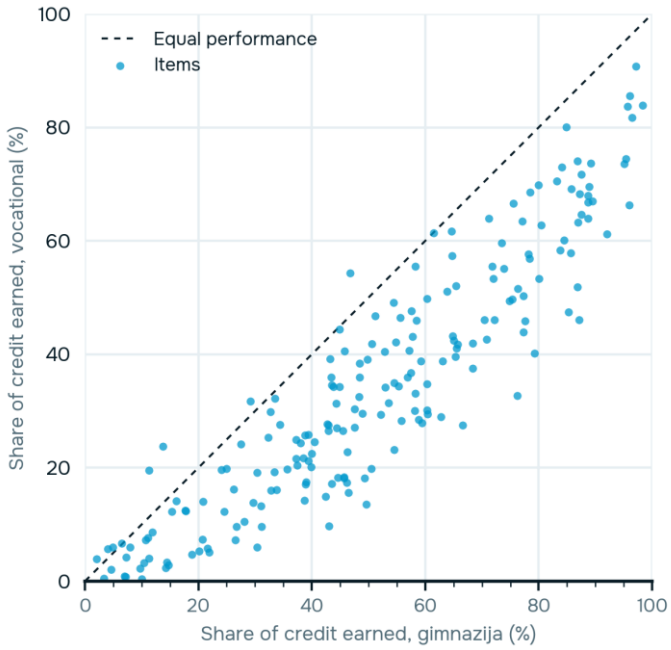
**Figure 2. Gap per item against item difficulty, with running means of the observed gap and of the gap implied by a single ability difference.**




Source: OECD PISA 2025 public use file (cognitive items) and Technical Report Annex A; author's calculations. Non-routed responses only. Shows 193 items with difficulty at or below 900 points; six harder items are omitted. Lines are running means over 25 items ordered by difficulty.

The formal screen detects nothing either. In a two-group calibration that frees each item's location for the two programmes while holding its slope equal for both and the ability distributions fixed, no item shows a shift for vocational students of equal ability at the nominal 1% level (the smallest p-value among the 149 testable items is 0.011). Fifty items could not be tested, mostly because they discriminate weakly in this sample. A null screen does not establish item equivalence or exclude contributions from format, context or language. Within those limits, the evidence points to a general difference in what students can do, which the companion report describes as half of vocational students remaining below Level 2.

**Figure 3. Share of available credit earned per item, gymnasium against vocational students.**




Source: OECD PISA 2025 public use file (cognitive items) and Technical Report Annex A; author's calculations. Non-routed responses only.

03

## Mathematics tells the same story

The same analysis on the 168 mathematics items, classified by content (quantity, space and shape, change and relationships, uncertainty and data), context and process, gives the same answer. In a two-group calibration, the screen flags none of the 132 items that could be tested. The model's ability difference between the programmes is 1.46 pooled within-group standard deviations, against 1.54 for the plausible values; the two estimates use different metrics, so the model gap is not converted to PISA points here. Across every content, context and process category, the model-based category-specific component is within 1.7 percentage points of available credit of zero.

