PISA 2025 · A NEUTRAL READ

PISA 2025: what the results show, how far they can be trusted, and what might explain them

Full version: report, United States annex, and supporting sections.

Published 10 September 2026. Based on the OECD’s release of 8 September 2026.

ABOUT THIS REPORT

This report was written by AI agents: separate instances of Claude Fable 5, a large language model made by Anthropic. Each agent worked from its written brief and the materials the brief listed, and the briefs say nothing about who commissioned the work or why. The software the agents ran in did supply basic account details (a name, an email address and a short description of Bill McCallum’s newsletter project), enough in principle to identify him, though the briefs did not ask the agents to use them. Bill McCallum commissioned the report and reviewed the briefs before the agents ran; the briefs are published here.

Figures come from the OECD’s own data tables rather than its summaries, with the OECD’s standard errors and significance tests. A separate agent then reviewed the draft for bias and omission from several opposed standpoints and checked numbers against the data; its review and the changes made in response are in the supporting sections.

Before publication, further agents checked about 300 quotations of named people and organizations against their sources. About 40 were wrong in some way, from a sentence cut short to words that could not be found in the source; all were corrected, paraphrased or removed. The list of corrections is with the supporting sections. The short summaries headed “In brief” at the top of the longer sections were also added before publication, by agents working from each section’s text, and checked against it twice; the same list explains how.

Summary

On 8 September 2026 the OECD published the results of the ninth cycle of its Programme for International Student Assessment, the three-yearly test of 15-year-olds’ reading, mathematics and science that began in 2000. The headline is a decline, and the decline is real in the narrow sense that it appears in the OECD’s tables, in every section of a reading test unchanged since 2018, and in most of the seventy-odd systems that can be compared with 2022. Across the 35 OECD members with comparable data since 2015, mean reading scores fell 14.3 points between 2022 and 2025 (95% confidence interval −16.9 to −11.8) and mathematics fell 9.4 points (−12.0 to −6.7); science, at −2.8 points with a standard error of 3.2, did not change significantly. Since 2018, reading is down 24.6 points and mathematics 24.2, and both are at the lowest levels recorded in the series.

A reader who stops here should carry away four qualifications. First, the decline did not begin with the pandemic: on the aggregate of 23 countries with data since 2000, reading was flat until 2012 and has fallen in each cycle since, most steeply in the most recent one, though the two pre-2018 steps (−4.1 and −4.4) are each smaller than the link errors that apply to comparisons across the 2015 move from paper to computer (6.1 points for reading from 2012 to 2025) and the combined 2012-to-2018 fall of 8.6 points is only marginally significant (z about 2.3 on the 3.74 link error published for that comparison); the pre-pandemic decline is clearest at country level, in Finland, Australia, the Netherlands and Korea. Mathematics was flat until 2018 and then fell in two steps. Second, the fall is not uniform: it is concentrated in rich, previously high-scoring systems, while eight systems, seven of them outside the OECD and most starting from low levels, improved in all three domains between 2022 and 2025, and in that window students from the most advantaged quarter of families fell more than those from the least advantaged, the reverse of 2018 to 2022. Third, part of what has changed is how students behave during a two-hour test with no consequences for them: self-reported effort fell, hasty wrong answers rose, and declines were largest on long texts. The OECD treats this as part of reading itself: its framework counts the capacity to sustain attention on a long text as something the reading test measures, so a fall in that capacity is a real fall in reading. A reader who separates the skill of reading from how hard a student chooses to try on a test that doesn’t count will read the same tables differently, as showing that some of the decline comes from students trying less on the test rather than from their reading less well. Fourth, the 2025 science comparison runs across a revised framework, a newly adaptive design, a new platform and a new scaling contractor, none of which the published volume analyses, and should be held more loosely than the reading and mathematics trends.

On causes, the sections synthesised here reached no single answer, and this report does not supply one. Fourteen explanations were each stated as predictions before the data were checked. Four survived without contradiction: digital distraction, declining reading habits, declining effort on the test itself, and the OECD’s composite of these; they are not independent of one another, and the third is as much a decision about what PISA measures as an empirical finding. Four are contradicted as the main or sole cause while surviving in a limited role: the pandemic, which plausibly accounts for a third to a half of the mathematics

decline since 2018 and less of reading but not for the pre-2018 decline, the acceleration after 2022 or the cross-country pattern; changes in the student population, chiefly a larger share of immigrant students, worth 1.5 to 3 points on the average and a third or more of Germany's and Israel's reading declines; spending, whose level runs the wrong way and whose change was not tested; and a reversal of the long-run rise in cognitive test scores, which fails on timing. Five fit some countries and not others. Of the explanations that dominate public debate, the evidence on smartphones is compatible with a large role and does not establish one; for funding and curriculum the direct evidence is small, untested on the quantity that matters, or specific to particular countries.

Terms used in this report

PISA and its cycles. The Programme for International Student Assessment tests 15-year-olds every three years (2000, 2003, 2006, 2009, 2012, 2015, 2018, 2022, 2025; the 2021 cycle was delayed a year by the pandemic). Each cycle tests reading, mathematics and science, with one of the three as the "major domain" that gets the most test time and the most detailed reporting. Science was the major domain in 2025, mathematics in 2022, reading in 2018.

PISA scale points. Scores are reported on a scale set, when each subject was first assessed, to have an OECD mean of 500 and a standard deviation of 100. A 10-point change is therefore a tenth of a standard deviation. The OECD sometimes converts points into "years of learning"; its own estimate of a year's progress ranges from about 10 points in Serbia to about 25 in the United Kingdom, so the conversion is rough and country-specific.

Standard error, confidence interval, significance. Every PISA figure is an estimate from a sample of a few thousand students per country, so it carries a standard error (SE). The 95% confidence interval is the estimate plus or minus about twice the SE. A change is called "significant" when it is more than 1.96 times its SE, which is the rule the OECD uses to print values in bold. "Not significant" means the data cannot distinguish the change from zero, not that the change is zero.

Link error. Each cycle uses a partly new set of test questions, so results have to be placed on the same scale as earlier cycles using the questions common to both (the "trend items"). That anchoring is itself estimated and carries its own error, called the link error. It is added to the sampling error whenever a change between cycles is tested, which is why a change of a few points can be non-significant even though each year's mean is known to within a point. Link errors are larger for comparisons that span more cycles or a design change; in this report the science link error for 2022 to 2025 is about three times the reading one, so science changes of the same size as reading changes come out as not significant.

Plausible values. PISA does not compute a single score per student. Each student answers only a subset of the questions, so their proficiency is estimated as a distribution, from which several "plausible values" are drawn; country statistics average over these draws. The consequence for a reader is that published estimates include measurement uncertainty as well as sampling uncertainty.

OECD averages. The OECD reports several averages depending on which members have comparable data. "OECD-38" is all current members, available only for 2025 in the trend tables. "OECD-35" excludes Costa Rica, Luxembourg and Spain and runs from 2015; it is the one the OECD's text uses. "OECD-23" is the 23 members with data in every cycle since 2000 and is the only series spanning the whole period. The three differ in level by several points, so a number should always be read with its aggregate.

Percentiles (p10, p90) and the inter-decile range. The 10th percentile is the score below which the lowest-scoring tenth of students fall; the 90th is the score above which the top tenth score. The difference between them, the inter-decile range, is the report's measure of spread within a country. The mean can be flat while the bottom falls and the top rises, which is what the United States shows.

Proficiency levels. PISA divides the scale into levels. Level 2 is the OECD's "baseline" of proficiency, the point at which students can apply basic skills to simple real-world situations; "low performers" are those below Level 2. Levels 5 and 6 are the top; "top performers" are those at Level 5 or above.

ESCS. PISA's index of economic, social and cultural status combines parents' occupations, parents' education and home possessions (including books and study space) into one scale per student. "Advantaged" and "disadvantaged" students in this report are the top and bottom quarters of the index within each country. The "socio-economic gap" is the difference in mean score between those two quarters.

Immigrant background. A student is first-generation if both the student and both parents were born abroad, second-generation if the student was born in the country but both parents were not. "Non-immigrant" means at least one parent was born in the country.

Sampling standards, response rates and asterisks. PISA requires that at least 85% of sampled schools take part (before replacement schools are substituted), that at least 80% of sampled students in participating schools take the test, and that no more than 5% of the target population be excluded (for disability, limited language or inaccessible schools). When a country misses these, an adjudication group decides whether to report its results, and if so with what warning. An asterisk in the OECD's tables marks a country that was reported despite missing a standard. Non-response tends to be higher among low-achieving students, so the usual direction of bias is upward.

Coverage Index 3. The share of all 15-year-olds in a country that the PISA sample represents, after removing those not in school (or not in grade 7 or above) and those excluded. In most wealthy countries it is 0.85 to 0.95; in some lower-income countries it is below 0.5, so PISA describes only the enrolled half of the age group. Changes in coverage between cycles can move a country's mean without any change in what students know.

Test effort and "hasty" responses. PISA has no consequences for the students who take it. Since 2018 the test has asked students how much effort they put in (the "effort thermometer", a 10-point scale) and has recorded response times. A "hasty" response is one that is both fast and wrong; the OECD uses the rate of hasty responses as an indicator of disengagement.

Adaptive testing. From 2018 (reading), 2022 (mathematics) and 2025 (science), the computer routes each student to easier or harder blocks of questions depending on how they did on the first, common section. This improves precision but means students in different countries and years see different questions, which raises the question of whether the design itself could create an apparent trend; the OECD's check on this is described in the report.

Mode effect. The change in difficulty when the same question is asked on paper versus on screen. PISA moved from paper to computer in 2015, and the adjustment for this makes the 2012-to-2015 comparison the weakest link in the series.

Item response theory, Rasch model, DIF. The statistical model that converts right and wrong answers into scale scores. The Rasch model is the simplest version, used by PISA until 2012; a more flexible model has been used since 2015. Differential item functioning (DIF) means a question is relatively harder or easier for students in one country than the model assumes, which is the basis of the main academic criticism of PISA's scaling.

Other assessments. TIMSS (Trends in International Mathematics and Science Study) tests grade 4 and grade 8 students every four years on curriculum-based content; its 2023 grade 8 cohort is the same birth cohort PISA tested in 2025. PIRLS tests grade 4 reading every five years. PIAAC tests adults' literacy and numeracy. NAEP is the United States' national assessment, with a "main" series at grades 4, 8 and 12 and a "long-term trend" (LTT) series at ages 9, 13 and 17 using instruments unchanged since the 1970s.

Correlations. A Pearson correlation measures a straight-line relationship and is sensitive to extreme cases; a Spearman correlation uses ranks and is not. Both run from −1 to +1. In this report correlations across countries are labelled, and all are observational: they describe an association among a few dozen countries and cannot show cause.

Standard deviation (SD) as a unit of effect. Research studies report effects in fractions of a standard deviation so that different tests can be compared. On the PISA scale 0.1 SD is 10 points. An effect of 0.02 to 0.05 SD, typical of the phone-ban studies cited here, is 2 to 5 PISA points.

What the 2025 results show

IN BRIEF (summary added before publication; see Appendix H)

On the OECD-23 average, reading fell 16.1 points and mathematics 10.9 in 2022-2025; the 4.9-point science change is not significant. Since 2018 reading and mathematics are each down 27.0 points and science 7.5. On the OECD-35 average, which the OECD's text uses, the share of students below Level 2 (the OECD's proficiency baseline) rose in 2022-2025 from 25.7% to 30.4% in reading and 30.0% to 33.9% in mathematics. The gap between the top and bottom socio-economic (ESCS) quarters narrowed in that period, which the OECD attributes to falling scores among advantaged students rather than gains among the disadvantaged. It had widened in 2018-2022, but only the narrowing rests on a consistent index, because the 2025 data restate the 2015 and 2018 ESCS values without documentation. Eight systems improved significantly in all three domains in 2022-2025 and nine declined in all three. The OECD attributes about 1.5 of the 25-point OECD-average reading decline since 2018 to the rising immigrant share.

All figures in this section come from the StatLink workbooks of *PISA 2025 Results (Volume I)* (OECD, 2026, https://doi.org/10.1787/73451bc5-en). Standard errors are the OECD's; for changes between cycles they already include the link error defined above, which the next section discusses further. "Significant" means the estimate exceeds 1.96 times its standard error, the rule the OECD uses to bold values in its tables.

The OECD publishes three "OECD averages". The 38-member average is available only for 2025 in the trend tables. The "OECD-35" average (all members except Costa Rica, Luxembourg and Spain) runs from 2015 and is the one the OECD's text uses. The "OECD-23" average (members with data in every cycle since 2000) is the only aggregate spanning the whole series. Table 1 gives it.

Table 1. OECD-23 mean scores by cycle (standard error in parentheses)

Cycle	Reading	Mathematics	Science
2000	499.5 (0.7)	–	–
2003	497.4 (0.6)	502.1 (0.6)	–
2006	495.5 (0.7)	501.1 (0.6)	502.7 (0.6)
2009	499.4 (0.6)	501.7 (0.6)	505.7 (0.6)
2012	501.4 (0.6)	498.7 (0.6)	504.7 (0.6)
2015	497.2 (0.6)	495.5 (0.5)	497.0 (0.5)
2018	492.8 (0.5)	496.4 (0.5)	493.5 (0.5)
2022	481.9 (0.6)	480.3 (0.5)	490.9 (0.5)

Cycle	Reading	Mathematics	Science
2025	465.8 (0.6)	469.4 (0.6)	485.9 (0.6)
Change, first cycle to 2025	−33.7 (6.8)	−32.7 (5.7)	−16.8 (4.9)
Change 2018 to 2025	−27.0 (2.0)	−27.0 (2.7)	−7.5 (3.6)
Change 2022 to 2025	−16.1 (1.3)	−10.9 (1.4)	−4.9 (3.2), not significant

Source: Tables I.B1.2a.36-38.

The three subjects have different shapes. Reading was flat from 2000 to 2012 and then declined in each cycle: −4.1 to 2015, −4.4 to 2018, −10.9 to 2022 and −16.1 to 2025. Mathematics stayed between 502 and 495 from 2003 to 2018, then lost 16.1 points to 2022 and 10.9 to 2025. Science peaked in 2009 and has moved little since 2018: the 2022-to-2025 change is not significant on either aggregate, though the OECD-35 decline over 2015 to 2025 (−6.9, SE 3.5) is; the fitted ten-year trends are −27.0, −23.9 and −6.7 points per decade (Table 2).

Table 2. OECD-35 mean scores and changes (the aggregate used in the OECD's text)

	Reading	Mathematics	Science
2015	490.3 (0.5)	487.2 (0.4)	490.8 (0.4)
2018	487.6 (0.4)	489.7 (0.4)	489.2 (0.4)
2022	477.4 (0.5)	474.8 (0.4)	486.8 (0.5)
2025	463.0 (0.5)	465.5 (0.5)	483.9 (0.5)
Change 2022 to 2025	−14.3 (1.3)	−9.4 (1.4)	−2.8 (3.2), n.s.
Change 2018 to 2025	−24.6 (1.9)	−24.2 (2.6)	−5.2 (3.6), n.s.
Change 2015 to 2025	−27.3 (3.9)	−21.8 (3.1)	−6.9 (3.5)
Fitted trend 2015-2025, points per decade	−27.0 (3.3)	−23.9 (3.1)	−6.7 (3.1)

Source: Tables I.B1.2a.36-38.

The OECD's headline sentences check against these cells, with one qualification: its statement that 2025 performance was "significantly below the mean performance observed in any earlier assessment (except for science in PISA 2022)" holds for the OECD-23 series, whereas on the OECD-35 series science in 2018 is also not significantly different from 2025 (−5.2, SE 3.6).

Table 3 counts the systems that moved significantly (the annex tables carry one further comparable system, Cyprus, whose 6.5-point mathematics decline is significant and which the text omits without saying why). Eight systems improved significantly in all three domains between 2022 and 2025: Cambodia, Georgia, Jordan, Montenegro, the Philippines, Thailand, Türkiye and the United Arab Emirates. Nine declined in all three: Argentina, Denmark, Guatemala, Hong Kong (China), Israel, Latvia, Malta, Serbia and Slovenia. The largest three-year reading decline is Latvia's (−44.9, SE 3.4); the largest since 2018 are Iceland (−51.7), Slovenia (−50.6), Latvia (−49.0), Norway (−46.6), Finland (−45.8), Cyprus (−44.9), Hong Kong (−44.2) and the Netherlands (−43.4).

Table 3. Systems with significant change in mean score (Tables I.B1.2a.36-38, Table I.2.6)

Base year → 2025	Domain	n	Up	Down	No significant change
2022	Science	74	15	11	48
	Reading	74	8	42	24
	Mathematics	74	8	38	28
2018	Science	74	17	21	36
	Reading	74	5	56	13
	Mathematics	75	9	56	10
2015	Science	61	11	23	27
	Reading	61	3	47	11
	Mathematics	62	6	40	16

Means conceal movement at the two ends of the distribution. In science, the OECD-35 10th percentile fell 12.6 points between 2018 and 2025 while the 90th did not move (+2.4, SE 3.6); the gap between them, 259 points in 2025, is the widest the OECD has measured, with the caveat that science became adaptive in 2025 and adaptive designs change precision at the tails, so percentile changes are more design-sensitive than means. In reading the whole distribution moved down between 2022 and 2025 (bottom −15.5, top −12.8); in mathematics the bottom fell 13.6 points against 5.5 at the top. The share below Level 2, the OECD's baseline of proficiency, rose on the OECD-35 average from 25.7% to 30.4% in reading and 30.0% to 33.9% in mathematics over 2022 to 2025, and did not change significantly in science; since 2015 it has risen by about 9.5 points in reading and 9.2 in mathematics while the top share fell by about 2.3 in each.

The socio-economic pattern reversed between the two most recent windows. PISA's index of economic, social and cultural status, or ESCS, combines parental occupation, parental education and home possessions into one scale, and the OECD compares the top and bottom quarters on it within each country. Between 2018 and 2022 the OECD-35 gap between these quarters widened in every domain (reading 87.9 to 93.4 points; mathematics 86.7 to 93.7). Between 2022 and 2025 it narrowed, by 12.8 points in science, 16.9 in reading and 11.1 in mathematics, and the arithmetic is dominated by the top: advantaged students lost 7.6 points in science, 20.4 in reading and 13.8 in mathematics, while disadvantaged students did not change significantly in science (+5.2, SE 3.2) or mathematics (−2.6, SE 1.5) and fell 3.5 in reading; the OECD's summary is that the gap narrowed "due to declining average scores among advantaged students rather than by improvements among disadvantaged students". Relative to 2015, both quarters are roughly 20 points lower in reading and mathematics. This finding is used repeatedly below and carries a caveat: the 2025 workbook restates the 2015 and 2018 ESCS-quarter values without documentation, the restated gaps differing from the 2022 workbook's by 1 to 3 points for most countries in 2018 (7 for Israel) and by more than 20 for Qatar and the United Arab Emirates in 2015, while the 2022 values are identical; the 2022-to-2025 comparison rests on a consistent index and the 2018-to-2022 widening does not.

Girls lead boys by 30 points in reading and boys lead girls by 13 in mathematics on the 2025 OECD average; both gaps widened between 2022 and 2025 because boys fell more in reading (−17.1 against −11.7) and girls more in mathematics (−11.2 against −7.4). The share of students with an immigrant background across 29 OECD members rose from 12.8% in 2015 to 18.0% in 2025, fastest in Portugal (7.0% to 18.3% over 2018 to 2025), Malta (8.8% to 17.1%), the United Kingdom (19.8% to 27.7%), Germany (22.2% to 29.2%), Norway (12.4% to 18.7%), Slovenia (8.9% to 14.1%) and Iceland (5.6% to 10.1%). The unadjusted science gap between immigrant and non-immigrant students widened from 37.2 to 42.7 points between 2022 and 2025, but after accounting for ESCS and home language the change (−2.7, SE 1.8) is not significant, which is what the OECD's "remained unchanged" refers to. The OECD attributes about 1.5 of the 25-point reading decline since 2018 to the rising immigrant share (Annex A1, p. 311); on the OECD-29 aggregate immigrant students' own reading fell 24.6 points over 2018 to 2025, about the same as non-immigrants' 22.3, so the compositional effect is a share effect, not a widening gap.

Several small discrepancies between the OECD's text and its tables were found. Annex A1 says the reading gap between ESCS quarters "narrowed in 36 countries and economies" over 2018 to 2025, whereas Table I.B1.2b.23 yields 31 significant narrowings, on a criterion the report does not state; the list of systems that raised their top-performer share in science since 2015 omits Cyprus, Luxembourg and Romania, which are significant in the table; and mean scores for earlier cycles are identical between the 2025 and 2022 workbooks except for the Palestinian Authority, whose 2022 results were recalculated to 2025 coverage, whereas the ESCS-quarter values are restated as described above.

How much weight the results can bear

> **IN BRIEF** (summary added before publication; see Appendix H)
>
> Science is the least secure trend. Its link error for comparisons with 2018 and 2022 is about three times reading's, so OECD-average science changes need 6 to 7 points for significance. The 2025 science test also had a revised framework, adaptive design and a new scaling contractor, whose effects on comparability await a Technical Report. Students reported less effort and gave more fast, wrong answers, which the OECD counts as part of what reading assesses. The report's validity section puts the part of the reading decline due to less effort at a small fraction to roughly a third, though the data cannot separate effort from skill. An unreplicated check on reused reading questions, absent for mathematics and science, is the strongest evidence that the reading decline is not a product of test design. No single measurement issue can explain a 22- or 25-point average decline, and several national declines carry sampling uncertainty about as large as their three-year changes.