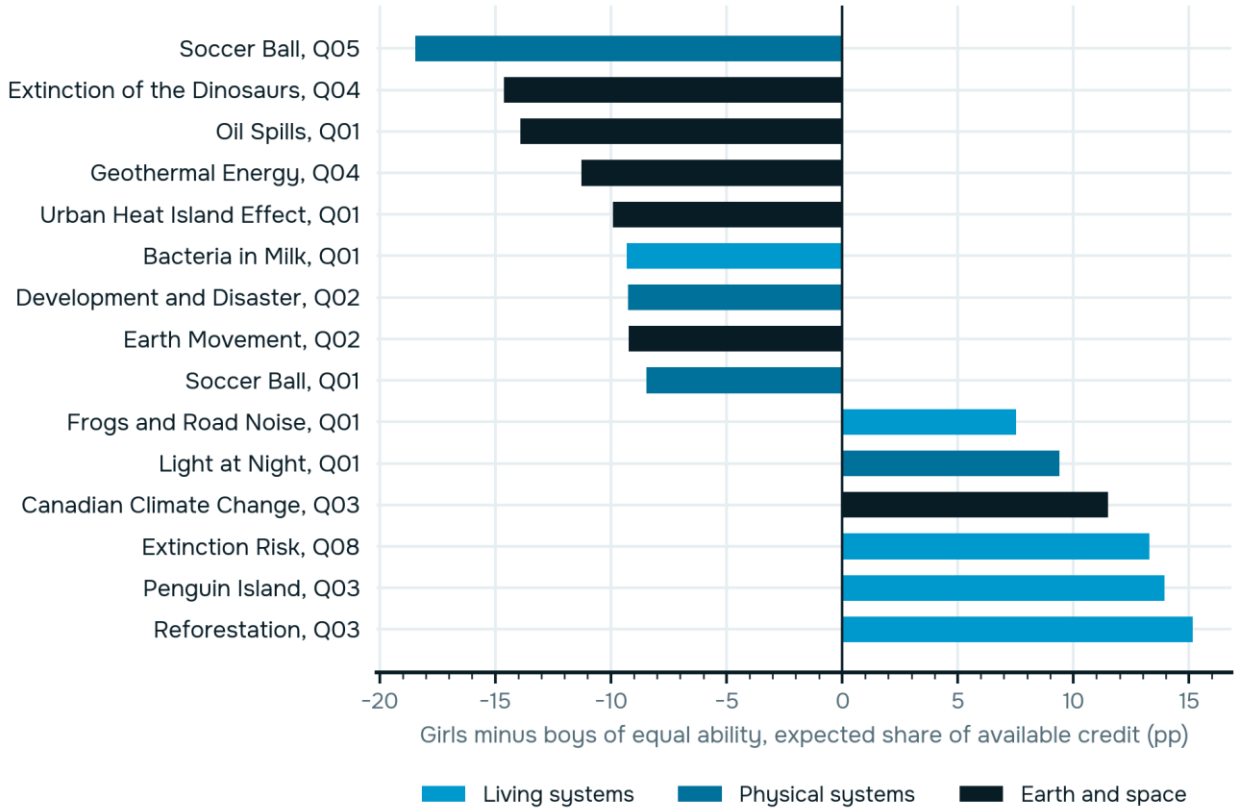
PISA 2025 mathematics is also partly adaptive, and the routing variables needed to separate routed from non-routed responses were not rebuilt for it here. The mathematics results are therefore reported only from the model, which can handle adaptive routing under its assumptions, and not as raw shares of credit.

04

## Girls and boys: where items do differ

The programme gap shows no item-specific pattern. The gender gap, smaller overall, does. Girls outperform boys in science by 12 points in Montenegro. The exploratory gender scan flagged 15 of 161 testable items at the nominal 1% level, compared with 1.6 expected under a calibrated global null; 4 remain under a Bonferroni correction and 13 under a false-discovery-rate correction at 5%. Individual flags require multiplicity-aware and model-robust confirmation.

**Figure 4. Science items flagged as functioning differently for girls and boys of equal ability, by scientific system.**




Source: OECD PISA 2025 public use file (cognitive items) and Technical Report Annex A; author's calculations. Bars: difference in expected share of available credit for girls and boys of equal ability; positive values favour girls. Nominal 1% level.

The pattern is largely consistent. Four of the six items that favour girls ask about living systems: penguin populations, reforestation, extinction risk, and the effect of road noise on frogs. Eight of the nine items that favour boys ask about physical or earth and space systems, among them the physics of a soccer ball, earth movement, oil spills, geothermal energy and urban heat islands. In the current item pool, estimated DIF components run in both directions and largely cancel on average. This is a model-based credit-component result, not a rescoring demonstration that the published 12-point gap is independent of item mix. Different content weights could change the gap; national assessments should check item functioning before reporting gender results.