PISA is not a test of what students were taught. Its framework assesses "how well students can extrapolate from what they have learned, both in and outside of school, and apply it to real-life problems" (Volume I, p. 47) and samples by age rather than grade; its curriculum-based counterpart, TIMSS, samples grades 4 and 8 and specifies content negotiated with participating countries, so a country's rank can differ between the two by design. Declines on PISA are declines on PISA's construct.

Scores come from an item-response model in which each student's proficiency is not a single number but a set of "plausible values", draws from a distribution reflecting the student's answers and background, which is why published estimates carry measurement as well as sampling uncertainty. Re-scalings under alternative models move country means by at most 3 to 4 points, and because one model is applied across cycles the error largely cancels in trends.

Three things specific to the 2025 cycle need more care. The first is the link between cycles. Each cycle's results are placed on an earlier cycle's scale using shared "trend items", and the uncertainty in that placement, the link error, is added to every cross-cycle comparison. The Reader's Guide gives 2.9 points as the science link error between 2022 and 2025. The pairwise link errors can be recovered from the published standard errors, since the implied value is identical across countries: for comparisons with 2025, from 2022, science 3.12, reading 1.09, mathematics 1.22; from 2018, 3.51, 1.83 and 2.55; from 2012, 6.06, 6.11 and 3.78. The science figure in the tables (3.1) is not the 2.9 quoted in the text, a small discrepancy whose source cannot be identified. The science link errors for the two most recent comparisons are about three times reading's, so a science change of 6 to 7 points is needed for significance at the OECD-average level where 2.5 to 3 points suffices in reading, which is why the OECD can describe science as stable while having reported significant science declines of similar size in earlier cycles. The academic literature holds that link errors understate trend uncertainty because they ignore country-specific item behaviour; if by a factor of two, a single country's 2018-to-2025 change would carry perhaps 5 to 8 points of linking uncertainty.

The second is the science test and who scaled it. Science was the major domain in 2025 and was assessed under a revised framework that merges two of the 2015 competencies, adds one on evaluating scientific information and adds an environmental-science subscale; the volume does not say how many items were carried over or how the new competency was linked to the old scale. The science test was also, on the volume's description, adaptive for the first time, with a linear version for 25% of test-takers. And the contractor changed: Volume I states that the Australian Council for Educational Research (ACER) was the Core 1 contractor for 2025, responsible for "instrument development, scaling and analysis, and all data products", with Open Assessment Technologies as subcontractor for the platform and the "migration of trend units", and Oxford University Press as contractor for the science framework (pp. 9, 322). ACER ran PISA from 2000 to 2012 under a one-parameter scaling model; ETS ran 2015 to 2022 under the two-parameter model with partial invariance described above. Volume I does not say whether the 2015-to-2022 scaling model, item parameters and invariance decisions were carried over, and analyses none of these changes as a comparability issue; all are deferred to a Technical Report that had not appeared at the time of writing. The large science link error is consistent with a scale re-anchored across a framework revision, though that is an inference from the tables rather than an OECD statement.

The third is sampling. PISA requires exclusions of at most 5% of the target population, school response of 85% before replacement, and student response of 80%. Six countries missed at least one standard in 2025 and carry an asterisk (Table 4). Beyond these, Catalonia's results were suppressed after within-school exclusions rose to 23.2%; Uzbekistan's reading and mathematics were withheld for undescribed "inconsistencies"; Viet Nam is excluded from all trends because it moved from paper instruments derived from PISA 2012 to computer; Guatemala and Paraguay made a similar transition and their trends are reported "with caution". No table of 2025 exclusion and response rates for all systems is published.

Table 4. Systems the OECD asterisks in 2025 (Reader's Guide, pp. 17-20)

System	Exclusions	School response, before / after replacement	Student response	OECD's statement on bias
Canada	–	79% / 82%	77%	upward bias "unlikely to exceed 15 points in science"
Netherlands	9.3%	66% / 88%	78%	up to "about 11 score points in science"
New Zealand	8.1%	45% / 52%	76%	"may be over-estimated by about 10 to 15 score points"
Norway	10.4%	–	–	exclusions "may bias performance results upwards"
Albania	–	76%	–	limited reporting; science lower bound 428; no trend
United States	6.9%	45% / 54%	76%	limited reporting; "not possible to rule out bias … nor to determine its most likely direction"; over 40% not given the questionnaire

A separate annex treats the United States. The 2022 baseline had more problems: thirteen systems fell short of standards in 2022 (including the six asterisked above plus Australia, Denmark, Hong Kong, Ireland, Latvia and the United Kingdom), with school response after replacement as low as 63% in the United States. Volume I compares all of them with 2025 without adjustment or a 2022 asterisk, judging only that New Zealand's bias "is unlikely to affect comparisons between PISA 2022 and PISA 2025 in meaningful ways". Where non-response bias has been quantified by linking to administrative records (England, New Zealand), non-respondents were lower-achieving and the bias upward, 7 to 10 points, the size of a flagged system's own three-year change; six such systems biased by 10 points in one cycle and not the other would move the OECD average by about 1.7 points (the validity section's calculation). Rising exclusions work the other way: excluded students are mostly low-achieving, so Norway's and the Netherlands' declines are, if anything, understated, while Denmark, whose exemptions fell from 11.7% in 2022 to 8.3% in 2025 as more dyslexic pupils were included, has the reverse problem, which its national researchers judge had a small effect on scores without explaining the 29-point reading decline (claims inventory B15.3).

One earlier baseline needs the same treatment. Sweden's 2018 sample had the highest exclusion rate of that cycle, 11.1% (PISA 2018 Table I.A2.1), which the OECD attributed at the time to "a large and temporary increase in immigrant and refugee inflows"; its Coverage Index 3 fell from 0.94 in 2015 to 0.86 in 2018 and recovered to 0.89 in 2022 and 0.92 in 2025 (Table I.A2.1). The Swedish National Audit Office (Riksrevisionen, RiR 2021:12) found deficiencies in the follow-up of exclusions and concluded that the migrant-inflow explanation offered by the agency, the OECD and the government "is not a valid explanation". Sweden's 2018-to-2025 declines (reading −39, mathematics −38) are therefore measured from the cycle with the narrowest tested population in its series, and Sweden is used below as a leading case against the pandemic, curriculum and spending explanations; its 2015-to-2025 declines (−34 and −30) start from a fully covered cycle. The baseline effect is unquantified and does not remove the Swedish decline.

Testing dates also changed between 2015 and 2025 in ten systems (Austria, Brunei Darussalam, Cambodia, Ireland, the Netherlands, Serbia, England, Northern Ireland, Scotland and the United States), which Box I.2.2 exploits to estimate yearly learning gains; a date shift changes the grade composition of the age-eligible sample, and these systems' trends run across it.

One further feature bears directly on interpretation. PISA is low-stakes; students get no feedback and face no consequences, and a financial incentive raised US students' scores on a PISA-style test by about 0.2 standard deviations in one experiment, with no effect in Shanghai. For trends, the evidence is the OECD's own: self-reported effort fell by 0.5 points on a 10-point scale between 2018 and 2025 across the OECD-35; "hasty" (fast and wrong) responses on reading items rose by more than 4 percentage points to about 9% while unreached items fell; hasty test-takers on the reading-fluency task rose from 6.6% to 11.4%, most of it after 2022; declines were larger on long texts and in later sections. Against this, questionnaire straightlining fell, first-hour versus second-hour differences did not widen, and fatigue within the linear science test is modest.

The OECD's position is that declining "capacity to direct and sustain attention" is "part of what the 'reading' test in PISA assesses" (p. 319), so it counts these changes as real changes in the construct. That is a defensible reading of the framework and an interpretive choice; a reader who regards willingness to try on a low-stakes test as separate from reading skill will see the same tables as evidence that part of the decline reflects less effort on the test rather than a loss of reading skill. The validity section's bound puts that part at anywhere between a small fraction and roughly a third of the 25-point reading decline; the data cannot separate the two, and the hasty-response indicators rest on timing thresholds re-normalised each year after the platform change, with no published sensitivity check.

The strongest evidence that the reading decline is not a design artefact is Annex A1's item-level analysis: the 2025 reading test reused the 2018 tasks, percent-correct on those items fell in the non-adaptive first section as well as the adaptive stages, and country-level changes in percent-correct correlate at 0.90 to 0.91 with changes in mean scores. This is a genuine check, unreplicated, with no counterpart for mathematics or science.

A different line of criticism concerns governance rather than measurement: Meyer and Benavot (2013) frame PISA as an instrument of "global educational governance"; the 2014 open letter to the OECD signed by about 80 academics argued that the three-year cycle shifts attention to short-term fixes and narrows curricula; Sjøberg (2015) argues that curriculum-free, translated items privilege a genre of reading; Zhao (2020) holds that these criticisms "have been largely ignored". The OECD has changed technical practice in response to its technical critics and has not conceded ground on construct or governance: Volume I reasserts that reading includes sustained attention, its Preface moves from results to prescriptions on curriculum, teachers and devices, and the volume that warns about screens introduces a "Learning in the Digital World" assessment domain. Two consequences for this report follow: its vocabulary ("baseline proficiency", "hasty readers") is the OECD's, with "hasty" resting on re-normalised timing thresholds; and its predict-then-check method can test only what PISA measures, which favours explanations PISA's questionnaire captures (effort, in-school device use, belonging) over those it does not (out-of-school phone hours, curriculum content, spending changes).

Taken together: no single measurement issue is large enough to account for a 22- or 25-point average decline; several national declines carry sampling uncertainty of the order of their own three-year changes; and the science trend is the least secure of the three.

The picture since 2000 and the country trajectories

> **IN BRIEF** (summary added before publication; see Appendix H)
>
> The trajectory analysis grouped 69 of 89 systems by significant changes in at least two subjects over 2012-2018, 2018-2022 and 2022-2025. Eleven show a long decline from 2012 and eight have risen since 2018. The long-decline, decline-since-2018 and gradual-decline groups are almost all OECD members; the rising, rebound and stable groups are all non-OECD. Systems scoring higher in 2018 tended to fall more (Spearman correlation −0.34 across 75 systems), but not significantly within the OECD (−0.16). Only three of sixteen systems that rose in some period after 2018 pass all checks. Of the features tested, only weeks of full pandemic school closure differ significantly across groups, and the groups that declined most had the shortest closures, because long closures were concentrated where several low-scoring systems rose; with GDP per capita controlled, the effect is indistinguishable from zero. No feature separates the groups within the OECD. Starting level and OECD membership do distinguish them; the tables cannot show why.

The trajectory analysis classified every system with comparable data by its changes in three windows (2012 to 2018, 2018 to 2022, 2022 to 2025) and over 2018 to 2025: a "fall" if at least two of three domains fell significantly and none rose, a "rise" symmetrically, "stable" otherwise, so that science alone, with its larger link error, cannot drive a verdict. Of 89 systems with some comparable data, 69 can be classified (Table 5).

Table 5. Trajectory groups, 2012-2025

Group	Rule	n (of which OECD)	Members
A. Long decline	fall 2012-18 and in a later window, no rise	11 (9)	Australia, Bulgaria, Costa Rica, Finland, Germany, Greece, Hong Kong (China), Hungary, Latvia, Netherlands, Switzerland
A2. Early fall, then stable	fall 2012-18, stable after	2 (2)	Italy, Korea
B. Decline since 2018	fall 2018-22 and 2022-25	8 (7)	Belgium, Canada, Denmark, France, Iceland, Morocco, Norway, Portugal
C. Fall, then stabilised	fall 2018-22, stable 2022-25	12 (5)	Cyprus, Estonia, Indonesia, Kosovo, Macao (China), Malaysia, Mexico, Moldova, North Macedonia, Poland, Slovak Republic, United Kingdom
D. Recent fall only	no fall 2018-22, fall 2022-25	10 (5)	Argentina, Austria, Chile, Croatia, Czechia, Ireland, Israel, Malta, Serbia, Singapore
D2. Gradual decline	no window falls, but 2018-25 falls	4 (4)	Colombia, Lithuania, New Zealand, United States

Group	Rule	n (of which OECD)	Members
E. Rise, then fall	rise earlier, fall 2022-25	5 (2)	Guatemala, Paraguay, Peru, Slovenia, Sweden
F. Rebound	fall 2018-22, rise 2022-25	2 (0)	Montenegro, Thailand
G. Rising	rise after 2018, no fall, 2018-25 rise	8 (0)	Brunei Darussalam, Cambodia, Chinese Taipei, Dominican Republic, Georgia, Philippines, Qatar, United Arab Emirates
G2. Rise not sustained	as G but 2018-25 not a rise	2 (2)	Japan, Türkiye
S. Stable	none of the above	5 (0)	Brazil, Kazakhstan, Romania, Saudi Arabia, Uruguay

Source: trajectory analysis, Table 2; data from Tables I.B1.2a.36-38 and PISA 2022 Volume I Tables I.B1.5.4-6. Twenty further systems have only partial data.

Some placements depend on the rules. Japan is in G2 because its 2018-to-2022 rise (reading +12, science +17.5) reversed in 2022 to 2025 (−13.1, −10.2, −9.0, reading significant), while its coverage index fell to 0.86, the lowest in its series. The United States is in D2: with standard errors of about 7 points only one of its window changes, reading 2022 to 2025 (−14.4, SE 7.3), reaches significance, and only just, while its 2018-to-2025 changes in reading (−15.8) and mathematics (−15.4) are significant (the trajectory section's −13.8, −15.6 and −15.0 are transcription errors; CSV values are used here). Table 6 gives the series for frequently discussed systems; B-S-J-Z (China), with data only for 2018 and 2025, fell 29 points in reading and rose 21 in mathematics, which the OECD does not explain.

Table 6. Reading and mathematics means for selected systems (bold: significant change to 2025)

System	Reading 2012	2018	2022	2025	2018→2025 (SE)	Maths 2012	2018	2022	2025	2018→2025 (SE)
Finland	524	520	490	474	**−46** (3.6)	519	507	484	469	**−39** (3.9)
Sweden	483	506	487	466	**−39** (4.7)	478	502	482	464	**−38** (4.5)
Netherlands*	511	485	459	441	**−43** (4.3)	523	519	493	483	**−36** (4.7)
Germany	508	498	480	465	**−33** (5.0)	514	500	475	464	**−36** (4.8)
Estonia	516	523	511	499	**−24** (3.6)	521	523	510	508	**−16** (3.8)
Poland	518	512	489	482	**−30** (4.3)	518	516	489	484	**−32** (4.6)
United Kingdom	499	504	494	494	**−10** (3.8)	494	502	489	488	**−14** (4.2)
Korea	536	514	515	501	**−13** (4.9)	554	526	527	522	−4 (5.6)
Japan	538	504	516	503	−1 (5.0)	536	527	536	525	−2 (5.5)
Singapore	542	549	543	535	**−15** (3.1)	573	569	575	563	−6 (3.4)
Canada*	523	520	507	490	**−30** (3.1)	518	512	497	485	**−27** (4.0)
United States*	498	505	504	490	**−16** (7.1)	481	478	465	463	**−15** (6.8)
Türkiye	475	466	456	472	+6 (3.7)	448	454	453	462	+8 (4.2)
Hong Kong (China)	545	524	500	480	**−44** (4.6)	561	551	540	522	**−29** (4.9)

Source: Tables I.B1.2a.36-37; asterisks mark systems the OECD annotates in 2025. Science is omitted because of its larger link error; the trajectory analysis gives the full series including science.

The first pattern in the groups is OECD membership: groups A, A2, B, D2 and G2 are almost entirely OECD, groups G, F and S entirely non-OECD, partly composition and partly the finding that the systems that rose after 2018 are, with the exceptions of Chinese Taipei, Qatar and Brunei, low- or middle-performing systems starting below 460 points. Across all 75 systems with 2018-to-2025 data the 2018 level correlates with the subsequent three-domain change at −0.34 (Spearman, $p = 0.003$; Pearson −0.47, the figure the trajectory section reports without labelling it), and within the 37 OECD members at −0.16 (Spearman, $p = 0.36$; Pearson −0.23). The Pearson value is pulled by Zambia and Cambodia, whose 2018 baselines are PISA-for-Development results on coverage of 0.36 and 0.28; without them the Spearman correlation is −0.29. Starting level discriminates, but the "regression to the mean" reading offered below rests on the weaker statistic.

The rises were checked for significance in more than one domain, consistency across windows, and stability of the coverage index (Coverage Index 3 is the weighted number of students the sample represents divided by the total number of 15-year-olds; where many are not in school, PISA describes only those who are). Only three of sixteen risers (the Philippines, Georgia, the United Arab Emirates) pass all three checks. Türkiye and Montenegro show real but recent rises against flat coverage; Cambodia's coverage index is 0.43; Chinese Taipei and Qatar rose between 2018 and 2022 and have been flat since; the Dominican

Republic's, Thailand's and Jordan's rises coincide with falls in coverage; Zambia's and Uzbekistan's cannot be assessed; and Georgia's, Türkiye's and Montenegro's rises are largest in science, where the link error is largest.

The trajectory analysis then asked whether documented system features discriminate between the groups: GDP per capita, cumulative spending per student, immigrant share, age of first tracking, weeks of pandemic school closure (from UNESCO's monitoring database), the share of decisions taken at school level, and national phone policy. Kruskal-Wallis tests across the seven groups with five or more members found one feature that differs significantly, full closure weeks ($p = 0.009$), in the direction opposite to the usual hypothesis: the groups that declined most, A and B, had the shortest closures (medians 15 and 9 full weeks) and the rising group the longest (25); Sweden and Iceland recorded no full closure weeks and the Philippines 75. This is a composition effect: long closures were concentrated in Latin America and South and Southeast Asia, where several low-scoring systems rose. Once log GDP per capita is controlled the closure coefficient falls to +0.20 points per week (SE 0.16), indistinguishable from zero, and within the OECD, where closures ranged from 0 to 53 full weeks, there is no relationship at all. The UNESCO measure is coarse (it records the United States as zero full weeks and 77 partial weeks), does not separate primary from secondary closure, and omits Hong Kong, Macao, Chinese Taipei and Kosovo, although the 2025 cohort was aged 10 to 11 when the pandemic began.

GDP per capita, spending per student and immigrant share all correlate with the size of the decline across all systems (Spearman −0.25, −0.28 and −0.32 with the 2018-to-2025 change), but this is largely the same fact as "higher-scoring systems fell more", and within the OECD the correlations weaken to non-significance. Governance and tracking do not discriminate at all: the share of decisions taken at school level ranges from 8% to 92% and correlates with the change at −0.03, and early-tracking systems appear in three groups. Group B, the systems that fell in both windows since 2018, combines wealth, short closures, late tracking and the highest immigrant share (median 18%), against a median three-domain change of −26 that the OECD's decomposition says composition cannot explain.

Phone policy cannot discriminate between trajectories that end in spring 2025, for a reason of timing: of the 46 classified systems with a coding, only three had a national ban in force before 2024 (France 2018, Italy 2007, the United Arab Emirates 2018), in groups B, A2 and G, and the bans that dominate current debate took effect after the test window; their effect will first be observable in PISA 2029.

Stated plainly: none of the features examined separates the trajectory groups once income is accounted for, and none separates them within the OECD. What does discriminate is where a system started and whether it is an OECD member; whether that reflects regression to the mean, a ceiling on what low-stakes testing can measure among high-performing adolescents, engagement effects concentrated in some populations, or construct changes that affected high-scoring systems differently, cannot be decided from the tables.

What has been claimed publicly, and by whom

The claims inventory compiled reactions from the OECD and from twenty-five countries in the day after the release. The OECD's own account should be kept separate from what others have said.

The OECD's Preface states that the weakening "predates the COVID-19 pandemic in many countries" and that "increased digitalisation, longer screen time and falling rates of reading for enjoyment have coincided with weaker literacy outcomes"; the Foreword says "we cannot say for sure" but that the shift towards digital reading "may be compromising the ability of students to engage with complex texts". The body is more careful: Box I.2.3 sets out four levels of explanation (attention and effort; classroom disruption; diversity and governance; reading habits, digital lives, the pandemic and absenteeism) that it calls "complementary rather than competing" and states that "the evidence does not point to one simple explanation"; Annex A1 says its attitude correlations "do not constitute proof of causality" and that demographic change is "never the main driving force". At launch events OECD officials named screen time, hasty reading and reading for pleasure as the leading considerations, while one senior analyst cautioned that no causal relationship between phone use and academic performance can be established.

Third-party explanations were counted by distinct speaker or organisation. Smartphones, social media and screen time were offered by 16, including ministers in Sweden ("screens are the poison in our society"), France (social media as a "weapon of mass intellectual destruction"), Spain, Denmark, Austria, Australia and Korea. Socio-economic inequality and underfunding were offered by 14, mostly unions, opposition parties and researchers, and dominated the German debate. Curriculum and pedagogy claims (knowledge-rich curricula and explicit instruction credited; skills-based reform blamed) came from nine speakers, clustered in England, Scotland, Wales, Australia, New Zealand and Finland. Reform churn (eight), teacher shortages (eight), digitalisation of schooling (six), the pandemic's "long tail" (six), AI chatbots (six) and declining reading habits and perseverance (nine) followed. Migration was named as the main cause by six commentators, chiefly in German-language and Nordic right-of-centre outlets, and rejected by the German national PISA lead ("the dividing line does not run between young people with and without a migration background, but between those from socially privileged and socially disadvantaged families") and the Finnish education minister. Sampling caveats were raised in every asterisked country and in England, Spain and Denmark. A Dutch reading researcher's "PISA tells us something is going on. Not what is going on" was the most careful non-OECD statement recorded.

Few speakers offered evidence beyond the OECD's descriptive tables, and several attached the results to policies that could not have affected the tested cohort, such as New Zealand's 2024 literacy reforms and Wales's new curriculum.

The candidate explanations, tested

IN BRIEF (summary added before publication; see Appendix H)

The report's checks can only rule explanations out. Pandemic closures are the best-identified single contributor, perhaps a third to half of the mathematics decline since 2018 and less of reading, but are contradicted as sole or main cause. Smartphones and social media fit more of the pattern than any other single account; PISA does not measure out-of-school use, so nothing decisive contradicts them. Direct causal estimates are mostly small; the evidence allows but does not establish a large role. Declining test effort is established as lowering scores, by an amount depending entirely on how the tested skill is defined. Native-born students in five stable, high-income systems fell 35 to 45 points over 2018-2025 (two on weaker samples), several of these systems did not fall on TIMSS, and advantaged students fell most in 2022-2025. A fitting explanation must work outside school, across social classes, on willingness or capacity to sustain effort on long, reading-heavy tasks, and more on PISA than TIMSS.

The causal section stated, for each explanation and before consulting the data, what pattern it would predict, then checked the predictions against the trend tables, the questionnaire tables and other assessment series. The pattern to be fitted is the one described above, plus one fact: the 2018-to-2022 and 2022-to-2025 changes are uncorrelated across systems (reading rho = 0.00).

Table 7. OECD-average questionnaire indicators (Volume I Tables I.B1.3.63-64, I.B1.4.30-161)

Indicator	2018	2022	2025
More than 1 hour/day on devices for leisure at school	–	34.5%	30.7%
More than 1 hour/day on devices for learning at school	–	55.8%	50.0%
More than 1 hour on social networks on a weekday	–	75.1%	80.1%
Students in schools banning cell phones	–	33.6%	49.4%
Sense-of-belonging index	0.00	−0.02	+0.09
Bullying index	0.00	−0.30	−0.07
Skipped a whole day in past two weeks	21.7%	19.8%	22.4%
Staff-shortage index	0.03	0.19	0.30
Disciplinary climate in science (2015 = 0.00)	–	–	+0.14

A caution applies throughout: PISA samples a new cohort each cycle, so a cross-sectional association cannot show causation, and a correlation between changes across 35 to 75 countries is an ecological association sensitive to a few cases. What these checks can do is falsify; an explanation whose predictions hold merely survives, which is much weaker.

Pandemic closures and learning loss. Prediction: a step in 2018 to 2022 that does not continue, larger where closures were longer, larger in mathematics than reading and among the disadvantaged, and mirrored in other assessments of the same cohorts. The timing prediction is half met: the largest single-cycle mathematics fall is 2018 to 2022 (−16.1) but 2022 to 2025 (−10.9) is not much smaller, and in reading the later fall (−16.1) exceeds the earlier (−10.9). The closure-length prediction fails on the trajectory analysis's specification (rho +0.24 across systems, the wrong sign; +0.14 points per week with GDP controlled, SE 0.13; −0.05 within the OECD), with qualifications. Jakubowski, Gajderowicz and Patrinos (*npj Science of Learning*, 2025), using the same UNESCO closure measure and PISA 2000-to-2022 data for 72 countries but netting out each country's pre-pandemic trend, find about 0.44 mathematics points lost per week of full closure, 9 to 12% of a standard deviation in the shortest-closure countries and 17 to 23% in the longest, concentrated among boys, immigrant and disadvantaged students; the difference is specification, and the cross-sectional regression here is not the last word. Hong Kong, whose three-domain 2018-to-2025 decline (−31.6) is among the largest in the set, is absent from the UNESCO closure data. And the absence of a gradient shows only that the marginal effect of additional weeks is invisible at country level with a coarse measure: the US annex reports that American schools in person all year still lost about 0.17 SD relative to pre-pandemic growth, so closure length is a poor proxy for pandemic exposure.The domain and distributional predictions fail after 2022. The cross-assessment check is the most informative: TIMSS 2023 grade 8 tested the same birth cohort as PISA 2025, and across 29 systems in both studies the TIMSS mathematics change correlates with the PISA change at rho = 0.26 (p = 0.17), and 0.00 within the OECD. Sweden's grade-8 mathematics rose 14 points on TIMSS while its PISA mathematics fell 38 (from the over-excluded 2018 baseline; from 2015 the fall is 30); Finland −5 against −39; France −4 against −38; Norway −2 against −49; Israel (−32 against −30) and the United States (−27 against −15) are consistent. Meta-analytic evidence (a pooled deficit of 0.14 SD, about 14 PISA points, larger in mathematics, unrecovered to mid-2022) and the Dutch natural experiment (Engzell, Frey and Verhagen, *PNAS*, 2021: 0.08 SD after eight weeks, up to 60% larger for children of less-educated parents) support a deficit of 10 to 15 points plausibly present in the 2018-to-2022 mathematics change. Verdict: part of the 2018-to-2022 window, perhaps a third to a half of the mathematics decline since 2018 and less of reading, which makes it the best-identified single component of the post-2018 decline; contradicted as the sole or main cause by the pre-2018 decline, the acceleration after 2022, the domain order, the concentration among advantaged students, the absence of a closure gradient on this specification, and the TIMSS divergence.

Smartphones, social media and digital distraction. Prediction: onset in the early-to-mid 2010s where adolescent smartphone ownership saturated first, reading most affected, heavier users scoring lower after controls, declines at least as large among advantaged students, hasty responding. The timing holds in outline; the country order partly, since the earliest and largest declines are in wealthy early adopters but Japan and Korea, with near-universal ownership, have the smallest declines among wealthy systems (reading −1 and −13 since 2018), as do Singapore and Chinese Taipei. The domain order, the behavioural signature and the concentration among advantaged students all hold, and the last is predicted by few other accounts. Within countries, an extra hour of leisure device use at school is associated with −9.7 science points before and −8.3 after adjusting for socio-economic profile, negative in 88 of 89 systems, an association that cannot separate distraction from selection. The predictions about change fail or are untestable: in-school leisure use fell and bans spread between 2022 and 2025 while reading fell 14 points; the change in leisure use correlates with the reading change at rho = +0.21, the wrong sign, the change in the banning share at −0.03 and the change in the social-network share at +0.26. The quantities the hypothesis is about, out-of-school hours and age at first phone, are not in the PISA tables; external series move as the hypothesis expects (WHO/HBSC problematic social-media use up from 7% in 2018 to 11% in 2022; Pew's 2024 survey finds nearly half of US teenagers online "almost constantly", against 24% a decade earlier) but are not linked to PISA scores. The causal literature on school bans gives test-score effects of 0.02 to 0.14 SD, concentrated among low achievers, the wrong end of the distribution for 2022 to 2025 and an order of magnitude short of a 25-point decline; Abrahamsson's Norwegian study also found improved girls' mental health and reduced bullying. One study finds much larger effects: Beneito and Vicente-Chirivella (*Applied Economic Analysis*, 2022), a synthetic-control and difference-in-differences analysis of the 2015 bans in Galicia and Castilla-La Mancha using PISA scores, report gains of 0.6 to 0.8 years of learning in mathematics and 0.7 to about one year in science, and reduced bullying; it is the one ban study on PISA scores and the one with an effect of the size the debate assumes, and as a single two-region study it should not carry the estimate alone. Verdict: consistent with more of the descriptive pattern than any other single account and contradicted by nothing decisive, because PISA does not measure what it is about; the direct causal estimates are mostly small; 5 to 15 reading points is an extrapolation; the evidence is compatible with a large role and does not establish one.

Declining reading for pleasure. Prediction: reading most affected, long-text and reflective items most, boys more, enjoyment declines followed by score declines in the same countries. The domain, item-type and reading-gender predictions hold. The check that matters most, the correlation between the change in enjoyment and the change in scores, cannot be done: the reading-habits questions were asked in 2009 and 2018 and the 2025 questionnaire has no reading-enjoyment index. Longitudinal cohort evidence supports the mechanism for book reading but not digital reading; the within-country difference of 30 to 40 points between frequent and non-readers and a 10-point shift in shares gives 3 to 4 points from composition alone. Verdict: consistent with reading, silent on mathematics unless PISA mathematics is treated as reading-intensive (which the OECD's 0.79 to 0.82 correlation between reading and mathematics trends supports, and which turns the explanation into a claim about the instrument); plausibly 3 to 8 of the 25 reading points directly, more if fragmented digital reading lowers the return to reading time; not distinguishable from the previous explanation.

Demographic composition. Prediction: declines concentrated where the immigrant share rose most, non-immigrant scores stable, low-immigration systems stable. The first holds weakly across all systems (−0.32) and not within the OECD (−0.17). The second fails: non-immigrant students' reading fell 13.0 points on average across 73 systems over 2018 to 2025 against 17.9 for all students, the two changes correlating at 0.97 to 0.98; on the OECD-29 aggregate non-immigrant reading fell 22.4 points against 24.6 for all students, so native-born students of native-born parents account for roughly nine-tenths of the decline. The difference is larger in Germany (all students −33.0, non-immigrant −20.4) and Israel (−34.6 against −19.4), and small in the Nordic countries and the Netherlands; Latvia (5% immigrant) and Finland (8%) are among the largest decliners. Two cautions apply: the native/all-student difference includes any larger decline among immigrant students themselves (substantial in Germany), and the background means use the sub-sample with valid data. Verdict: 1.5 to 3 points on the OECD average; roughly a third of Germany's reading decline and two-fifths of Israel's, a tenth or less in most Nordic systems. The claim that the link with asylum migration is "unmistakable" is contradicted by the native-student series wherever it has been made; the claim that composition is irrelevant is contradicted by Germany and Israel.

Curriculum, standards and pedagogy. These claims are testable only country by country, and the test must respect timing, since a reform affects only cohorts that passed through the reformed grades. England's 2010-to-2014 reforms (synthetic phonics with a check from 2012, a knowledge-based national curriculum from 2014, mathematics mastery) reached the cohort fully from about 2022; the United Kingdom's 2018-to-2025 changes (−10, −14, +7) are among the smallest in the OECD and its TIMSS grade-8 mathematics rose 10 points. Two readings fit. England's 2022 sample was estimated to be biased upward by 7 to 8 points, so its true fall from 2018 to 2022 was larger than published and larger than the OECD average, and its flat 2022-to-2025 figures conceal a recovery of about the same size: on this reading England was not protected through the pandemic but recovered afterward. Or the reforms protected England. The 2018-to-2025 changes are unaffected by the 2022 bias, and the coverage index (0.85 in 2018, 0.97 in 2022, 0.85 in 2025) marks 2022 as the anomalous sample. The within-UK contrast is the strongest natural experiment in the set: Scotland (Curriculum for Excellence from 2010) and Wales (whose new curriculum reaches PISA only after 2025) declined more than England over 2012 to 2022, and Scotland's 2022-to-2025 change is flat; but it is confounded by sampling (Scotland's 2022 respondents had higher prior attainment than non-respondents) and by differences in immigration, funding and poverty. Türkiye aligned its high-school entrance examination with PISA-style items from 2018 and rose 8 to 18 points in all domains between 2022 and 2025 on flat coverage, the clearest case of a national policy that predicts its PISA trend, which a reformer will read as policy working and a critic of test-driven teaching as students practising PISA-format items. Finland's decline began with the 2009 cohort, before the 2016 curriculum could affect any PISA cohort; Sweden, Norway, France, Estonia and Australia have no reform at the right time. The TIMSS divergence is the most awkward fact for this account: a curriculum that had stopped teaching mathematics would show on the curricular test, and in Sweden, Finland and France it did not (with the Swedish caveat). A limitation is that this check examined the reforms named in public debate, most of which do not fit the timing, and not the reform histories of the largest decliners, Latvia, Slovenia, Iceland and Denmark, beyond Norway's 2020 curriculum, which reached the 2025 cohort only for its last five years; inclusion policies were not tested. Verdict: a few points in particular systems, in either direction; not the synchronised decline across systems with opposite reform histories, on a test not yet applied to the countries where a reform-timing fit is most likely.

Digitalisation of schooling itself. Prediction: the earliest and fullest one-to-one adopters fell most, mathematics and science fell as much as reading, and the effect shows on curricular tests. The Nordic and Dutch geography fits; Japan (small decline, low use) and Germany (large decline, the lowest learning-device use in the OECD in 2022) do not. The 2025 share with more than an hour of learning-device use correlates with the reading change at rho = 0.00 across systems and −0.07 within the OECD; the domain prediction fails; the most digitalised systems did not fall on TIMSS; and a difference-in-differences study of 57,000 Swedish students in one-to-one schools found no average effect on national tests. Verdict: device programmes did not raise achievement and may have widened gaps; nothing supports their lowering averages by tens of points; zero to a few points in the most digitalised systems.

Teacher supply. The staff-shortage index rose and its change correlates with score changes at −0.25 to −0.27 across all systems, the predicted direction, but at −0.08 and 0.00 within the OECD; the student-teacher ratio fell from 13.3 to 12.7, against the prediction; certification was flat; and the largest decliners include the systems with the highest teacher pay and qualification standards. Verdict: 2 to 5 points; not distinguishable from a common cause.

Accountability and testing regimes. Two opposite versions exist. The strong-accountability version is contradicted by England, Singapore and Korea (small declines); the weak-accountability version by Sweden and Denmark (expanded testing, large declines). Governance features are uncorrelated with change. Türkiye fits the first version and equally the narrower reading that students practised PISA-style items. The literature puts large regime changes at 0.05 to 0.10 SD. Verdict: a few points in either direction; the two versions cancel across the decliners.

Declining effort on a low-stakes test. Prediction: effort fell most where scores fell most, hasty responses rose, declines are larger in later sections and on sustained items, and larger on PISA than on shorter curricular tests. Most predictions hold. Effort fell in 66 of 71 systems; the 2018-to-2025 effort change correlates with the reading change at rho = 0.32 across 69 systems ($p = 0.008$) and 0.46 within the OECD ($p = 0.004$), with mathematics at 0.35 and 0.52. A regression of the reading change on the effort change gives 17 points per effort point (SE 7.3, $R^2 = 0.07$), which at the OECD-35 effort change of −0.52 attributes about 9 of the 24.6 points to effort; a slope explaining 7% of the cross-country variance is being used to attribute more than a third of the average decline, so this is an upper bound, inflated by any common cause. The cross-sectional literature cuts the other way at country level: a 2026 multilevel analysis of PISA 2022 across 36 OECD countries found self-reported effort associated with achievement within students and schools ($\beta = 0.18$) but not between countries ($\beta = -0.01$, $p = 0.93$), and Michaelides and Ivanova found rapid guessing in PISA 2015 positively correlated with country means. The cross-test prediction holds strongly for the Nordic and French cases (TIMSS flat, PISA down 38 to 49) and not for Israel, the United States, Hungary, Chile or Hong Kong, where both fell. The OECD reports that “a group of students from relatively affluent families, but who struggle academically, are contributing most to the overall decline”, which fits disengaged students with nothing at stake. Verdict: established that test-taking behaviour changed in a way that lowers scores, most where scores fell most, in every section and domain; contested whether this is measurement or construct; a few points to a third of the reading decline on the validity review’s bound, about 9 points on the self-report regression, more in the Nordic and French cases if the TIMSS divergence is read as a PISA-specific component. The explanation does not say why students stopped trying, and its size depends entirely on a definitional choice.

PISA’s own instrument, mode and sampling. An artefact of a design change should appear as a step in the cycle it was introduced, in the domain affected, similar across countries. The reading decline is progressive across four cycles with different designs and runs from −52 (Iceland) to +26 (the United Arab Emirates); the flagged systems’ biases are upward, so their declines are understated; the Annex A1 check places the reading decline in the non-adaptive section on unchanged items. Where the instrument story gains ground is the other-assessment check: PISA mathematics fell where TIMSS grade-8 mathematics did not, and the OECD attributes the high correlation of reading and mathematics trends to the reading load of PISA mathematics items; that is not an error but a statement that PISA mathematics measures something that includes reading and persistence, and that this fell. Verdict: 1 to 2 points on the average; 7 to 15 for flagged systems and baselines; unknown for science 2025, where the domain ordering would change if the science link is off by 5 points either way; contradicted as the main cause; open for science until the Technical Report appears.

Spending and austerity. Cumulative spending per student correlates with the 2018-to-2025 change at −0.28 and spending per lower-secondary student at −0.45: the highest spenders fell most. The United Kingdom, Italy and Spain, whose real spending per student fell through the 2010s, are among the smallest decliners; Finland, which cut after 2015, and Norway and Denmark, which did not, are among the largest. But the prediction that matters, that systems which cut spending in the 2010s fell more, was not tested: the causal review did not construct a change-in-spending variable, so what was tested is level against change, largely the same fact as “higher-scoring systems fell more”. The causal literature (0.032 SD per $1,000 per pupil over four years) implies 3 to 5 points for a sustained 10% cut. The feature set contains no measure of private-school share or grade inflation, though Sweden’s school market is a long-standing target of public criticism; and UNICEF’s Report Card 18 (2023) shows child income poverty rising between 2012-2014 and 2019-2021 by 20% in the United Kingdom and by 10% or more in France, Iceland, Norway and Switzerland, while falling by 38% in Poland and by more than 30% in Slovenia, Latvia and Lithuania, so the poverty trend runs the predicted way for some large decliners and the opposite way for others. Verdict: not supported by the level data, which run the wrong way; the change prediction is untested; a few points where real cuts occurred, in systems that are not the ones that fell.

Absenteeism. Self-reported whole-day absence in the previous two weeks was essentially unchanged on the OECD average (21.7%, 19.8%, 22.4%); the compositional bound is 0.2 points. It rose most in the United States (+16 points to 40%) and Australia (+11 to 44%), and fell in Latvia, the largest 2022-to-2025 decliner. Administrative data show large rises in England (persistent absence from 10.9% to 22.3%) and the United States, and the causal effect is well estimated (about 0.05 SD per ten days). Verdict: 5 to 15 points in the English-speaking systems with documented surges, which are among the smaller decliners; about zero elsewhere; not consistent with the overall pattern.

Mental health and well-being. The PISA well-being indicators did not worsen: the belonging index rose from 0.00 (2018) to +0.09 (2025), loneliness fell, bullying remains below 2018. The bullying change correlates with the reading change at −0.43 ($p < 0.001$), in the predicted direction, but bullying is as plausibly a marker of school-climate disruption as a cause; the gender prediction fails for reading. Verdict: contradicted by PISA’s own well-being indicators; a few points at most; not distinguishable from digital distraction, with which it shares a timing and a proposed cause.

Reversal of the Flynn effect. Prediction: a gradual, roughly linear decline across birth cohorts from the mid-1970s, uniform across domains. PISA was flat in reading for cohorts born 1985 to 1997 and in mathematics for those born 1988 to 2003; the Norwegian conscript series on which the reversal claim rests fell about 0.2 IQ points per birth year from the 1975 cohort, which would predict roughly 20 PISA points of decline between the 2000 and 2015 cohorts, against an observed Norwegian change of +8. Verdict: contradicted on timing; a few points per decade of background trend at most, though the finding that the Norwegian decline was environmental and operated within families fits a period effect on test-taking rather than a change in schooling.

AI chatbots, classroom discipline, family support. Generative AI reached students in late 2022 and can bear only on the last window; across systems the 2025 AI-use index correlates positively with the 2022-to-2025 reading change (rho = 0.24), so systems whose students use AI more did not fall more, and the within-country association between AI summarising and lower science scores has the same selection problem as device use. The disciplinary-climate index improved and its change is uncorrelated with score changes, while TALIS reports more time keeping order. The family-support index fell 0.18 over 2022 to 2025 and its change is uncorrelated with score changes.

Table 8. Explanations, most informative check, plausible size and verdict

Explanation	Most informative check	Plausible size (reading −25 since 2018; mathematics −22 since 2015)	Verdict
Pandemic	Closure weeks vs change (≈0 with GDP here; ≈0.44/week in a trend-adjusted published study); 2018-22 vs 2022-25 changes uncorrelated	A third to a half of mathematics since 2018; less of reading; none pre-2018	Part of 2018-22 and the best-identified single component; contradicted as sole or main cause
Smartphones / social media	Leisure-use change vs score change (rho +0.2, wrong sign); within-country −8/hour after ESCS in 88/89 systems	Ban studies 0.02-0.14 SD among low achievers (one two-region PISA study finds 0.6-1 year); 5-15 reading points is an extrapolation	Consistent; compatible with a large role and does not establish one; not testable with PISA's measures
Reading habits	None available (enjoyment asked 2009/2018 only)	3-8 reading points as composition; more if kind of reading matters	Consistent with reading; silent on mathematics; not distinguishable from smartphones
Composition	Non-immigrant vs all-student change (r = 0.97)	1.5-3 points on average; a third of Germany's, two-fifths of Israel's reading decline	Some countries; contradicted as general explanation
Curriculum / pedagogy	Reform timing vs cohort; within-UK contrast (confounded); TIMSS divergence	A few points in particular systems, either direction	Some countries; not the synchronised decline; largest decliners not yet examined
Digitalised schooling	Learning-device level vs change (rho 0 to −0.3, n.s. within OECD)	0 to a few points in the most digitalised systems	Some countries; contradicted as a large effect
Teacher supply	Staff-shortage change vs change (−0.25 all; ≈0 OECD)	2-5 points	Part; not distinguishable from common cause
Accountability	Decision-locus features (rho ≈ 0)	A few points, either direction	Some countries; the two versions cancel
Test effort	Effort change vs score change (rho 0.32-0.52; R^2 0.07); TIMSS g8 vs PISA (0.26; 0.00 in OECD); country-level cross-sectional null in 2022	A few points to a third of reading (self-report bound ≈9); more in Nordic/French cases	Consistent; size depends on construct vs contamination
Instrument, sampling	Link errors; Annex A1 percent-correct (r 0.90-0.91)	1-2 points on average; 7-15 for flagged systems; science 2025 unknown	Contradicted as main cause; open for science 2025 and individual countries
Spending	Spending level vs change (rho −0.28 to −0.45, wrong sign); change in spending not tested	3-5 points where real cuts occurred	Level runs the wrong way; change untested; not the systems that fell
Absenteeism	Compositional bound 0.2 points; change vs change ≈ 0 within OECD	5-15 points in English-speaking systems; ≈0 elsewhere	Some countries; contradicted as general explanation
Mental health	Belonging change vs change (rho ≈ 0)	A few points at most	Contradicted by PISA well-being indicators; not distinguishable from smartphones on timing
Flynn reversal	OECD-23 series 2000-12 flat	A few points per decade of background trend	Contradicted on timing

Source: causal review, section 16.