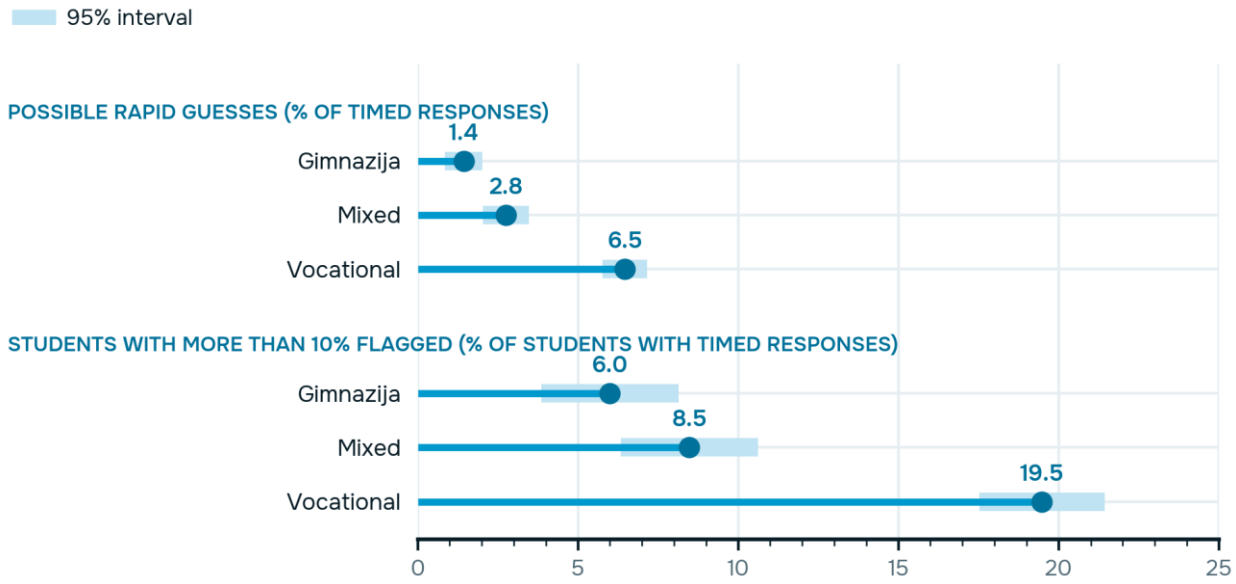
05

## Effort on the test is part of the picture, but a small part

Report 1 found that the effort students say they put into the test is associated with their score. Response times allow a closer look. The public process file records how long each student spent on each item for the 80 science trend items (the new 2025 items carry no times). A very short response signals a disengaged rapid guess rather than an attempt to solve the item (Wise, 2017). Responses were flagged as possible rapid guesses when their time was below the smaller of 10 seconds and one tenth of the item median time in Montenegro. This is a study-specific median-based rule, not the conventional mean-based NT10

definition and not a demonstrated minimum reading time. Only students with observed timed responses are covered: 397 of 1,159 in gymnasium and 1,474 of 3,578 in vocational schools, with a median of 10 timed responses each.

**Figure 5. Possible rapid guesses by school programme, PISA 2025 science trend items.**




Source: OECD PISA 2025 public use files (cognitive items and response times); author's calculations. Left: share of timed responses below the median-based threshold. Right: share of students with timed responses for whom more than one timed response in ten is flagged. 95% intervals from 80 replicate weights.

Possible rapid guesses are rare in gymnasium (1.4% of timed responses) and four to five times more common in vocational schools (6.5%). One vocational student in five (19.5%) has more than a tenth of timed responses flagged, against one in seventeen (6.0%) in gymnasium. Responses this fast are concentrated where scores are lowest.

On the 80 timed science items, excluding responses flagged by the median-based timing rule, as in the effort-moderated approach (Wise & DeMars, 2006), changes the gap, standardised separately by the gymnasium distribution in each fit, from 1.40 to 1.37 (a 1.8% relative difference in these standardised estimates). This sensitivity result does not identify a change on a common PISA scale, what students would score with greater effort, or the contribution of all forms of disengagement.