Three facts constrain any combination. Native-born students in high-income systems with stable schooling, short or no closures, high spending and no major reform (Finland, Norway, Denmark, Sweden, the Netherlands) fell 35 to 45 points in reading and mathematics between 2018 and 2025 (Sweden from an over-excluded baseline, the Netherlands on asterisked samples); the same cohort in several of these systems did not fall on TIMSS grade-8 mathematics; and students in the most advantaged quarter fell most in the most recent window, while belonging, discipline, absence at 15 and in-school device use moved in the benign direction. An explanation that fits all three has to operate outside school, across social classes, on the willingness or capacity to sustain effort on long, reading-heavy tasks, and more on PISA than on TIMSS.

What remains open, and what would settle it

Several questions cannot be answered from the published volume and are stated here with the evidence that would resolve them.

The science trend. Whether the 2025 science scale is comparable with 2015 across the framework revision, the adaptive design, the platform change and the move of scaling from ETS to ACER is unaddressed in Volume I; the Technical Report should document how many trend items span the framework change, how the new competency was linked, whether the platform affected scores, and whether the 2015-to-2022 scaling model, item parameters and partial-invariance decisions were carried over. Until then "science held up" is provisional, and the domain ordering that several explanations are asked to fit depends on it.

The flagged samples. The volume describes non-response bias analyses for the asterisked systems but does not tabulate them, and publishes no 2025 exclusion or response rates for the others. Linkage to administrative achievement records, as England and New Zealand have done, would settle the direction of the US bias, and a table of bias estimates for 2022, 2025 and Sweden's 2018 baseline would show whether flagged changes are inside their uncertainty.

Effort versus skill. An incentive experiment embedded in a PISA cycle, or a linked administration of PISA and a curricular test to the same students, would measure the PISA-specific component of the decline rather than bound it; the TIMSS comparison used here is across separate samples with their own sampling issues, which the causal section did not verify.

The out-of-school mechanisms. PISA does not measure out-of-school smartphone hours, age at first phone or reading volume at home, and the reading-enjoyment index was dropped; reinstating the 2009 and 2018 reading-habits questions in 2029 would allow the one test the reading-habits explanation most needs. The post-2024 phone bans will first be observable in PISA 2029.

The early-childhood signal. TIMSS grade-4 changes over 2019 to 2023 correlate with the PISA 2022-to-2025 change at 0.42 to 0.46, and PIRLS 2016-to-2021 changes at 0.34, indicating that something happened to primary-age children that is visible three years later; identifying it would need longitudinal cohorts with measures of reading, screen exposure and schooling.

How to read the results

The safest statement the evidence supports is the one the OECD's Annex makes and its Preface does not: the decline in PISA reading since 2018 is present on every section and item type of a test that has not changed, is larger than any measurement issue reviewed here can account for, and is accompanied by measurable changes in how students behave during the test. It predates the pandemic at country level, is concentrated in rich systems that previously scored well, has in the latest window affected advantaged students more than disadvantaged ones, and is not mirrored on the curriculum-based mathematics test given to the same cohort in several of the countries where it is largest. Mathematics has fallen almost as far since 2015; science less, on the least secure of the three comparisons.

The explanations that survive the checks are not rivals in the ordinary sense. Digital distraction and declining reading habits are proposed reasons why students engage less with long texts; declining effort on the test is the form that reduced engagement takes when PISA measures it; and the OECD's composite is these three seen together. The effort explanation has the strongest correlations within the dataset because it is the only one PISA measures directly, though they explain little of the cross-country variance, and its size, from a few points to most of the decline, depends on a decision the data cannot make: whether the willingness and capacity to sustain attention on a long, low-stakes, reading-heavy task is part of what "reading" means, as the OECD's framework holds, or a nuisance variable that contaminates the measurement of a skill. On the first reading PISA has detected a real change in a competence that matters, as large as the tables say; on the second PISA has become a test of something 15-year-olds in rich countries no longer do for free, and the underlying skill has fallen less. These are one hypothesis seen from the student's side and from the instrument's side, and the assembled data cannot separate them.

A reader is therefore left with a decision rather than a discovery, and with constraints on it: no explanation acting inside schools fits the cross-country pattern; the pandemic is the best-identified single component and explains part of one window, not the trend; composition and the cognitive-trend literature are contradicted as general accounts while contributing something somewhere; spending is untested on the quantity that matters; and of the explanations most voiced after the release, screens have evidence compatible with a large role that does not establish one, while funding and curriculum have direct evidence that is small, untested or country-specific. None of this selects an explanation. It does rule out the confident ones.

Some practical guidance follows. Three-year changes for individual countries should be read against standard errors that include the link error and, for flagged systems and baselines, against sampling uncertainty of the same order as the change; rises coinciding with falling coverage should be discounted; science trends should be read against a link error three times reading's and an undocumented framework revision; reforms or bans enacted after 2022 cannot be checked against these results; and the OECD's "20 points equals one year" conversion, re-estimated in its own report at about 10 points (Serbia) to 25 to 28 (the United Kingdom or Scotland, on two readings of Table I.2.7), should not be used to translate any number above into years lost.

How this report was produced

This report was written by AI agents: separate instances of Claude Fable 5, a large language model made by Anthropic, each given a written brief and the materials the brief lists. The software the agents ran in also supplied basic account details (a name, an email address and a short description of Bill McCallum's newsletter project), enough in principle to identify him; the briefs did not ask the agents to use them. The work was commissioned by Bill McCallum, who reviewed the briefs before the agents ran. The briefs do not say who commissioned the work or what that person thinks about the questions involved; they are published alongside this report. The numerical checks and source checks described below were carried out by the agents. Before publication on 10 September 2026, further agents checked the quotations of named people and organizations against their sources; about forty were corrected, paraphrased or removed, and the check is described in the supporting sections.

The sections synthesised in this report were written independently of one another from the OECD's published data files, the text of *PISA 2025 Results (Volume I)*, national assessment agencies' reports, the peer-reviewed literature and, for the inventory of public claims, press and official sources published up to 9 September 2026; each stated the source of every figure and carried the OECD's standard errors and significance flags. The section on causes used a predict-then-check structure in which each explanation's predictions were written down before the data were examined, so that the checks could falsify rather than merely illustrate. A draft of this synthesis was then subjected to an adversarial review against the source sections, the data files and the cited literature, which found numerical errors in the United States material, an unlabelled correlation, a summary that characterised the pandemic and spending verdicts more harshly and the smartphone verdict more vaguely than the section beneath it, and omissions that ran mostly one way (Sweden's 2018 baseline, the change of scaling contractor, the qualifying statistics on effort, the untested spending prediction, the within-UK contrast, the governance critique). This version corrects and restores these and records the changes in an accompanying changelog. The refusal to select an explanation is unchanged.

Annex: The United States

Scope and sources

This annex sets out the United States' PISA record from 2000 to 2025 with confidence intervals, position relative to the OECD average and the sampling adjudication for every cycle; places it alongside the NAEP long-term trend (LTT) at ages 13 and 17, the main NAEP at grades 8 and 12 including 2024, and TIMSS grade 8 through 2023; checks each explanation in the causal review (`E/causal_review.md`) that has been applied to the United States against national and state evidence; and reports what the US student questionnaire shows on devices, absence and reading enjoyment.

PISA figures are from the OECD trend tables (PISA 2025 Results Volume I, Tables I.B1.2a.36-44, https://stat.link/mrq53f), the Reader's Guide (pp. 17-19), the OECD's US country note (https://www.oecd.org/en/publications/pisa-2025-results-volume-i-country-notes_2d4ff9ea-en/united-states_0c8cbc7c-en.html) and the NCES web report *Highlights of U.S. PISA 2025 Results* (NCES 2026-032, https://nces.ed.gov/surveys/pisa/pisa2025/) with its technical notes (https://nces.ed.gov/surveys/pisa/pisa2025/technical-notes/index.asp). NAEP figures were downloaded from the NAEP Data Service API of the NAEP Data Explorer (https://www.nationsreportcard.gov/DataService/GetAdhocData.aspx; documentation at https://www.nationsreportcard.gov/api_documentation.aspx) with `QCData=true` for standard errors (`fetch_naep.py` ; raw JSON in `raw/naep/`); every value used is in `us_series.csv` with its query URL. TIMSS values are from the TIMSS 2023 trend exhibits as parsed in `E/computed/timss_trends.csv` . State checks are in `state_checks.py` and `computed/state_checks_log.txt` ; the change estimates quoted are in `computed/annex_numbers.txt` (`build_series.py`). "Significant" means |estimate/SE| > 1.96; PISA change SEs include the OECD's link error; NAEP change SEs are sqrt($SE1^2 + SE2^2$), which for differences of about a point can differ from NCES's published calls. PISA scales have SD 100; NAEP scales have SDs of about 35-42 (grade 8 reading 37.8, grade 8 mathematics 39.7, grade 12 reading 42.3, grade 12 mathematics 35.6, grade 8 science 34.8, LTT age 13 mathematics 35.3 and reading 37.2), so 4 NAEP points is roughly 0.1 SD, about 10 PISA points.

1. The US PISA record, 2000-2025

> **IN BRIEF** (summary added before publication; see Appendix H)
>
> The US has never been significantly above the OECD average in mathematics; in 2025 it is 2.6 points below the OECD-35 average, not significant, because the average fell faster, so its better relative position reflects declines elsewhere. The 2025 reading and mathematics means are the lowest in the US series, though reading is not significantly different from 2000, 2009, 2012 or 2015; the declines since 2018 of 15.8 and 15.4 points are significant, while science is flat throughout. Compared with 2009 and 2015, the 10th percentile in reading and science is lower and the 90th higher, and the reading inter-decile range of 312 points is the largest of the 91 systems. The US has missed the 85% school response standard in every cycle. In 2025 school response was 45%, exclusions 6.9% and results carry an asterisk; the OECD's range of possible bias, about 15 points either way, is about the size of the 2018-2025 changes, with direction unknown.

1.1 Means, confidence intervals and position relative to the OECD average

Table 1 gives the US mean, SE and 95% confidence interval in each domain and cycle and the difference from the OECD average. As in the OECD's own text, the aggregate is "OECD average-23" for 2000-2012 and "OECD average-35" from 2015; the two differ by 6-8 points in level, so the difference column is not a single series across 2012-2015. The SE of the difference ignores the small covariance between the US and an average that includes it.

Table 1. United States PISA means, 95% confidence intervals and difference from the OECD average

Cycle	Reading: mean (SE), 95% CI	US − OECD (SE)	Mathematics: mean (SE), 95% CI	US − OECD (SE)	Science: mean (SE), 95% CI	US − OECD (SE)
2000	504.4 (7.0), 490.6-518.2	+4.9 (7.1) n.s.	scale established 2003		scale established 2006	
2003	495.2 (3.2), 488.9-501.5	−2.2 (3.3) n.s.	482.9 (2.9), 477.1-488.7	−19.2 (3.0) sig		
2006	withheld (printing error)		474.4 (4.0), 466.5-482.2	−26.7 (4.1) sig	488.9 (4.2), 480.6-497.2	−13.8 (4.3) sig
2009	499.8 (3.7), 492.7-507.0	+0.4 (3.7) n.s.	487.4 (3.6), 480.4-494.4	−14.3 (3.6) sig	502.0 (3.6), 494.9-509.1	−3.7 (3.7) n.s.
2012	497.6 (3.7), 490.2-504.9	−3.8 (3.8) n.s.	481.4 (3.6), 474.3-488.4	−17.3 (3.6) sig	497.4 (3.8), 490.0-504.8	−7.3 (3.8) n.s.
2015	496.9 (3.4), 490.3-503.6	+6.7 (3.4) n.s.	469.6 (3.2), 463.4-475.8	−17.6 (3.2) sig	496.2 (3.2), 490.0-502.5	+5.5 (3.2) n.s.
2018	505.4 (3.6), 498.4-512.3	+17.7 (3.6) sig	478.2 (3.2), 471.9-484.6	−11.4 (3.3) sig	502.4 (3.3), 495.9-508.9	+13.2 (3.3) sig
2022	503.9 (4.3), 495.5-512.4	+26.6 (4.4) sig	464.9 (4.0), 457.0-472.7	−10.0 (4.0) sig	499.4 (4.3), 490.9-507.9	+12.7 (4.3) sig
2025	489.5 (5.8), 478.1-501.0	+26.5 (5.9) sig	462.8 (5.4), 452.3-473.4	−2.6 (5.4) n.s.	501.6 (5.8), 490.2-513.0	+17.7 (5.8) sig

Sources: Tables I.B1.2a.36-38. The 2006 US reading result was withheld because students “were incorrectly instructed to refer to the passage on the ‘opposite page’”; NCES estimated the effect on mathematics and science at about one point (NCES 2008-016, p. iii, https://nces.ed.gov/pubs2008/2008016.pdf). The 2025 SEs (5.4-5.8) are the largest since 2000, from a sample of 4,625 students in 144 schools.

Three features stand out. The United States has never been significantly above the OECD average in mathematics and was significantly below it in every cycle from 2003 to 2022; in 2025 it is 2.6 points below the OECD-35 average, not significant, because the average fell faster. The country note’s “close to the OECD average in mathematics, and higher than the OECD average in reading and science” and the NCES’s equivalent statement both check out against the 38-member average (461, 463, 482). Second, the US position relative to the average has improved by 20-30 points in every domain since 2012 while the US level has not: reading went from 3.8 below to 26.5 above with the US score down 8 points, science from 7.3 below to 17.7 above with the US score up 4. The relative improvement is a statement about other countries. Third, the 2025 reading (489.5) and mathematics (462.8) means are the lowest in the US series, which is the basis of the country note’s “among the lowest ever observed in PISA in the United States” and of the NCES’s statement that 2025 reading was “lower than its scores in the two most recent assessment cycles”.

1.2 Changes over time, and where in the distribution they occur

Table 2. United States: change to PISA 2025 from earlier cycles (SE includes link error) and changes at the tails

Domain	Base	Mean change (SE)	95% CI	p10 change (SE)	p90 change (SE)	Below L2, pp (SE)	L5+, pp (SE)
Reading	2000	−14.9 (11.4) n.s.	−37.2 to +7.4	−31.2 (14.3)	+8.0 (12.7)	+7.8 (3.4)	+1.1 (2.5)
Reading	2009	−10.3 (8.4) n.s.	−26.7 to +6.1	−40.4 (8.2)	+18.9 (11.1)	+8.1 (2.4)	+3.5 (2.1)
Reading	2012	−8.1 (9.2) n.s.	−26.2 to +10.1	−46.5 (9.4)	+29.9 (11.3)	+9.1 (2.8)	+5.4 (2.1)
Reading	2015	−7.4 (7.7) n.s.	−22.6 to +7.8	−32.2 (8.5)	+20.2 (10.1)	+6.7 (2.3)	+3.7 (1.9)
Reading	2018	−15.8 (7.1) sig	−29.7 to −2.0	−28.7 (7.7)	+1.1 (9.6)	+6.4 (2.0)	−0.2 (1.8)
Reading	2022	−14.4 (7.3) sig	−28.8 to −0.0	−23.9 (8.2)	−4.1 (10.3)	+5.6 (2.0)	−0.9 (1.9)
Mathematics	2003	−20.0 (8.4) sig	−36.4 to −3.6	−22.1 (8.8)	−12.9 (11.7)	+9.5 (3.0)	−2.1 (1.9)
Mathematics	2009	−24.5 (7.9) sig	−39.9 to −9.2	−34.1 (8.0)	−12.1 (11.4)	+11.8 (2.8)	−1.9 (1.9)
Mathematics	2012	−18.5 (7.5) sig	−33.2 to −3.8	−33.2 (7.4)	−5.9 (11.0)	+9.3 (2.7)	−0.8 (1.8)
Mathematics	2015	−6.8 (6.9) n.s.	−20.4 to +6.8	−20.5 (7.0)	+10.0 (10.8)	+5.8 (2.6)	+2.1 (1.7)
Mathematics	2018	−15.4 (6.8) sig	−28.7 to −2.1	−22.5 (7.2)	−3.4 (10.7)	+8.1 (2.5)	−0.3 (1.7)
Mathematics	2022	−2.0 (6.8) n.s.	−15.4 to +11.3	−10.1 (6.5)	+4.4 (11.2)	+1.3 (2.6)	+0.6 (1.8)
Science	2006	+12.7 (8.7) n.s.	−4.3 to +29.6	+2.0 (9.6)	+21.2 (11.5)	−2.1 (2.6)	+3.8 (2.1)
Science	2009	−0.4 (9.6) n.s.	−19.2 to +18.4	−23.0 (9.9)	+20.0 (12.6)	+4.2 (2.7)	+3.8 (2.4)
Science	2015	+5.3 (7.5) n.s.	−9.3 to +20.0	−16.2 (7.8)	+23.8 (10.8)	+2.0 (2.1)	+4.4 (2.0)
Science	2018	−0.8 (7.6) n.s.	−15.6 to +14.0	−19.1 (8.4)	+20.2 (10.8)	+3.7 (2.1)	+3.8 (2.0)
Science	2022	+2.2 (7.9) n.s.	−13.3 to +17.6	−5.1 (8.3)	+10.6 (11.2)	+0.4 (2.2)	+1.9 (2.1)

Sources: Tables I.B1.2a.33-41. The 2018→2022 changes in the PISA 2022 tables (I.B1.5.4-6, https://stat.link/wh9d4z) were mathematics −13.4 (SE 5.6, significant), reading −1.4 (5.8) and science −3.0 (5.7), not significant.

The tails are more informative than the means. In reading the mean is not significantly different from 2000, 2009, 2012 or 2015, but the 10th percentile has fallen 31-47 points from every base before 2018 while the 90th has risen 19-30 points since 2009-2015; the NCES puts this as “the reading score of higher-performing (90th percentile) students in 2025 was not measurably different or was higher than in all earlier cycles while the scores of lower-performing (10th and 25th percentile) students in 2025 were lower than in all earlier cycles.” The US 90th-10th range in reading, 312 points, is the largest of the 91 systems (OECD average 264), the fact behind the NCES commissioner’s and NYU statements in the claims inventory (B10.2). Science shows the same shape: the mean is flat over every window, the 10th percentile down 16-26 points since 2009-2018, the 90th up 20-30, and the range (298) second only to Luxembourg (Volume I, p. 73). In mathematics the whole distribution has moved down since 2003-2012, the bottom (p10 −22 to −34) more than the top (p90 −6 to −13, none significant), and the share below Level 2 has risen from 25.7% in 2003 to 35.2%. Over 2015-2025 the Level 5+ share in science rose 4.4 points to 12.9%, which is why Volume I lists the United States among systems that raised their top share (p. 115); the country note’s “six percentage points in mathematics; seven percentage points in reading; did not change significantly in science” for low performers since 2015 corresponds to +5.8, +6.7 and +2.0 (n.s.). The OECD’s fitted 2015-2025 decennial trends for the US are −6.9 (reading; SE 7.0, p = 0.33), −10.5 (mathematics; SE 6.5, p = 0.11) and +3.6 (science; SE 6.8), none significant, against OECD-35 trends of −27.0, −23.9 and −6.7. Over 2018-2025 the US declines of 15.8 (reading) and 15.4 (mathematics) are significant and about two-thirds of the OECD-35 declines (24.6, 24.2); science (−0.8) is not significant. Converting these to time is hazardous, and the OECD says so: Box I.2.2

warns against “using a single estimate to convert any difference in terms of years-of-schooling equivalents”, and its US-specific estimates of the yearly learning gain around age 15 in Table I.2.7 are 13.8 points in science (SE 4.4), 12.0 in reading (SE 4.6) and 13.5 in mathematics (SE 3.6), among the lowest of the ten systems estimated (England 24.5, 24.0 and 27.0; Serbia 13.7, 11.4 and 7.4). On the OECD’s own US figure the 14-point 2022-2025 reading fall corresponds to more than a year of learning, not the “seven months” in Secretary McMahon’s statement (claims inventory B1.1), which uses the generic 20-points-a-year rule; on the SEs of Table I.2.7 the US estimate is compatible with anything from about eight months to several years.

1.3 Sampling: the US record and the adjudication outcome, cycle by cycle

PISA’s standards require overall exclusions of at most 5%, a weighted school response rate of at least 85% before replacement (or an acceptable rate after replacement; the OECD’s floor has historically been 65% before replacement), and student response of at least 80% (Reader’s Guide, p. 17). The United States has missed the 85% school standard in all nine cycles.

Table 3. United States PISA sampling record and adjudication outcome, 2000-2025

Cycle	School response before / after replacement (weighted %)	Student response (%)	Exclusions (%)	Adjudication outcome and notes
2000	56 / 70	85	not reported	Below the 65% floor; US, UK and Netherlands submitted bias analyses; the Technical Advisory Group judged remaining bias “likely to be minimal” and included the US and UK but not the Netherlands (NCES 2002-115, pp. 66-67, https://nces.ed.gov/pubs2002/2002115.pdf).
2003	65 / 68	83	7	“Intermediate” category; a second testing window was opened from September through November 2003 “to improve response rates (a response rate of approximately 50 percent was projected for the end of the data collection period) and better accommodate school schedules”; two bias analyses; reported without annotation (NCES 2005-003, pp. 3-4 and 47-49, https://nces.ed.gov/pubs2005/2005003.pdf).
2006	69 / 79	91	not reported	Intermediate; reading withheld (printing error); NCES notes the international formula “is not consistent with NCES standards” and a stricter one gives 67.5% (NCES 2008-016, pp. 4, 29-30).
2009	68 / 78	87	5	Participating schools had a slightly higher free-lunch share (35.4% vs 34.1%); reported without annotation (NCES 2011-004, pp. 45-46, https://nces.ed.gov/pubs2011/2011004.pdf).
2012	67 / 77	89	5	Bias analysis required; reported without annotation (NCES 2014-024rev, pp. 28-29, https://nces.ed.gov/pubs2014/2014024rev.pdf).
2015	67 / 83	90	3	Reported without annotation (NCES 2017-048, p. 29, https://nces.ed.gov/pubs2017/2017048.pdf).
2018	65 / 76	85	3.8	The Adjudication Group “did not consider the violation … as major adjudication issues” but “noted several limitations in the data used in non-response-bias analyses”; reported with annotation, and an “OECD average-35b” excluding Spain and the United States was used for some comparisons (PISA 2018 Vol. I, Annex A4; Tables I.A2.1, I.A2.6, https://doi.org/10.1787/888934028862).
2022	51 / 63	80	6.1	Asterisked, in the group where “it is not possible to exclude the possibility of more than minimal bias”; participation “particularly low among private schools (about 7% of the student population)”; exclusions “showed a marked increase, with respect to 2018 (3.8%), in exclusion rates for students with functional or intellectual disabilities”; “not possible to exclude the possibility of bias, nor to determine its most likely direction” (PISA 2022 Vol. I, Reader’s Guide, pp. 17-20). NCES found 10th-graders (73.7% vs 73.1%) and students without special needs (95.4% vs 95.0%) slightly over-represented (https://nces.ed.gov/surveys/pisa/pisa2022/technical-notes/index.asp).
2025	45 / 54	76	6.9 (5.4 excluding limited-English students)	“Limited reporting with asterisk” (with Albania). Range-of-bias simulations: school non-participation “up to about ±15.4% of a standard deviation”, student non-response “up to about ±15.9%”; the bias analysis used observed characteristics “though not direct measures of school performance, which were not available”; “it is not possible to rule out bias in PISA performance results, nor to determine its most likely direction.” Over 40% of respondents were not given the questionnaire (nine states declined; some schools tested during a national suspension), so questionnaire results are “below the line” (Reader’s Guide, pp. 18-19). NCES: city and town schools under-represented, participating schools smaller (mean enrollment 1,323 vs 1,399); at student level only mean age differed, by 0.07 months (NCES technical notes).

Two things matter for the trend. Response rates have fallen steadily, from 65-69% of schools and 83-91% of students in 2003-2018 to 45% and 76% in 2025, and the direction of any bias is unknown: where linked records exist (New Zealand, England) non-respondents were lower achieving and the bias upward (validity review), but no achievement data were available for the US analyses in 2022 or 2025. The 2025 bounds of about ±15 points are the same order as the 2018-2025 changes in reading (−15.8) and mathematics (−15.4) and larger than the 2022-2025 changes; Volume I reports US trends without adjustment and drops the 2022 asterisk in the 2025 trend tables (technical note). Exclusions (6.9%) exceed 5% for the second cycle; the 2022 Adjudication Group had asked the national centre to investigate and expected rates “will fall again”, which did not happen. Since excluded students are mostly low achievers, this component works against the reported declines. One further change is specific to 2025: the United States had tested in the autumn in every cycle since 2003, with permission, “to avoid the large amount of state and local testing occurring in the spring”; for 2025 the PISA Governing Board limited data collection to February-September, and the US assessed from 9 April to 2 June 2025 (NCES technical notes). Birth-date ranges are adjusted so that the age range is unchanged, and NCES’s 2003 spring-autumn comparison found no significant difference, but the grade composition of the age cohort shifts, which is why Box I.2.2 lists the United States among ten systems whose testing-date changes it uses to estimate yearly learning gains; the move into the state-testing season is also a plausible contributor to the fall in school participation from 51% to 45%.

2. PISA alongside NAEP and TIMSS

IN BRIEF (summary added before publication; see Appendix H)

PISA agrees with the other US series (NAEP, the US national assessment, and TIMSS) that scores flattened or turned down from 2013-2015, before the pandemic, and that each series shows its largest drop between successive assessments in the pandemic period. Unlike PISA, the other series show US middle-school mathematics rising through 2013. On every NAEP series the 10th percentile has fallen far more than the 90th, so the widening US inter-decile range is not produced by PISA itself. Where the series diverge, PISA is the outlier. In standard-deviation units its mathematics fall since 2018 (0.15) is smaller than on NAEP grade 8 (0.21) or TIMSS (0.27), and its reading fall is concentrated in 2022-2025 (14.4 points), where NAEP spreads it over 2013-2024. Given the standard errors and the notes on sampling, neither divergence supports a claim that PISA and NAEP show a different kind of US trend, and every US series is lower in 2023-2025 than in 2019.

2.1 What each series measures and whom it samples

PISA samples 15-year-olds (15 years 3 months to 16 years 2 months) in grade 7 or above, public and private; in the US about 74% are in grade 10, 19% in grade 11 and 7% in grade 9 (PISA 2018 Table I.A2.8), coverage was 0.85 of the 15-year-old population in 2025 (Table I.A2.1), and the items are “literacy” tasks not tied to a curriculum. Main NAEP samples grade cohorts (4, 8, 12) in public and private schools with school participation above 85% and student participation of 90% or more at grades 4 and 8 (grade 12 participation is lower and is published with each release), on frameworks set by NAGB and tied to US curricula; LTT samples by age (9, 13, 17) with instruments essentially unchanged since the 1970s; TIMSS samples grade 8 (and 4) with a curriculum framework. None has consequences for students. Grade-12 NAEP is closest to PISA in age but has the weakest participation and the greatest risk of selective absence; grade-8 NAEP and LTT age 13 are the strongest samples. Table 4 aligns birth cohorts.

Table 4. Approximate birth-cohort alignment of the US series

PISA cycle	Birth year	NAEP grade 8	LTT age 13	TIMSS grade 8	NAEP grade 12	LTT age 17
2009	1993	2007	2008	2007	2009/2013	2012
2012	1996	2009/2011	2008/2012	2011	2013/2015	none
2015	1999	2013	2012	2015	2015/2019	none
2018	2002	2015/2017	none	2015/2019	2019	none
2022	2006	2019/2022	2020	2019	2024	none
2025	2009	2022/2024	2023	2023	2027	none

The PISA 2025 cohort was in grade 5-6 when schools closed in March 2020, sat TIMSS grade 8 and LTT age 13 in 2023, and straddles the 2022 and 2024 grade-8 NAEP samples. The PISA 2022 cohort sat grade-12 NAEP in 2024.

2.2 The series

Table 5. US national series, 2000-2025 (means, SE in parentheses)

Year	PISA reading	PISA maths	PISA science	NAEP g8 reading	NAEP g8 maths	NAEP g8 science	NAEP g12 reading	NAEP g12 maths	LTT 13 reading	LTT 13 maths	LTT 17 reading	LTT 17 maths	TIMSS g8 maths	TIMSS g8 science
1999/2000	504 (7.0)	–	–	263 (0.8) [1998]	273 (0.8)	–	290 (0.6) [1998]	–	259 (1.0)	276 (0.8)	288 (1.3)	308 (1.0)	502 (3.9)	515 (4.4)
2003/04	495 (3.2)	483 (2.9)	–	263 (0.3)	278 (0.3)	–	287 (0.7) [2002]	–	257 (1.0)	279 (1.0)	283 (1.1)	305 (0.7)	504 (3.4)	527 (3.2)
2005-08	– [2006]	474 (4.0)	489 (4.2)	263 (0.2) [2007]	281 (0.3) [2007]	–	286 (0.6) [2005]	150 (0.6) [2005]	260 (0.8) [2008]	281 (0.9) [2008]	286 (0.9) [2008]	306 (0.6) [2008]	508 (2.9)	520 (2.9)
2009	500 (3.7)	487 (3.6)	502 (3.6)	264 (0.3)	283 (0.3)	150 (0.3)	288 (0.7)	153 (0.7)	–	–	–	–	–	–
2011-13	498 (3.7) [2012]	481 (3.6)	497 (3.8)	268 (0.3) [2013]	285 (0.3) [2013]	152 (0.3) [2011]	288 (0.6) [2013]	153 (0.5) [2013]	263 (1.0) [2012]	285 (1.1) [2012]	287 (0.9) [2012]	306 (0.8) [2012]	509 (2.7) [2011]	525 (2.4) [2011]
2015	497 (3.4)	470 (3.2)	496 (3.2)	265 (0.2)	282 (0.3)	154 (0.3)	287 (0.5)	152 (0.5)	–	–	–	–	518 (3.1)	530 (2.8)
2017/18	505 (3.6)	478 (3.2)	502 (3.3)	267 (0.3) [2017]	283 (0.3) [2017]	–	–	–	–	–	–	–	–	–
2019/20	–	–	–	263 (0.3)	282 (0.3)	154 (0.5)	285 (0.5)	150 (0.4)	260 (0.7) [2020]	280 (0.8) [2020]	cancelled	cancelled	515 (4.8)	522 (4.7)

Year	PISA reading	PISA maths	PISA science	NAEP g8 reading	NAEP g8 maths	NAEP g8 science	NAEP g12 reading	NAEP g12 maths	LTT 13 reading	LTT 13 maths	LTT 17 reading	LTT 17 maths	TIMSS g8 maths	TIMSS g8 science
2022/23	504 (4.3)	465 (4.0)	499 (4.3)	260 (0.3)	274 (0.3)	–	–	–	256 (0.9) [2023]	271 (0.9) [2023]	–	–	488 (4.2)	513 (3.9)
2024	–	–	–	258 (0.4)	274 (0.3)	150 (0.5)	283 (0.5)	147 (0.5)	–	–	–	–	–	–
2025	490 (5.8)	463 (5.4)	502 (5.8)	–	–	–	–	–	–	–	–	–	–	–

Sources: PISA Tables I.B1.2a.36-38; NAEP Data Service API (`us_series.csv` carries the query URL for each cell; 2004 LTT values are the revised-format sample; pre-2002 main NAEP values are accommodations-permitted samples); TIMSS 2023 Exhibits 1.2.10 and 2.2.10. Grade-12 mathematics and science scales run 0-300; others 0-500. The 2020 LTT age-17 assessment was cancelled and not repeated.

2.3 Where the series agree

The series agree on three things. Mathematics improved at the middle-school level through the 2000s: NAEP grade 8 rose from 273 (2000) to a peak of 284.6 (2013), +7.0 points (about 0.18 SD) over 2003-2013; LTT age 13 from 276 (1999) to 285 (2012); TIMSS grade 8 from 502 (1999) to 509 (2011) and 518 (2015). PISA mathematics shows neither a rise nor a fall in that decade (483 in 2003, 481 in 2012, difference −1.5). Second, everything flattened or turned down from 2013-2015, before the pandemic: NAEP grade 8 mathematics −2.6 (SE 0.4) and reading −4.4 (SE 0.4) over 2013-2019; grade 12 reading −2.9 and mathematics −3.2 over 2013-2019; LTT age 13 reading −2.7 and mathematics −5.3 over 2012-2020 (the 2020 LTT was administered before the closures); PISA mathematics −11.7 over 2012-2015 followed by +8.6 to 2018. Loveless's 2016 comparison of groups of states points the same way: slight progress in reading and small declines in grade 8 mathematics over 2009-2015, whatever a state's Common Core status ("None of the states are setting the world on fire"). Third, all series show their largest single-interval decline in the pandemic window: NAEP grade 8 mathematics −7.7 (SE 0.4) over 2019-2022, the largest in the series' history, and reading −2.7; LTT age 13 mathematics −9.0 over 2020-2023 and reading −4.4; PISA mathematics −13.4 (SE 5.6) over 2018-2022; TIMSS grade 8 mathematics −27 (SE 6.4) over 2019-2023 and science −9 (SE 6.1, n.s.).

The tails agree even more closely than the means. On every NAEP series the 10th percentile has fallen far more than the 90th: grade 8 reading p10 −15.3 (SE 0.8) over 2009-2024 against p90 +1.5; grade 8 mathematics p10 −16.5 against p90 −0.4; grade 12 reading p10 −13.7 against −0.4; grade 12 mathematics p10 −10.1 against −0.6; LTT age 13 mathematics p10 −26.7 (SE 2.0) against p90 −6.4 over 2012-2023, reading −12.3 against −3.7. The share of grade-8 students below NAEP Basic rose from 27.4% (2019) to 32.9% (2024) in reading and is 39.4% in mathematics; at grade 12 in 2024, 31.8% were below Basic in reading and 44.8% in mathematics. This is the shape of Table 2. The widening US inter-decile range that Volume I and the NCES report is not a PISA artefact: it is present on the national assessments at ages 13, 15 and 17, and it began before the pandemic.

2.4 Where the series diverge, and what the divergences imply

Reading since 2018. PISA reading was flat over 2018-2022 (−1.4, n.s.) and then fell 14.4 points (0.14 SD) over 2022-2025. NAEP grade 8 reading fell in both windows, −2.7 over 2019-2022 and −2.4 over 2022-2024 (about 0.07 and 0.06 SD), and LTT age 13 reading fell 4.4 points (0.12 SD) over 2020-2023. The cohort PISA tested in 2025 sat LTT age 13 in 2023, so on the same cohort PISA's reading decline is of the order of NAEP's; what differs is timing, with PISA showing none of it in 2022. Sampling and administration are the obvious candidates: the 2022 US reading mean (503.9) came from a sample with a 63% school response rate whose bias direction the OECD could not determine, and if that figure was a few points high the 2022-2025 fall is correspondingly overstated; and 2025 was the first US administration in the spring since 2003 (section 1.3). PISA also misses the 2013-2019 slippage that NAEP grade 8 reading (−4.4, 0.12 SD) and grade 12 reading (−2.9) show, because US PISA reading rose 8.4 points between 2015 and 2018 on an annotated sample; with SEs of 3.2-7.0 plus link errors of 2-9 points, US PISA cannot resolve changes of the size NAEP measures with SEs of 0.2-0.4.

Mathematics since 2018. PISA is the series that fell least. PISA mathematics fell 15.4 points (0.15 SD) over 2018-2025, 13.4 of it by 2022. NAEP grade 8 mathematics fell 8.2 points (0.21 SD) over 2019-2024, LTT age 13 mathematics 14.3 points (0.40 SD) over 2012-2023 of which 9.0 (0.25 SD) came in 2020-2023, TIMSS grade 8 mathematics 27 points (0.27 SD) over 2019-2023, and grade 12 mathematics 3.3 points (150.28 to 146.95; 0.09 SD) over 2019-2024 to the lowest value since its scale began in 2005. For the 2009-born cohort the curricular tests show declines of 0.25-0.40 SD against PISA's 0.15; the United States is one of the systems (with Israel, Hungary, Chile) where the causal review notes TIMSS fell as much as or more than PISA, the opposite of the Nordic and French pattern. The apparent NAEP-PISA divergence (−8 against −15) is a matter of scale: compared in SD units PISA has fallen less than every national mathematics series since 2018.

Science. PISA science is flat across 2006-2025 (every change to 2025 not significant) and the US is 18 points above the OECD-35 average. NAEP grade 8 science rose from 150 (2009) to 154 (2015, 2019) and fell 3.9 points (SE 0.8, 0.11 SD) to 149.9 in 2024, its 2009 level; grade 12 science was flat over 2009-2019 and has not been assessed since; TIMSS grade 8 science fell 9 (n.s.) over 2019-2023 and 17 (significant) over 2015-2023. The curricular tests show a modest post-2019 decline that PISA does not, though a US PISA science change of about ±8 would be needed for significance and the 2025 science comparison is the least secure of the three domains (technical note).

Grade 12 versus age 15. Grade-12 NAEP is closest to PISA in age but shows the smallest changes (reading −2.8, mathematics −3.3 over 2019-2024, about 0.08 SD), partly because the lowest achievers are less likely to be present and tested in grade 12 than in grade 8; PISA at 15 precedes that attrition, which is one reason its US low-performer share moves more.

What the divergences imply. Where the series agree, the agreement does not depend on PISA's sampling: US achievement at ages 13-17 rose in mathematics through 2013, slipped in reading and mathematics from 2013 to 2019, fell sharply in mathematics in 2020-2022 without recovery, has continued to fall in reading through 2024-2025, and the fall is concentrated in the lowest-achieving third while the top decile is stable. Where they diverge, PISA is the outlier in two directions: less mathematics decline than the curricular tests (consistent with PISA mathematics being a reading-heavy literacy test whose reading component held up better in the US than elsewhere, and with the ±15-point sampling uncertainty) and a reading fall concentrated in 2022-2025 that NAEP spreads over 2013-2024. Neither divergence, given the SEs and sampling notes, supports a claim that the US trend differs in kind between PISA and NAEP. Chalkbeat's framing, that the US moved up the rankings because other countries declined (B5.1), is true only of relative position; every US series in Table 5 is lower in 2023-2025 than in 2019.

3. The explanations, applied to the United States

> **IN BRIEF** (summary added before publication; see Appendix H)
>
> The pandemic explanation fits the US better than the OECD overall. On NAEP the timing fits, mathematics fell more than reading, and losses grew with poverty and time spent remote, though even largely in-person states lost ground. The pandemic does not explain the 2013-2019 decline or the continuing fall in reading. Chronic absenteeism, one of the best-identified causes, could explain a material part of the missing recovery since 2022 but not the pre-2019 decline. Common Core effects lie between −0.10 and +0.10 standard deviations, more often slightly negative, and cannot explain a decline also found in states that never used the standards. Demographic change explains about a tenth of the NAEP decline since 2013 (a lower bound) and none of PISA's. Falling effort predicts larger declines on PISA than on other tests, but PISA mathematics fell less than NAEP or TIMSS in standard-deviation units, and reading about as much. Smartphones fit the US timing but are untestable with PISA's questionnaire.

Each explanation below is one the causal review treats generically and that has been applied to the United States in the OECD's materials, the claims inventory or the literature. For each: the prediction, the check, the sources. State checks use the NAEP state public-school series for 50 states and DC and the COVID-19 School Data Hub's district shares of 2020-21 spent in-person, hybrid and virtual (https://www.covidschooldatahub.com/data-resources), aggregated to states with 2020-21 enrollment weights.

3.1 Pandemic closures, their length by state, and the Education Recovery Scorecard

Prediction. States and districts that stayed remote longer in 2020-21 should have fallen more on NAEP over 2019-2022; the fall should be larger in mathematics and in high-poverty schools; recovery should be partial. The PISA 2025 cohort (grade 5-6 in 2020) and the 2022 cohort (grade 7-8) should both show the loss.

National check. Timing holds on every series (section 2.3); domain order holds (mathematics fell two to three times more than reading in SD terms on NAEP and LTT, and 13.4 against 1.4 on PISA over 2018-2022); recovery has not occurred (grade 8 mathematics 274.3 in 2022, 273.8 in 2024; reading down a further 2.4; the 2024 grade-12 cohort, in grade 8 when schools closed, is at the lowest level in both series). On PISA the 2022-2025 mathematics change is −2.0 (n.s.), so the 2009 cohort is at the level of the 2006 cohort: an unrecovered loss, not recovery.

State check. Across the 47 states and DC with schooling-mode data (Iowa, Montana and Oklahoma are absent from the Data Hub file), the enrollment-weighted share of 2020-21 not spent in person correlates with the 2019-2022 NAEP change at $r = -0.38$ (Spearman −0.42, $p = 0.003$) in grade 8 mathematics, −0.45 ($p = 0.001$) in grade 4 mathematics, −0.35 in grade 4 reading and 0.08 (n.s.) in grade 8 reading; the slope is about 0.3 NAEP points (0.008 SD) per 10 percentage points of the year not in person (`computed/state_checks_log.txt`). The most in-person states (Florida 97%, South Dakota 96%, Wyoming 93%, Texas 89%, Arkansas 87%) lost 5.4-7.5 grade-8 mathematics points; the least in-person (DC 0%, Maryland 2%, Arizona 2%, Kentucky 3%, Washington 5%, North Carolina 5%) lost 8.7-11.4. By 2024 the correlation has weakened ($r = -0.23$, $p = 0.12$, for 2019-2024; +0.13 for 2022-2024), so the longer-remote states have partly closed the gap, but no group of states is back to 2019. The district evidence points the same way: Jack, Halloran, Okun and Oster found 2019-2021 pass rates down 12.8 points in mathematics and 6.8 in ELA across 11 states with fully in-person districts showing significantly smaller declines (AER: Insights, 2023, https://www.aeaweb.org/articles?id=10.1257%2Faeri.20210748); Goldhaber, Kane, McEachin, Morton, Patterson and Staiger, on 2.1 million NWEA test-takers, found high-poverty schools remote all year lost 0.36 SD more in mathematics than high-poverty schools in person, hybrid instruction cost 0.03 SD in low-poverty and 0.15 SD in high-poverty schools, and schools in person all year still lost about 0.17 SD relative to pre-pandemic growth (https://files.eric.ed.gov/fulltext/ED620476.pdf). That last figure matters: even the most in-person states show losses, so closure length explains variation across states, not the national floor.