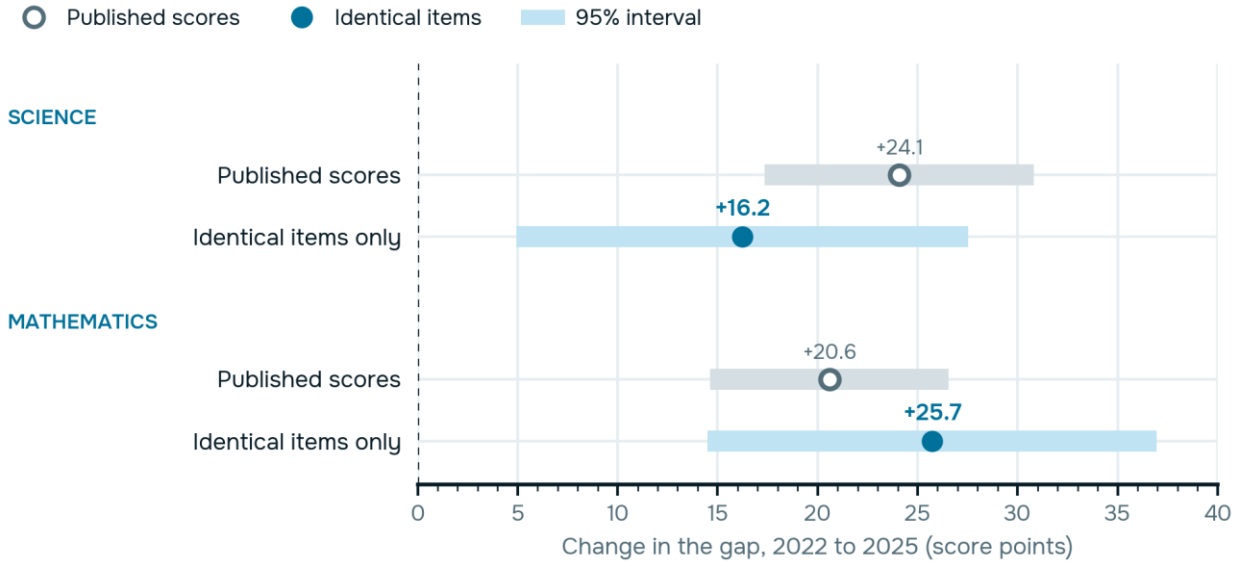
06

## The widening holds on identical items

Of the items shared across the two cycles, 79 science and 157 mathematics items met the scoring and response-count criteria. Removing 16 and 14 items flagged for drift by a uniform DIF scan left 63 and 143 items for concurrent calibration, which places gymnasium and vocational students in each cycle on one scale. On the 2022 plausible-value metric, the gap widened by 16.2 points in science (standard error 5.8) and 25.7 in mathematics (5.7). The approximate standard errors combine student-sampling replication and delete-one-item sensitivity, conditional on the retained item set and the fixed score conversion estimated

from the 2022 plausible-value gap. They are not the official OECD linking errors and do not cover all sources of model or anchor uncertainty. The published plausible values show widening of 24.1 and 20.6 points. The gap also widens in this common-item model, which does not use the 2025 plausible values. This supports the direction of the change, but does not quantify how much the revised imputation method contributes to the published increase.

**Figure 6. Change in the gymnasium-vocational gap from 2022 to 2025: published scores and identical items.**




Source: OECD PISA 2022 and 2025 public use files; author's calculations. Published change: sampling error only (the official link error is not yet released). Identical items: sampling error and item-sampling sensitivity combined, conditional on the retained items and the score conversion. 95% intervals.

CENTER FOR PSYCHOLOGY

## Why this matters beyond PISA

**These methods (calibration with sampling weights, linking to published item parameters, tests of whether items function equally across groups, and same-items comparisons over time) can inform national assessments that retain suitable item-response data and a defensible linking design. Adapted to repeated testing, they can help assess comparability and monitor performance on common tasks. Establishing whether a reform caused a change would additionally require an appropriate evaluation design.**

# Methods note

Model: generalised partial credit model (Muraki, 1992), two-parameter logistic for dichotomous items, scaling constant D = 1.7, weighted marginal maximum likelihood by EM on 61 quadrature points (41 for the common-item trend model). The

Montenegrin calibration correlates 0.84 with the OECD's international item locations across all 199 items; the RP62 thresholds in Annex A recover a score transform of about 168.3 points per theta unit, with a root mean squared residual of about 0.3 points. Uniform DIF: group differences in item location with slopes held equal throughout calibration and purification; two-group concurrent calibration; Wald test per item with the full covariance of each item's threshold differences, conditional on the other items and the group distributions; nominal 1% level; anchor purification; every calibration stage met the convergence criterion. Descriptive statistics: final student weights, 80 Fay-adjusted replicate weights. Standard errors for the item analysis are model-based and ignore the clustered design; the trend standard errors combine replicate refits with a delete-one-item jackknife, as described above. Missing codes: not administered, not reached and technical failure are treated as not presented; a score of zero is a zero.



**ABOUT THIS REPORT**

Independent analysis of the OECD's public PISA microdata. Analysis code and tables are available from the author.

Cite as: Kankaraš, M. (2026). Where vocational schools fall behind: item-level evidence from PISA 2025. Center for Psychology. https://doi.org/10.5281/zenodo.23389928

Dr Milos Kankaras is a psychometrician and policy analyst based in Podgorica, with twenty years of work on international large-scale assessments.

**Dr Milos Kankaras** · milos@centerforpsychology.me · centerforpsychology.me · miloskankaras.com · **ORCID** · **Google Scholar** · **LinkedIn**

# References

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159-176. https://doi.org/10.1177/014662169201600206

OECD. (2026a). PISA 2025 results (Volume I): Future-ready students. OECD Publishing. https://doi.org/10.1787/73451bc5-en

OECD. (2026b). PISA 2025 technical report [Draft chapters]. OECD. https://www.oecd.org/en/data/datasets/pisa-2025-database.html

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. Educational Measurement: Issues and Practice, 36(4), 52-61. https://doi.org/10.1111/emip.12165

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. Journal of Educational Measurement, 43(1), 19-38. https://doi.org/10.1111/j.1745-3984.2006.00002.x