The Education Recovery Scorecard (Harvard CEPR and Stanford; Kane, Reardon and colleagues) links state tests to NAEP at district level. Its February 2025 release found "the average U.S. student remained nearly half a grade level behind pre-pandemic achievement in math and reading", 83% of students in districts still below 2019 in mathematics and 89% in reading, high-income districts "nearly 4 times more likely" than low-income ones to have recovered in both subjects (14.1% vs 3.9%), the affluent-poor mathematics gap up 11% since 2019, and federal relief reducing the loss "by 10 percent of a grade equivalent" in the highest-poverty districts (https://cepr.harvard.edu/news/education-recovery-scorecard). Its May 2026 release, using 2025 state tests, describes a "U-shaped recovery" (highest- and lowest-income districts improving most since 2022, middle-income districts lagging), names 108 districts up at least 0.3 grade levels in both subjects since 2022, and estimates that if student absence rates had returned to pre-pandemic levels "the recovery would have been meaningfully larger (0.03–

0.05 grade equivalents)" (https://www.gse.harvard.edu/ideas/news/26/05/new-education-scorecard-finds-u-shaped-recovery). These are the Scorecard's model-based statistics in its own "grade equivalent" unit.

Assessment. For the United States the pandemic explanation fits better than for the OECD as a whole (causal review, section 1): timing, domain order, poverty gradient and closure gradient all hold on NAEP, and PISA's 2018-2022 mathematics fall of 13.4 is the right size for a 0.1-0.2 SD unrecovered deficit. It does not explain the 2013-2019 slippage or the continuing reading decline through 2024-2025; NCEE's "long, long tail" (B5.1) describes non-recovery without explaining it. The marginal effect of closure length is real but modest (about 0.01 SD per 10 points of the year), and most of the loss occurred in states that were largely in person.

3.2 Chronic absenteeism

Prediction. Absence should have risen and stayed high, absent students should score lower, effects should be larger in mathematics and among the disadvantaged, and recovery should track the fall in absence.

Check. Chronic absenteeism (missing 10% or more of days) in public schools was about 15% in 2018-19, 28.3% in 2021-22, 25.7% in 2022-23, 23.5% in 2023-24 and 22.6% in 2024-25 across 45 states and DC (Malkus, AEI, June 2026, https://www.aei.org/research-products/report/from-attendance-crisis-to-chronic-condition-tracking-post-pandemic-chronic-absenteeism-into-2025/); the rise was proportionally similar in high- and low-achieving districts (10%→20% and 19%→35%), and children who entered school after the closures show rates "48 percent above the 2019 baseline". Dee's estimate from the federal 2021-22 data is +13.5 points, a 91% increase (PNAS, 2024, https://www.pnas.org/doi/10.1073/pnas.2312249121). PISA shows the same for 15-year-olds: the share of US students who skipped a whole day in the previous two weeks rose from 19.9% (2018) to 29.0% (2022) and 35.7% (2025), +15.9 points (SE 2.1), the largest rise of the 38 OECD countries and the third largest of 77 systems, against an OECD average of 21.7%, 19.8% and 22.4% (Table I.B1.4.161; 2025 US value with the OECD's dagger for 50-75% coverage). The causal literature gives 0.03-0.055 SD per ten days of absence (Aucejo and Romano 2016; Gershenson et al. 2017; causal review, section 12), so five to ten extra days a year sustained since 2021 implies 0.05-0.15 SD, or 5-15 PISA points, for affected students; the Scorecard's own estimate of aggregate recovery foregone is 0.03-0.05 grade equivalents.

Assessment. Absenteeism is one of the best-identified causes available for the United States and is of the right size for a material part of the non-recovery since 2022; the Scorecard finds high-absence districts recovering more slowly. It does not explain the pre-2019 decline, and the cross-country point stands: the United States had one of the largest documented rises in absence and one of the smaller PISA declines.

3.3 Smartphones and social media

Prediction. Declines should begin with adolescent smartphone saturation (2012-2015), be larger in reading and at the bottom, heavy users should score lower after controls, and bans should be followed by gains.

Check. Timing fits: NAEP grade 8 and grade 12 reading peaked in 2013 and LTT age 13 reading in 2012; 95% of US teenagers had a smartphone by 2024 and "nearly half" were online "almost constantly" against 24% a decade earlier (Pew, September-October 2024, https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/). Domain order holds only partly: reading turned down first (2013-2019), but mathematics has fallen more since 2019 on every US series, which a general-distraction account can accommodate and a reading-displacement account cannot alone. On PISA the US share spending more than an hour a day on social networks on a weekday was 75.4% (2022) and 80.3% (2025), the OECD averages being 75.1% and 80.1%; video games 59.5%→63.8% against 54.2%→54.9% (Table I.B1.4.39); the US time-on-device items (Tables I.B1.4.21, I.B1.4.30) are reported as missing for both years. Schleicher's "biggest increase in the U.S. is … browsing social networks" (A2.3) corresponds to +4.9 points, the OECD-average increase. Bans are less common in US schools: 13.6% of US students were in banning schools in 2022 and 31.0% in 2025, against OECD averages of 33.6% and 49.4% (Table I.B1.4.49); most state phone laws took effect from the 2024-25 school year, after or during the PISA window (I could not retrieve a verified count of states). The only US causal estimate, Figlio and Özek on Florida's 2023 ban, found second-year gains of 0.6-1.1 percentile points and a 5-10% fall in unexcused absences (NBER WP 34388, https://www.nber.org/system/files/working_papers/w34388/w34388.pdf), within the 0.02-0.14 SD range of the international ban studies.

Assessment. The US timing and distributional facts are consistent with the explanation, as in most rich countries, and the questionnaire cannot test it: the relevant quantities (out-of-school hours, age at first phone) are not measured and the US 2025 questionnaire is incomplete. Nothing in the US data separates it from other explanations with the same onset, and the ban evidence is small.

3.4 Common Core State Standards

Prediction. The standards were adopted by 45 states and DC in 2010-2011 and implemented from about 2012-2015 in K-12 ELA and mathematics; Alaska, Nebraska, Texas and Virginia never adopted them and Indiana (2014), Oklahoma (2014) and South Carolina (2015) withdrew. If the standards raised achievement, adopting states should have gained relative to non-adopters on NAEP grades 4 and 8 in reading and mathematics from about 2013, with the largest effect for fully exposed cohorts (kindergarten 2012, grade 8 in 2020-21); on PISA the 2018 cohort was the first with most of its middle-school years under the standards and the 2022 and 2025 cohorts had all of them. If the standards lowered achievement, adopting states should have fallen relative to non-adopters, in the two covered subjects and not in science.

Published evaluations. Loveless (Brown Center, 2014-2016) sorted states into non-adopters, medium and strong implementers; for 2009-2015 the largest advantage for adopters was +0.61 NAEP points (strong implementers vs non-adopters, grade 8 mathematics) and the largest for non-adopters +0.96 (grade 4 reading), which he called “trivial”, adding that the method “cannot determine causality” and the non-adopter group “is too small” (Brown Center Report 2016, pp. 12-13, https://www.brookings.edu/wp-content/uploads/2016/03/Brown-Center-Report-2016.pdf). Song, Garet, Yang and Atchison (AIR; American Educational Research Journal, 2022, 59(3), 610-647; https://files.eric.ed.gov/fulltext/ED620442.pdf), using a comparative interrupted time series on state NAEP 1990-2017 with treatment defined by the rigor of prior standards, found “significant negative effects for Grade 4 reading, with effect sizes ranging from −0.10 to −0.06” and “a significant negative effect for eighth graders 7 years after the adoption” in mathematics (−0.10 SD, significant under one of two classifications), most other estimates being “small, negative, and non-significant”. Xu and Cepa (CALDER WP 127, 2015, https://www.air.org/sites/default/files/2021-06/Kentucky-Common-Core-State-Standards-March2015.pdf), following three Kentucky cohorts from grade 8 to 11, found students exposed to the standards “made faster progress in learning than similar students who were not exposed”, while noting it was “not conclusive whether cross-cohort improvement was entirely attributable to the standards reform”. Polikoff (AERA Open, 2017, https://journals.sagepub.com/doi/full/10.1177/2332858417691749) concluded no convincing causal evidence existed either way: adoption dates do not mark implementation (41 states reported full implementation only in 2013-14 or 2014-15), adopters “certainly differed” from non-adopters, and NAEP is not aligned to the standards. Bleiberg (AERA Open, 2021, https://files.eric.ed.gov/fulltext/EJ1323929.pdf), comparing early with late implementers in NAEP microdata, found a positive initial mathematics effect of about 0.04-0.10 SD concentrated among economically advantaged students and none for disadvantaged students. Arold and Shakeel (PEPG 21-03, 2021, https://files.eric.ed.gov/fulltext/ED613142.pdf) estimate negative spillovers of about −0.08 SD in non-targeted subjects, largest for disadvantaged students.

Check on the state series (Table 6). Grade 8 mathematics fell in all three groups over 2009-2019: adopters −1.6 (mean of 43 states, SD 3.6), the four never-adopters −3.5 (SD 5.1), the three withdrawers −1.5. Grade 8 reading: adopters −0.9, never-adopters −4.7, withdrawers +0.2. Grade 4 mathematics: −0.4, +1.8, +1.3; grade 4 reading: −0.8, −3.2, −0.6. With four and three states in the comparison groups and within-group SDs of 2-5 points, none of these is distinguishable from zero, and the groups differ in demographics and other policies. The largest declines among the seven are in Alaska and Oklahoma, both with documented budget and staffing problems; the smallest in Indiana, whose replacement standards are widely described as similar to the ones withdrawn.

Table 6. NAEP grade 8 public-school means in states that did not adopt or withdrew from the Common Core, with national public and two comparison states

State	Status	Maths 2009	2013	2019	2024	Reading 2009	2013	2019	2024
Alaska	never adopted	283.0	281.6	274.3	264.0	259.4	261.3	252.4	246.4
Nebraska	never adopted	284.3	285.1	285.0	279.9	267.1	269.2	264.0	256.1
Texas	never adopted	286.7	288.2	279.7	269.4	260.4	263.7	255.7	252.3
Virginia	never adopted	286.1	288.1	287.1	275.0	265.6	267.6	261.8	256.3
Indiana	withdrew 2014	286.8	287.8	285.6	278.2	265.7	267.3	266.0	261.7
Oklahoma	withdrew 2014	275.7	275.5	276.5	264.5	259.5	261.9	258.2	248.8
South Carolina	withdrew 2015	280.4	279.8	276.4	268.0	257.3	261.4	259.0	253.7
National public		281.7	283.6	281.0	272.2	262.3	266.0	262.0	256.7
Massachusetts	adopted	298.9	300.6	294.5	283.5	273.6	277.0	273.1	268.0
Mississippi	adopted	265.0	271.2	273.7	269.1	251.3	253.1	256.4	253.5

Source: NAEP Data Service API, public schools (`us_series.csv`).

Assessment. The positive prediction fails: no evaluation finds NAEP gains above a few hundredths of an SD, and the best-identified estimates are small negatives (AIR, −0.06 to −0.10 SD in grade 4 reading and, less robustly, grade 8 mathematics). The large-negative prediction also fails: non-adopting states did not do better, and the national decline appears in non-adopters and in science, which the standards did not cover (Arold and Shakeel’s spillover notwithstanding). The 2013-2015 turn coincides with implementation but also with the end of recession-era cuts, the ESSA transition and smartphone diffusion, none of which the state comparison can separate. The evidence supports a bounded statement: effects between −0.10 and +0.10 SD, more often slightly negative, unable to account for a decline present in states that never used the standards. On PISA that would be at most 5-10 points against a 2012-2025 mathematics change of −18.5 (SE 7.5).

3.5 Reading instruction, the science-of-reading laws, and Mississippi

Prediction. If the shift to explicit, phonics-based early-reading instruction raises achievement, states that legislated it earliest should show grade 4 reading gains relative to others within about four years of enactment and grade 8 gains four years later; Mississippi’s Literacy-Based Promotion Act (2013; first retention cohort 2014-15) predicts grade 4 gains from 2015-2017 and grade 8 gains from 2019-2021. Because a third-grade retention gate removes the weakest readers from the grade 4 cohort for a year, part of a grade 4 gain could be compositional and would not carry to grade 8.

Check. Mississippi's NAEP grade 4 reading rose from 208.5 (2013, rank 49 of 51) to 219.3 (2019, rank 29) and 218.5 (2024, rank 9) while the national public mean fell from 220.7 to 214.3; grade 4 mathematics rose from 231.1 to 240.7 (2019) and 239.2 (2024) against 241.2→237.3 nationally. Grade 8 reading went 253.1 (2013) → 256.4 (2019) → 253.5 (2024) against 266.0→256.7, so Mississippi's grade 8 reading is unchanged over eleven years while the nation fell 9 points and its rank moved from 50th to 41st; grade 8 mathematics 271.2→273.7→269.1 against 283.6→272.2. Louisiana, with a later similar package, shows a similar grade 4 rise (210.5 to 216.0) and flat grade 8 (257.4→256.5). On components: Slungaard Mumma and Winters, using a regression discontinuity at the 2015 retention cut-off, found retained third-graders scored 1.2 SD higher in sixth-grade ELA than promoted students just above the cut-off, with no effect on mathematics, absences or special-education placement (Wheelock, 2023, https://wheelockpolicycenter.org/wp-content/uploads/2023/02/WEPC-MS-Retention-Policy-Brief-02-03-2023.pdf); Berne, Jacob, Weiland and Strunk, with the same design in Michigan, found being flagged for retention raised next-year reading by 0.045 SD even in districts that retained no one, so "the implied effect of retention itself … would be an implausibly large 1.3 SD" and previous retention estimates may be biased upward by the support flagged students receive (EdWorkingPaper 25-1188, 2025, https://edworkingpapers.com/sites/default/files/ai25-1188.pdf); Spencer, using synthetic difference-in-differences on state NAEP, estimates the LBPA "meaningfully improved grade 4 reading and math test scores" (Economics of Education Review, 2024, https://ideas.repec.org/a/eee/ecoedu/v103y2024ics027277572400092x.html). On composition, Fordham's Todd Collins argued in 2019 that a K-3 retention rate of 8% could inflate grade 4 scores and in an August 2022 update concluded from NAEP age data that retention was "not a major contributor" because test-taker ages were "almost identical in 2002 and 2017" (https://fordhaminstitute.org/national/commentary/mississippi-rising-partial-explanation-its-naep-improvement-it-holds-students).

Across states, the Scorecard's May 2026 release reports that all eight states it classifies as having comprehensive early-literacy reforms by January 2024 (DC, Indiana, Kentucky, Maryland, Minnesota, Mississippi, Louisiana, Tennessee) improved in reading over 2022-2025 while ten states without such reforms did not. On NAEP grade 4 reading those eight averaged +1.9 points over 2022-2024 against −2.0 for the other 43 jurisdictions, and −2.6 against −5.9 over 2019-2024; at grade 8, +0.3 against −2.2 and −3.1 against −5.9 (`computed/state_checks_log.txt`). The classification is the Scorecard's, made after the outcomes were known, and Louisiana's grade 4 rise predates its 2021 law; this is an association.

Assessment. Mississippi's grade 4 series is one of the few in the country that rose through 2013-2024, the timing fits the 2013 law, and the retention RD shows large gains for retained students, with the Michigan replication indicating the gain comes at least partly from the intervention attached to being flagged rather than the repeated year. Grade 8 has held flat while the nation fell, an improvement in relative terms only. PISA cannot see it: the first LBPA cohort reached 15 in 2021-22 and PISA reports no state results. The literacy laws postdate the national reading decline and cannot explain it; they are the best-documented US case of a state policy that predicts its own NAEP trend at the grade it targets.

3.6 Accountability: NCLB and ESSA

Prediction. The strong-accountability version says NCLB (signed January 2002) raised achievement in tested grades and subjects (3-8 reading and mathematics), more for low achievers, and that the 2011-12 waivers and ESSA (December 2015, in force 2017-18) removed the pressure, predicting a flattening from 2013-2017. The narrowing version says the NAEP gains should be confined to tested content and absent from PISA and grade 12.

Check. Dee and Jacob's comparative interrupted time series, using states with pre-2002 consequential accountability as the comparison, found NCLB "generated statistically significant increases in the average math performance of fourth graders (effect size = 0.23 by 2007)", eighth-grade mathematics gains "particularly among traditionally low-achieving groups and at the lower percentiles", and "no evidence that NCLB increased fourth-grade reading achievement" (JPAM, 2011, https://tom-dee.github.io/files/jpam2011.pdf). The series fit: grade 8 mathematics rose 7.0 points (0.18 SD) over 2003-2013 and reading 4.3 (0.11 SD), both peaking in 2013, the last year before the waiver-era and ESSA changes; LTT age 13 mathematics rose 276→285 over 1999-2012 on an unchanged instrument. Grade 12 mathematics rose 3.3 over 2005-2009 and reading 2.3, then both were flat to 2013. PISA mathematics did not rise over 2003-2012 (483→481), which the narrowing version predicts, though US PISA changes in that window have SEs of about 7 and cannot resolve 0.1 SD. After 2013 every series flattened or fell, which the weak-accountability version predicts, as do the smartphone, Common Core and recession-spending explanations for the same years; no study separates them. The Bergbauer-Hanushek-Woessmann cross-country panel (causal review, section 8) runs in the same direction.

Assessment. The NCLB gains are among the better-identified effects in the US record, 0.1-0.2 SD in mathematics and concentrated at the bottom. They are not, however, the only period in which the bottom of the US distribution rose: the NAEP grade 8 mathematics 10th percentile rose from 215.0 (1990) to 229.9 (2003) before NCLB took effect, a larger rise than the 229.9 to 237.4 of 2003-2013, and the LTT age 13 mathematics 10th percentile rose from 213.3 (1978) to 233.8 (1999) and 240.1 (2012); grade 8 reading's 10th percentile rose from 212.5 (1992) to 223.1 (2013) (`us_series.csv`). The bottom of the distribution rose through the 1990s and 2000s under several policy regimes and turned down after 2013. Whether the post-2013 flattening is the unwinding of accountability pressure cannot be established from the aggregate series. Secretary McMahon's claim that the "one-size-fits-all federal education bureaucracy has shortchanged our children" (B13.1) is a policy position; the evidence offered (the reading low, the widest gap) is accurate but does not bear on federal structure, and the period of strongest federal accountability is the period of NAEP gains.

3.7 Demographic composition of the tested cohort

Prediction. The decline should shrink substantially within demographic groups, and the groups whose share has grown should be lower scoring.

Check. On NAEP grade 8 the White share of tested students fell from 62.7% (2003) to 54.5% (2013) and 45.5% (2024), the Hispanic share rose from 14.6% to 22.4% and 28.3%, English learners from 4.6% to 9.6%, students with disabilities from 9.5% to 13.3% (`computed/naep_g8_composition.csv`). Holding the 2013 race/ethnicity composition fixed, the 2024 grade 8 reading mean would be 259.1 instead of 258.1, so composition accounts for about 1.0 of the 9.5-point decline since 2013, and 1.3 of 10.9 in mathematics. The within-group declines are general: White −9.0 (reading) and −8.3 (mathematics), Hispanic −10.1 and −13.1, Black −7.3 and −11.5, Asian/Pacific Islander −0.2 and −1.2. On PISA the US immigrant-background share has been stable, 23.1% (2015), 23.0%, 23.7%, 25.2% (2025; SE 3.7), while the OECD-29 share rose from 12.8% to 18.0% (Table I.B1.2d.1), and in 2025 US immigrant students scored 7-18 points above non-immigrants before adjustment (Tables I.B1.2d.7-9, on the questionnaire sub-sample). The rise in the tested shares of English learners and students with disabilities reflects both population change and NAEP's own March 2010 inclusion policy, adopted by NCES with the Governing Board "to maximize the participation of students with disabilities (SD) and English learners (EL)"; in 2024 about 90% of such students were assessed at grades 4 and 8 (https://nces.ed.gov/nationsreportcard/about/inclusion.aspx). Greater inclusion of lower-scoring students works in the same direction as the race shift, so the composition effect above is a lower bound, while PISA's higher exclusion rate works the other way.

Assessment. Composition explains about a tenth of the NAEP decline since 2013 and none of the PISA decline. The fall is within every group except Asian/Pacific Islander students and is largest among Hispanic and Black students in mathematics, where the bottom-decile falls of sections 1.2 and 2.3 are concentrated.

3.8 Test-taking effort, the instrument and the sample

Prediction. If US students exert less effort on PISA than on other tests or than before, PISA should understate US achievement relative to NAEP and TIMSS, and the US decline should be larger on PISA than on the high-participation national tests.

Check. The experimental evidence is US-specific: Gneezy and colleagues found a financial incentive raised US students' scores on a PISA-style test by about 0.2 SD with no effect in Shanghai, enough to move the United States from 36th to about 19th in the 2012 mathematics ranking (AER: Insights, 2019; validity review). That concerns levels. On changes the prediction fails: in SD units PISA mathematics fell less (0.15 over 2018-2025) than NAEP grade 8 (0.21), LTT age 13 (0.25 over 2020-2023) or TIMSS (0.27), and PISA reading fell about as much as LTT age 13 reading on the same cohort. The United States is therefore not a case where the PISA-specific component looks large, unlike the Nordic countries (causal review, section 9). The sampling issue is different: the ±15-point bounds in 2022 and 2025 are wider than the 2022-2025 changes and of unknown sign; NAEP's grade 8 samples, with school participation above 85% and student participation above 90%, do not have this problem, which is the main reason to weight NAEP above PISA for US trends. The inventory's sampling claim (B15.1) is correct as far as it goes; it does not make the US decline an artefact, because the strongly sampled national series show declines at least as large.

3.9 Other explanations applied to the United States

Spending. Real current expenditure per pupil in US public schools fell from $14,714 (2008-09) to 13, 952(2012−13)*inconstant*2022 − 23*dollars*, *a*5.214,894) and reached $16,280 in 2020-21 (Digest of Education Statistics 2023, Table 236.55, https://nces.ed.gov/programs/digest/d23/tables/dt23_236.55.asp). Jackson, Wigger and Xiong's estimate of about 0.03 SD per $1,000 per pupil applies to cohorts schooled during the cuts, so the children who were in grades K-6 in 2009-2015 are the grade 8 NAEP cohorts of 2013-2022 and the PISA cohorts of 2018-2025; on that estimate a cut of about $760 per pupil sustained for five or six years would lower those cohorts' scores by roughly 0.02-0.03 SD, or 1 NAEP point and 2-3 PISA points. The timing therefore fits the post-2013 flattening as a small contributor, not the pandemic-era fall, which coincided with the largest federal transfer in US education history (ESSER; the Scorecard estimates it reduced the loss by about 0.1 grade equivalents in high-poverty districts). Spending may explain part of the state variation (Alaska, Oklahoma), which I have not tested.

Teacher relationships and growth mindset. NCEE's suggestion that "relatively positive relationships with teachers, higher growth mindset, and more access to peer-to-peer tutoring" explain the smaller US decline (B8.5) rests on questionnaire data the OECD reports below the line because 40% of respondents were not surveyed; it cannot be checked on the 2025 data.

4. What the US student questionnaire shows on devices, absence and reading enjoyment

A caveat first. More than 40% of US respondents were not given the questionnaire, "in roughly equal parts" because nine states declined to administer it, which "introduce[s] systematic gaps in coverage", and because some schools tested while administration was suspended nationally, which the OECD treats as unrelated to student characteristics (Reader's Guide, p. 19). The OECD prints US questionnaire results "below the line" with a dagger for 50-75% coverage, the country note omits them, and the NCES report carries none. The figures below describe the covered part of the sample.

Device use. The US share spending more than an hour on social networks on a weekday was 75.4% in 2022 and 80.3% in 2025 (+4.9, SE 1.5), identical to the OECD averages (75.1%, 80.1%); more than an hour on video games 59.5%→63.8% against 54.2%→54.9%; browsing the internet 70.0%→74.3% against 65.6%→62.9% (Table I.B1.4.39). The share reporting distraction from devices in most or every science lesson was 29.3% against 28.4% (Table I.B1.4.43). The US time-on-device items (learning and leisure use at school; Tables I.B1.4.21, I.B1.4.30) are missing for 2022 and 2025, so Schleicher's "1.7 hours per day on devices for learning and then another hour for leisure while in school" (A2.3) is the OECD average (1.73 and 1.06 hours, Table I.B1.4.21), not a US figure. Phone bans reach fewer US students than elsewhere (31.0% in banning schools in 2025 against 49.4%; 13.6% against 33.6% in 2022), and 62% of US students were in schools with subject-specific guidelines but no ban (Table I.B1.4.49).

Absenteeism. The United States has the largest rise in self-reported whole-day absence of any OECD country: 19.9% (2018), 29.0% (2022), 35.7% (2025), against OECD averages of 21.7%, 19.8% and 22.4%; skipping classes rose from 18.8% to 21.7% and lateness from 43.3% to 51.4% (Table I.B1.4.161). In level the US ranks 7th of 38 OECD countries and 29th of 92 systems in 2025 (Italy 63%, Costa Rica 54%, Türkiye 53%, Australia 44%, Ireland 40% are higher within the OECD). The administrative series (section 3.2) peaks in 2021-22 and declines slowly; the PISA two-week self-report has continued to rise to 2025.

Reading enjoyment. PISA 2025 has no reading-enjoyment index, so the latest US data are from 2018. On the OECD index (OECD mean 0, SD 1 in 2009) the US value was −0.04 in 2009 and −0.07 in 2018 (change −0.03, SE 0.03, not significant), against an OECD average of 0.00 and −0.06; the largest declines were Germany (−0.36), Norway (−0.32) and Finland (−0.30) (OECD, *21st-Century Readers*, 2021, Figure 4.2, https://doi.org/10.1787/888934239857). US 15-year-olds reported slightly below-average enjoyment in both years and no measurable change over the decade in which enjoyment fell in the countries that have since had the largest PISA reading declines. Outside PISA, the American Time Use Survey shows the share of Americans reading for pleasure on a given day falling about 40% between 2003 and 2023 (Wei and Lee, iScience 2025, https://www.ucl.ac.uk/news/2025/aug/proportion-americans-reading-pleasure-fell-40-over-20-years), an adult measure.

5. Summary

The US PISA means in reading and science in 2025 are not significantly different from any earlier cycle; mathematics is 20 points below 2003 and 15 below 2018; and in all three domains the US position relative to the OECD average has improved by 20-30 points since 2012 because other countries fell faster. Beneath the flat means the distribution has pulled apart: the 10th percentile is 30-47 points below its 2009-2012 level in reading and 20-34 in mathematics, the 90th is stable or higher, and the US has the widest inter-decile range in reading of any system. All of this is also present on NAEP at ages 13 and 17 and on TIMSS grade 8, with far stronger samples, and the national series date the turn to 2013-2015 in reading and 2019-2022 in mathematics. US PISA samples have never met the school-response standard, response has fallen to 45% of schools and 76% of students, the OECD bounds the possible bias at about ±15 points of unknown sign, and the 2022 baseline carries the same caveat, so US PISA changes between adjacent cycles should not be read to a precision better than 10-15 points. Of the explanations examined, the pandemic (with closure length as a modifier only: schools that stayed in person all year still lost about 0.17 SD, so closure weeks are a poor proxy for pandemic exposure), chronic absenteeism and the NCLB accountability gains have US-specific causal evidence of the right size and timing for parts of the record; Common Core has been evaluated repeatedly with effects between −0.10 and +0.10 SD and cannot account for a decline present in non-adopting states and untargeted subjects; the Mississippi literacy package predicts its own grade 4 series and holds at grade 8 only in relative terms; composition explains about a tenth of the NAEP decline and none of the PISA decline; and smartphones fit the timing without any US measure that could test them. The continuing decline of reading at the bottom of the distribution from 2013 to 2025, across all series, demographic groups and states, is the fact that none of the explanations with strong US evidence covers.

Changelog (revision after the adversarial review)

Finding 3(b): the statement that the NCLB years were “the only period in which the bottom of the US distribution rose” was false against `us_series.csv` and has been replaced in section 3.6 with the 1978-2013 10th-percentile figures (grade 8 mathematics p10 215.0 in 1990, 229.9 in 2003, 237.4 in 2013; LTT age 13 mathematics p10 213.3 in 1978, 233.8 in 1999, 240.1 in 2012; grade 8 reading p10 212.5 in 1992, 223.1 in 2013). Finding 3(c): the grade 12 mathematics 2019-2024 change is −3.33 (150.28 to 146.95); the annex used −3.3 throughout, and the unrounded values are now given in section 2.4 so that the rounded published means (150.3, 146.9) do not suggest 3.4. Finding 15: Table I.2.7’s US yearly learning gains (science 13.8, reading 12.0, mathematics 13.5; verified in vol1.txt) and Box I.2.2’s caution are added in section 1.2 where the 14-point reading fall is converted to time, against the “seven months” in claims inventory B1.1. Finding 14: the US change from autumn to spring testing in 2025 (verified in the NCES 2025 and 2022 technical notes) is added to sections 1.3 and 2.4. Finding 2(d): the Goldhaber et al. finding that schools in person all year still lost about 0.17 SD is now carried into the summary’s statement on closures. Finding 7: the spending paragraph (3.9) now uses the Digest of Education Statistics Table 236.55 series (verified by direct fetch) and applies Jackson, Wigger and Xiong’s estimate to the cohorts actually exposed, which places the effect in the post-2013 flattening rather than at the NAEP peak. Finding 16: NAEP’s 2010 inclusion policy (verified at nces.ed.gov) is added to section 3.7 as a compositional factor. Not changed: the reviewer’s Common Sense Census figures were not fetched and are not cited; no other numerical check of the annex failed (A1-A8, A10 pass).

Appendices

The appendices are the supporting sections as the agents wrote them, retitled, with the quotation corrections listed in Appendix H. References to working files such as `A/pisa_trends.csv` point to the dataset published with this report (`pisa_trends.csv`); line numbers in `vol1.txt` refer to a plain-text extraction of the OECD's *PISA 2025 Results (Volume I)* PDF, not to its printed pages.

Appendix A. What the PISA 2025 trend tables show (factual summary)

IN BRIEF (summary added before publication; see Appendix H)

This section summarises what the trend tables of PISA 2025 Results (Volume I) show and does not interpret causes. On the OECD-35 average, mean scores fell between 2022 and 2025 by 14.3 points in reading and 9.4 in mathematics, while the 2.8-point fall in science was not significant. On the OECD-23 series, 2025 performance is the lowest in all three subjects, significantly below every earlier cycle except science in 2022. Among the 74 systems comparable with 2022, reading fell significantly in 42 and mathematics in 38, while science rose in 15 and fell in 11. In mathematics the OECD-35 decline over 2022–2025 was 13.6 points at the 10th percentile against 5.5 at the 90th, and the share of low performers reached 30.4% in reading and 33.9% in mathematics. Over the same period the OECD-35 socio-economic gap narrowed in all three subjects, mainly because advantaged students lost ground rather than because disadvantaged students improved. Six systems carry asterisks for sampling shortfalls.

All numbers below are taken from the StatLink workbooks of *PISA 2025 Results (Volume I): Future-Ready Students* (OECD, 8 September 2026, https://doi.org/10.1787/73451bc5-en), as assembled in `pisa_trends.csv` (see `data_dictionary.md` for table references). Standard errors (SE) are the OECD's; for changes between cycles they already include the link error (Annex A3). "95% CI" is estimate ± 1.96 SE. "Significant" means |estimate/SE| > 1.96, which is the rule embedded in the workbooks' conditional formatting for bold values. Counts of countries are computed from the annex tables over all entities the OECD treats as comparable (values present rather than "m"); where the OECD's own count in the text differs, both are given. OECD averages over time use the OECD's consistent-membership aggregates: "OECD-23" (the 23 members with data in every cycle since 2000) and "OECD-35" (all members except Costa Rica, Luxembourg and Spain, comparable since 2015). Nothing here interprets causes.

1. OECD-average trends over the full series

Table 1 gives the OECD-23 series, which is the only aggregate the OECD reports for every cycle (Tables I.B1.2a.36-38). The OECD-35 series since 2006/2012 is in Table 2.

Table 1. OECD-23 mean scores by cycle (SE in parentheses) and change to 2025

Cycle	Reading	Mathematics	Science
2000	499.5 (0.7)	–	–
2003	497.4 (0.6)	502.1 (0.6)	–
2006	495.5 (0.7)	501.1 (0.6)	502.7 (0.6)
2009	499.4 (0.6)	501.7 (0.6)	505.7 (0.6)
2012	501.4 (0.6)	498.7 (0.6)	504.7 (0.6)
2015	497.2 (0.6)	495.5 (0.5)	497.0 (0.5)
2018	492.8 (0.5)	496.4 (0.5)	493.5 (0.5)
2022	481.9 (0.6)	480.3 (0.5)	490.9 (0.5)
2025	465.8 (0.6)	469.4 (0.6)	485.9 (0.6)
Change, first cycle → 2025	−33.7 (SE 6.8; CI −47.1 to −20.3)	−32.7 (5.7; −43.9 to −21.5)	−16.8 (4.9; −26.4 to −7.2)
Change 2018 → 2025	−27.0 (2.0)	−27.0 (2.7)	−7.5 (3.6)
Change 2022 → 2025	−16.1 (1.3)	−10.9 (1.4)	−4.9 (3.2), not significant

Source: Tables I.B1.2a.36, I.B1.2a.37, I.B1.2a.38.

Table 2. OECD-35 mean scores and changes (the aggregate used in the OECD's text)

	Reading	Mathematics	Science
2015	490.3 (0.5)	487.2 (0.4)	490.8 (0.4)
2018	487.6 (0.4)	489.7 (0.4)	489.2 (0.4)
2022	477.4 (0.5)	474.8 (0.4)	486.8 (0.5)
2025	463.0 (0.5)	465.5 (0.5)	483.9 (0.5)
Change 2022→2025	−14.3 (1.3; CI −16.9 to −11.8)	−9.4 (1.4; −12.0 to −6.7)	−2.8 (3.2; −9.1 to +3.4)

	Reading	Mathematics	Science
Change 2018→2025	−24.6 (1.9)	−24.2 (2.6)	−5.2 (3.6), n.s.
Change 2015→2025	−27.3 (3.9)	−21.8 (3.1)	−6.9 (3.5)
Average decennial trend 2015-2025 (points per 10 years)	−27.0 (3.3; p < 0.001)	−23.9 (3.1; p < 0.001)	−6.7 (3.1; p = 0.03)

Source: Tables I.B1.2a.36-38. The 38-country "OECD average" for 2025 is 482 in science, 461 in reading and 463 in mathematics (Table I.B1.2a.1-3); the OECD does not report it for 2018 or 2022 in the trend tables because not every current member has comparable data in those cycles.

The shape of the three series differs. Reading on the OECD-23 aggregate is flat between 2000 and 2012 (all cycles within 6 points of 500), then declines in each subsequent cycle: −4.1 (2012→2015), −4.4 (2015→2018), −10.9 (2018→2022), −16.1 (2022→2025). Mathematics stays within 502.1-495.5 from 2003 to 2018 and then loses 16.1 points to 2022 and a further 10.9 to 2025. Science peaks in 2009 (505.7), falls to 497.0 by 2015 and 493.5 by 2018, and then changes little: 490.9 in 2022 and 485.9 in 2025; the 2022→2025 change of −4.9 (SE 3.2) is not significant on the OECD-23 aggregate, nor is −2.8 (SE 3.2) on the OECD-35 aggregate, nor is the OECD-35 2018→2025 change of −5.2 (SE 3.6). Over 2015→2025 the OECD-35 science decline of −6.9 (SE 3.5) is significant, and the fitted decennial trend is −6.7 points per decade (p = 0.03).

The OECD's headline statements check out against these cells. "On average across 35 OECD countries, mean performance did not change significantly in science between 2022 and 2025 … dropped by 9 score points in mathematics and about 14 score points in reading" (p. 91) corresponds to −2.8, −9.4 and −14.3. "Performance in PISA 2025 was the lowest in all subjects, significantly below the mean performance observed in any earlier assessment (except for science in PISA 2022)" (p. 96) holds for the OECD-23 aggregate on which Figure I.2.6 is based: every change-to-2025 cell is significant except science 2022. On the OECD-35 aggregate, however, science 2018→2025 (−5.2, SE 3.6) is also not significant, so the "except 2022" qualifier is specific to the OECD-23 series. Annex A1's "25 points in total between 2018 and 2025" for reading corresponds to −24.6.

2. Distribution of country-level changes

Table I.2.6 of the report lists 74 countries/economies that can compare 2022 and 2025 in all three domains, plus the OECD-35 average. The annex trend tables carry one further comparable entity in all three domains, Cyprus, and one in science only, Uzbekistan; the counts below are given both ways.

Table 3. Number of countries/economies with significant change in mean score, by base year

Base → 2025	Domain	n	Significantly up	Significantly down	No significant change
2022 (Table I.2.6 set, n = 74)	Science	74	15	11	48
	Reading	74	8	42	24
	Mathematics	74	8	38	28
2022 (all comparable in annex tables)	Science	76	16	11	49
	Reading	75	8	42	25
	Mathematics	75	8	39	28
2018	Science	74	17	21	36
	Reading	74	5	56	13
	Mathematics	75	9	56	10
2015	Science	61	11	23	27
	Reading	61	3	47	11
	Mathematics	62	6	40	16

Source: Tables I.B1.2a.36-38 and Table I.2.6; significance from the reported SE. The OECD text (p. 91) gives "48 out of 74" stable in science and "42 and 38 out of 74" declining in reading and mathematics; the annex tables add Cyprus (mathematics −6.5, SE 2.2, significant; reading −1.6 and science +0.3, not significant) and Uzbekistan (science +82.7, SE 4.6). The Reader's Guide (p. 19) notes that Uzbekistan's reading and mathematics data are withheld because of response-data inconsistencies; its science data are reported without an asterisk.

Between 2022 and 2025, eight systems improved significantly in all three domains: Cambodia (science +34.8, reading +18.4, mathematics +29.4), Georgia (+38.1, +10.3, +26.3), Jordan (+32.4, +20.4, +26.3), Montenegro (+32.1, +13.4, +5.5), the Philippines (+17.2, +20.4, +15.9), Thailand (+23.0, +13.2, +12.8), Türkiye (+18.3, +16.0, +8.4) and the United Arab Emirates (+26.1, +15.5, +21.5). Nine declined significantly in all three: Argentina, Denmark, Guatemala, Hong Kong (China), Israel, Latvia, Malta, Serbia and Slovenia. The largest three-year declines are in Latvia (reading −44.9, SE 3.4; science −25.5, SE 4.4; mathematics −22.7, SE 2.9), Israel (−38.0, −22.7, −24.5), Guatemala (reading −37.0), Malta (reading −30.7, mathematics −26.9) and Denmark (reading −28.8).

Among OECD countries, science rose significantly only in Costa Rica (+13.0), the Slovak Republic (+12.6), Türkiye (+18.3) and the United Kingdom (+11.7), as the executive summary states.

Over the longer 2018→2025 window the picture is more one-sided: reading fell significantly in 56 of 74 comparable systems and rose in five (Brunei Darussalam, Cambodia, the Philippines, Qatar and Zambia), which matches Annex A1's "56 out of 74 … and only five"; mathematics fell in 56 of 75; science fell in 21 and rose in 17. Only Cambodia, the Philippines and Zambia rose significantly in all three domains over 2018→2025 (Zambia's 2018 value is the PISA-for-Development 2017 result, and Cambodia's likewise), while 20 systems fell in all three. The eight largest 2018→2025 reading declines are Iceland (−51.7, SE 3.0), Slovenia (−50.6), Latvia (−49.0), Norway (−46.6), Finland (−45.8), Cyprus (−44.9), Hong Kong (China) (−44.2) and the Netherlands (−43.4).

3. The top and the bottom of the within-country distribution

Changes in the mean conceal different movements at the two ends. Table 4 shows OECD-35 percentiles.

Table 4. OECD-35 10th and 90th percentiles and changes (SE)

	Science p10	Science p90	Reading p10	Reading p90	Mathematics p10	Mathematics p90
2006/2012*	371.3 / 378.9	616.5 / 614.7	– / 370.2	– / 609.2	373.2 / 372.9	607.7 / 609.1
2015	366.2	611.9	362.0	610.4	370.3	601.2
2018	365.2	609.4	355.2	614.2	370.7	605.0
2022	357.6	613.5	342.9	605.8	356.1	593.5
2025	352.6	611.8	327.3	593.0	342.5	588.1
Change 2022→2025	−5.0 (3.2) n.s.	−1.7 (3.2) n.s.	−15.5 (1.5)	−12.8 (1.4)	−13.6 (1.5)	−5.5 (1.5)
Change 2018→2025	−12.6 (3.6)	+2.4 (3.6) n.s.	−27.9 (2.1)	−21.2 (2.0)	−28.3 (2.7)	−16.9 (2.7)
Change 2006/2012*→2025	−18.6 (4.9)	−4.7 (4.9) n.s.	−42.9 (6.2)	−16.2 (6.2)	−30.8 (4.4)	−19.6 (4.4)

* Science and mathematics from 2006, reading from 2012 (the first cycle for which the OECD-35 aggregate is reported). Source: Tables I.B1.2a.39-41.

In science the OECD-35 mean did not change significantly between 2022 and 2025, and neither percentile did either; but over 2018→2025 and 2006→2025 the 10th percentile fell significantly (−12.6 and −18.6) while the 90th did not move (+2.4 and −4.7, both within their SE). The inter-decile range in science is 259.2 points in 2025, up 3.3 from 2022 and about 15 from 2018 (244.2), and the OECD calls it "the largest measured in PISA so far" (p. 92; Table I.B1.2a.42); the 2025 10th percentile, 352.6, is the lowest in the science series. In reading the whole distribution shifted down between 2022 and 2025, slightly more at the bottom (−15.5) than at the top (−12.8), so the 90-10 range widened by 2.7 points; since 2012 the bottom has fallen by 42.9 points and the top by 16.2. In mathematics the 2022→2025 decline is 13.6 at the 10th percentile against 5.5 at the 90th, widening the range by 8.1 points; the OECD's "−7 and −5 score points" for the 75th and 90th percentiles correspond to −6.9 and −5.5.

At the country level, the decoupling is most visible in science and mathematics. Between 2022 and 2025 the 10th percentile in science fell significantly in 18 systems (Argentina, Bulgaria, Croatia, Denmark, Guatemala, Hong Kong (China), Israel, Japan, Latvia, Malta, Morocco, North Macedonia, the Palestinian Authority, Paraguay, Serbia, Singapore, Slovenia, Spain) and rose in 11, while the 90th percentile fell in only 6 and rose in 15; in 13 systems the 10th percentile fell without a significant fall at the 90th, against one case of the reverse. In mathematics the 10th percentile fell in 45 of 75 systems and the 90th in 24; in 27 systems only the bottom fell. In reading both ends fell together in 31 systems, the bottom alone in 6 and the top alone in 11. The inter-decile range widened significantly between 2022 and 2025 in 24 systems in science (narrowing only in Peru), in 20 in mathematics (narrowing in Cambodia, Guatemala and Paraguay) and in 10 in reading (narrowing in 11), which reproduces the OECD's counts on p. 94 (Tables I.B1.2a.42-44). Over 2018→2025 the science 10th percentile fell in 31 systems and the 90th in 6, and the 90th rose in 22.

4. Low performers and top performers

Table 5. OECD-35 share of students below Level 2 and at Level 5 or above (%, SE)

	Science < L2	Science ≥ L5	Reading < L2	Reading ≥ L5	Mathematics < L2	Mathematics ≥ L5
2015	22.1 (0.2)	7.5 (0.1)	20.8 (0.2)	8.1 (0.1)	24.6 (0.2)	10.4 (0.1)
2018	21.9	6.9	22.4	8.8	23.9	11.0
2022	23.8	7.7	25.7	7.4	30.0	9.0
2025	25.0 (0.2)	7.5 (0.1)	30.4 (0.2)	5.8 (0.1)	33.9 (0.2)	8.1 (0.1)

	Science < L2	Science ≥ L5	Reading < L2	Reading ≥ L5	Mathematics < L2	Mathematics ≥ L5
Change 2022→2025 (pp)	+1.2 (1.0) n.s.	−0.3 (0.5) n.s.	+4.7 (0.5)	−1.7 (0.2)	+3.8 (0.5)	−0.9 (0.2)
Change 2018→2025	+3.2 (1.1)	+0.6 (0.5) n.s.	+7.9 (0.7)	−3.0 (0.3)	+10.0 (1.0)	−2.9 (0.4)
Change 2015→2025	+3.0 (1.1)	−0.1 (0.5) n.s.	+9.5 (1.4)	−2.3 (0.5)	+9.2 (1.1)	−2.3 (0.5)

Source: Tables I.B1.2a.33-35. The 38-country OECD average for 2025 is 25.7% below Level 2 and 7.2% at Level 5-6 in science (Table I.B1.2a.10), which is the basis of the executive-summary statement that "around three-quarters of students demonstrated at least baseline proficiency".

The OECD's statements that the science shares "did not change significantly between 2022 and 2025", that low performers in science rose "by three percentage points between 2015 and 2025" with top performers unchanged, and that low performers rose "by 10 percentage points" in reading and "about 9" in mathematics since 2015 while top shares fell by about two points (pp. 94-96) all match the cells above (+1.2 n.s.; +3.0; +9.5; +9.2; −2.3; −2.3). Across systems, between 2022 and 2025 the share of low performers in science rose significantly in 14, fell in 16 and did not change in 46 (the text says 45, again omitting Cyprus); the top share rose in 7, fell in 7 and was unchanged in 62 (text: 61). Only Denmark and Latvia combined more low and fewer top performers, while Jordan, Montenegro, Saudi Arabia, Türkiye, the United Arab Emirates, the United Kingdom and Uzbekistan combined fewer low and more top performers, as the report states. In reading the low-performer share rose in 40 systems and fell in 8, and the top share fell in 29 and rose in one; in mathematics the low share rose in 33 and fell in 8, the top share fell in 17 and rose in 3.

Over 2015→2025 the annex tables show 13 systems with rising low-performer shares and unchanged top shares in science and 8 with rising low and falling top shares, matching the text; five (Montenegro, Qatar, the Slovak Republic, Türkiye, the United Arab Emirates) both cut low performers and raised top performers, and twelve cut low performers. The text adds Colombia, Korea, Macao (China), Chinese Taipei and the United States as systems that raised the top share; the table also shows significant increases for Cyprus (+1.1, SE 0.4), Luxembourg (+1.7, SE 0.8) and Romania (+1.2, SE 0.4), all three of which simultaneously saw rising low-performer shares.

Coverage matters for these shares: Coverage Index 3 in 2025 ranges from 0.32 (Rwanda) and 0.43 (Cambodia) to 0.90 or more in 34 systems (Table I.A2.1), as the report states on p. 112; it is below 0.70 in Guatemala (0.55), the Palestinian Authority (0.54, down from 0.78 in 2022 because conflict-affected schools were excluded), the Dominican Republic (0.68), Thailand (0.69) and Mexico (0.69). The OECD's Figure I.2.8 addresses this by tracking the score attained by the top 25% of the whole 15-year-old cohort for low-coverage countries; those tables (I.B1.2a.48-50) are not folded into the CSV.

5. The socio-economic gap and what is moving it

Table 6. OECD-35 mean score by national ESCS quarter (Tables I.B1.2b.22-24)

	Science bottom Q	Science top Q	Gap	Reading bottom Q	Reading top Q	Gap	Maths bottom Q	Maths top Q	Gap
2015	450.4	536.9	86.6 (0.8)	450.4	535.1	84.7 (0.9)	448.7	531.0	82.2 (0.9)
2018	448.0	534.0	86.1 (0.9)	445.9	533.8	87.9 (0.9)	448.2	534.9	86.7 (0.9)
2022	441.5	538.2	96.7 (0.9)	433.8	527.2	93.4 (0.9)	431.4	525.1	93.7 (0.8)
2025	446.7	530.6	83.8 (1.0)	430.3	506.8	76.5 (1.0)	428.8	511.4	82.5 (0.9)
Change 2022→2025	+5.2 (3.2) n.s.	−7.6 (3.3)	−12.8 (1.3)	−3.5 (1.4)	−20.4 (1.5)	−16.9 (1.3)	−2.6 (1.5) n.s.	−13.8 (1.6)	−11.1 (1.3)
Change 2015→2025	−3.6 (3.5) n.s.	−6.4 (3.6) n.s.	−2.8 (1.3)	−20.2 (3.9)	−28.3 (3.9)	−8.1 (1.4)	−19.9 (3.1)	−19.6 (3.2)	+0.3 (1.3) n.s.

The gap between the top and bottom ESCS quarters narrowed in all three domains between 2022 and 2025, by 12.8 points in science (CI −15.4 to −10.2), 16.9 in reading and 11.1 in mathematics. In each case the arithmetic is dominated by the top quarter: advantaged students lost 7.6 points in science, 20.4 in reading and 13.8 in mathematics, while disadvantaged students did not change significantly in science (+5.2, SE 3.2) or mathematics (−2.6, SE 1.5) and fell 3.5 points in reading. This reproduces the OECD's summary: "the socio-economic gap has been reduced by 13 score points on average, due to declining average scores among advantaged students rather than by improvements among disadvantaged students" (p. 119), and the reading and mathematics figures of −17/−20/−4 and −11/−14 (pp. 121-122). The narrowing is, on this aggregate, a reversal of a widening between 2018 and 2022 (science 86.1→96.7; reading 87.9→93.4; mathematics 86.7→93.7): relative to 2015, the 2025 gap is 2.8 points smaller in science, 8.1 points smaller in reading and unchanged in mathematics, and in reading and mathematics both quarters are roughly 20 points lower than in 2015. Annex A1's figure for reading over 2018→2025 ("from 88 score points in 2018 to 77 points in 2025", −11) corresponds to 87.9→76.5.

Across systems, between 2022 and 2025 the science gap narrowed significantly in 24 and widened in 3 (the Dominican Republic, Guatemala, the Philippines); in 14 of the 24 narrowing cases the top quarter fell significantly and in 11 the bottom quarter rose significantly. Disadvantaged students improved significantly in 22 systems (the text says 21) and declined in 5 (Guatemala, Hong Kong (China), Latvia, Morocco, Paraguay), as the report lists; advantaged students declined in 18 and improved in 12. In reading the gap narrowed in 37 systems (in 28 of them with a significant fall among advantaged students) and widened only in the Philippines; in mathematics it narrowed in 29 (23 with falling advantaged scores) and widened in Brazil, the Philippines and Türkiye. All of these match the OECD's counts on pp. 121-122. One count does not: Annex A1 says the reading gap "narrowed in 36 countries and economies, and widened only in two (Croatia and Zambia)" over 2018→2025; Table I.B1.2b.23 yields 31 significant narrowings and the same two widenings. The next five largest negative changes (Bulgaria −17.3, SE 9.0; Poland −13.5, SE 7.8; Uruguay −13.4, SE 7.7; Greece −11.7, SE 7.4; Guatemala −11.7, SE 10.0) are not significant at the 5% level, so the 36 appears to use a different criterion; the report does not say which.

The share of low performers by ESCS quarter moves the same way. On the OECD average, 37% of disadvantaged and 12% of advantaged students were below Level 2 in science in 2025; in reading the low-performer share among advantaged students rose to 16% (five points above 2022) and among disadvantaged students to 42% (two points above 2022); in mathematics, 48% of disadvantaged students were below Level 2, not significantly different from 2022, while the advantaged share rose four points to 18% (pp. 122 and 129; Tables I.B1.2b.29-30). Among advantaged students the top-performer share fell by 2 points in science, 4 in reading and 3 in mathematics on the OECD-35 average (Table I.2.11 and Tables I.B1.2b.31-33).

6. Gender

Table 7. OECD-35 boys-minus-girls difference in mean score (SE) and change (Tables I.B1.2c.25/28/31)

	Science	Reading	Mathematics
2015	+3.2 (0.6)	−27.3 (0.7)	+7.4 (0.6)
2018	−2.4 (0.6)	−29.7 (0.6)	+5.1 (0.6)
2022	−0.6 (0.6)	−24.8 (0.6)	+8.9 (0.6)
2025	−1.7 (0.6)	−30.2 (0.6)	+12.7 (0.6)
Change 2022→2025	−1.1 (0.8) n.s.	−5.4 (0.9)	+3.8 (0.8)
Change 2015→2025	−4.9 (0.8)	−2.9 (0.9)	+5.2 (0.8)

In 2025 girls lead boys by 30 points in reading and boys lead girls by 13 points in mathematics on the 38-country OECD average (−29.6 and +13.1; Tables I.B1.2c.2-3), while the science difference is 1 point in girls' favour (−1.2), as the chapter summary states. In science the gender gap did not change significantly on average between 2022 and 2025, but the OECD-35 girls' and boys' means both fell slightly (girls 487.1→484.8; boys 486.5→483.1; Tables I.B1.2c.23-24); at the 10th percentile girls scored 12 points above boys and at the 90th boys scored 10 above girls (p. 130). In reading the gap widened in girls' favour by 5.4 points because boys fell more (465.1→448.0, −17.1) than girls (489.9→478.2, −11.7); in mathematics it widened in boys' favour by 3.8 points because girls fell more (470.4→459.2, −11.2) than boys (479.2→471.8, −7.4) (Tables I.B1.2c.26-27, 29-30). Since 2015 the reading gap has moved 2.9 points towards girls and the mathematics gap 5.2 points towards boys, while science moved 4.9 points from a small boys' advantage to a small girls' advantage.

Across systems, between 2022 and 2025 the science gap changed significantly in 12 of 76 (the text: 11 of 75), moving towards girls in Chinese Taipei, Kosovo, Latvia and Mongolia and towards boys in Brunei Darussalam, Cyprus, the Dominican Republic, Hong Kong (China), Iceland, Jordan, Malaysia and the United Arab Emirates; the OECD describes Hong Kong (China) as the only case where the gap "widened in favour of boys" (girls fell 33 points, boys less) and the others as narrowings of a girls' lead. In reading the boys-minus-girls difference became significantly more negative in 19 systems and less negative in 7, matching the text's "widened in favour of girls in 19 … narrowed in 7"; in mathematics it became significantly more positive in 13 systems (the OECD counts these as widening in boys' favour in 5, reversing to boys in 3, and narrowing a girls' lead in others) and more negative in one (Serbia). Girls' reading scores fell significantly in 34 systems and boys' in 48; girls' mathematics scores fell in 36 and boys' in 27 (pp. 133-134).

The 2025 low-performer shares by sex on the OECD average are 27% of boys and 24% of girls in science, 37% of boys and 25% of girls in reading, and 33% of boys and 36% of girls in mathematics (p. 134; Tables I.B1.2c.14-16).

7. Immigrant background

On the OECD-29 aggregate (the 29 members with usable immigrant-background data), the share of students with an immigrant background rose from 12.8% in 2015 to 14.2% in 2018, 15.3% in 2022 and 18.0% in 2025 (+2.7 percentage points 2022→2025, SE 0.3; Table I.B1.2d.1), with first-generation students at 8.0% and second-generation at 10.0% in 2025. Before any adjustment, immigrant students scored 42.7 points (SE 1.2) below non-immigrant students in science in 2025, 41.7 below in reading and 30.5 below in mathematics (Tables I.B1.2d.7-9). Between 2022 and 2025 the unadjusted science gap widened from 37.2 to 42.7 points: non-immigrant students' mean was unchanged (497.3 in both cycles) while immigrant students' mean fell from 460.1 to 454.6. Table I.B1.2d.10, which uses the sub-sample with valid ESCS and language data, gives the change in the science gap as −4.3 points (SE 1.7; significant) before adjustment, −5.4 (SE 1.6; significant) after accounting for ESCS, and −2.7 (SE 1.8; not significant) after accounting for ESCS and language spoken at home. The OECD's statement that disparities "remained unchanged, on average across 29 OECD countries, after accounting for students' socio-economic status and language spoken at home" (p. 144) therefore refers to the fully adjusted gap; the raw gap and the ESCS-adjusted gap both widened significantly.

Annex A1 also quantifies the compositional contribution to the reading trend: with the immigrant share up 3.8 points between 2018 and 2025 and a gap of about 40 points, roughly 1.5 of the 25-point OECD-29 reading decline is attributable to composition (p. 311), and about 3 of Portugal's 30 points. Over 2018→2025 the immigrant/non-immigrant reading gap widened significantly in seven systems with negative overall trends (Argentina, Croatia, Israel, Latvia, Lebanon, North Macedonia, Serbia) and was stable in 22 of the 31 systems meeting the OECD's conditions of at least 5% immigrant students and a negative overall trend (Table I.B1.2d.8). In 2025, 41% of immigrant students and 21% of non-immigrant students were in the bottom ESCS quarter on the OECD-29 average, and after accounting for ESCS and language immigrant students outscored non-immigrants in science in 13 systems, scored lower in 21 and did not differ in 10 (p. 144).

8. Entities the OECD flags, and other checks

The Reader's Guide (pp. 17-21) annotates six countries with an asterisk because sampling standards were not met in 2025. Full reporting with an asterisk: Canada (school response 79% before/82% after replacement, student response 77%; upward bias judged unlikely to exceed 15 points in science), the Netherlands (exclusion rate 9.3%; school response 66%/88%, student 78%; up to about 11 points of possible bias from student non-response), New Zealand (exclusions 8.1%; school response 45%/52%, student 76%; likely upward bias of 10-15 points, "similar in size" to 2022 and "unlikely to affect comparisons between PISA 2022 and PISA 2025"), Norway (exclusions 10.4%, up from 7.3%; possible upward bias). Limited reporting with an asterisk: Albania (school response 76%; schools with fewer than 21 eligible students not contacted; no trend reporting, questionnaire results below the line, plausible lower bound of 428 for the science mean) and the United States (exclusions 6.9%; school response 45%/54%, student 76%; "not possible to rule out bias … nor to determine its most likely direction"; over 40% of students not given the questionnaire). Without an asterisk, Uzbekistan's reading and mathematics results are withheld, Viet Nam is excluded from trend analyses because it moved from PISA 2012-derived paper instruments to computer-based testing (its 2025 reading mean of 391.5 compares with 461.9 in 2022 in the 2022 tables), Guatemala and Paraguay's paper-to-computer transition is flagged for caution, the Palestinian Authority's 2022 results were recalculated to the 2025 coverage, and Lebanon and the Kurdistan Region (Iraq) have restricted coverage. In Spain, within-school exclusions rose to 23.2% in Catalonia (not reported separately) and 12.7% in Murcia. For earlier cycles, the 2022 trend tables asterisked Australia, Canada, Denmark, Hong Kong (China), Ireland, Jamaica, Latvia, the Netherlands, New Zealand, Panama, the United Kingdom, the United States and Viet Nam; 2022 student response rates were below 80% in Australia (76.1%), Canada (77.0%), Hong Kong (75.3%), Ireland (76.8%), Jamaica (67.6%), Malta (79.1%), New Zealand (71.7%), Panama (76.8%), the United Kingdom (75.2%) and the United States (79.9%), and 2022 school response after replacement was below 85% in Hong Kong (79.9%), New Zealand (72.4%), Chinese Taipei (83.8%), the United Kingdom (81.8%) and the United States (63.3%) (Table I.A2.6 of PISA 2022 Volume I). For 2018 the corresponding cases were Hong Kong (78.6% schools), Portugal (76.1% students) and the United States (76.4% schools). Volume I 2025 publishes no country-by-country response-rate table for 2025; it refers to Chapter 14 of the forthcoming Technical Report.

Discrepancies between the OECD's text and its tables found in this exercise are small: the "74" comparable systems in Table I.2.6 exclude Cyprus, whose inclusion raises the count of mathematics declines from 38 to 39 and shifts several "unchanged" counts by one; the "36" reading gap narrowings over 2018-2025 in Annex A1 correspond to 31 significant narrowings in Table I.B1.2b.23; the list of systems that raised their top-performer share in science over 2015-2025 omits Cyprus, Luxembourg and Romania, which are significant in Table I.B1.2a.33; and the "except for science in PISA 2022" qualifier for the lowest-ever OECD average applies to the OECD-23 series, whereas on the OECD-35 series science 2018 is also not significantly different from 2025. Mean scores for 2000-2022 are identical in the 2025 and 2022 workbooks (1,230 overlapping cells, maximum difference 0.0001) except for the Palestinian Authority, but ESCS-quarter means for 2015 and 2018 differ between the two workbooks by up to about 4 points for individual countries (e.g. Moldova's 2018 science gap is 88.2 in the 2025 workbook and 92.0 in the 2022 workbook), which the 2025 volume does not comment on.

Appendix B. Technical note on comparability in the 2025 cycle

IN BRIEF (summary added before publication; see Appendix H)

This technical note summarises what PISA 2025 Results (Volume I) says about the 2025 cycle's design and comparability with earlier cycles, and what it defers to the forthcoming Technical Report. Science was assessed under a revised framework and with adaptive testing; the volume presents no evidence on the effect of either on the science scale. The 2022-to-2025 science link error, calculated from the published tables as 3.1 points, is roughly three times reading's, so a science change of about 6 to 7 points is needed for significance on the OECD-35 average. Annex A1 shows the reading decline in every section and question type, alongside lower self-reported effort and more hasty responses (timed against thresholds reset each year), but this does not by itself distinguish declining skill from declining willingness to engage. Three systems moved from paper to computer with no mode-effect study presented, six missed sampling standards, and no table of 2025 response rates for all systems is published.

This note summarises what *PISA 2025 Results (Volume I)* (OECD, 2026, https://doi.org/10.1787/73451bc5-en) itself says about the design of the 2025 cycle and about the comparability of its results with earlier cycles, drawing on the Reader's Guide (pp. 17-24), "What is PISA?" (pp. 47-51), Annex A1 (pp. 308-320), Annex A2 (pp. 321-325) and Annex A3 (pp. 326-331), and on the StatLink workbooks. For each item it reports what the OECD states, whether Volume I presents evidence for it, and what is deferred to the PISA 2025 Technical Report, which the volume cites throughout as "forthcoming" and which was not available on 9 September 2026. Where a claim is my inference from the tables rather than an OECD statement, it is labelled as such.

Test design and adaptive testing

Volume I describes the 2025 cognitive assessment as a two-hour computer-based test in which "a multi-stage adaptive approach was applied in computer-based tests whereby students were assigned a block of test items based on their performance in the preceding blocks" in science, mathematics and reading (p. 49). Reading has been adaptive since 2018 and mathematics since 2022 (this is from the PISA 2018 and PISA 2022 Technical Reports, not from Volume I 2025); the 2025 volume does not say in so many words that 2025 is the first cycle in which science was adaptive, but Annex A1 describes the science test as having a "linear version, which was used for 25% of test-takers", with three rotated testlets of 11-13 items (p. 318), and Annex A3 states that "for science and mathematics, only students assigned to linear tests are included in the computation of percent-correct scores" (p. 329). The reading design is described in more detail: a first "Core" non-adaptive section followed by two adaptive stages, with items in harder testlets assigned with probabilities 0.1/0.5/0.9 to low-, medium- and high-performing students and the reverse for easier testlets (Annex A3, p. 330). More than 11 hours of items were covered across the four domains (p. 49). Detailed design, routing rules and item counts are deferred to the Technical Report.

The volume presents evidence on one specific worry, that adaptivity could create the appearance of a trend. Annex A1 computes percent-correct scores on the 2018-2025 reading trend items separately for the non-adaptive Core section and for Stages 1 and 2 (Table I.A1.7) and reports that each subset shows the same direction of change as the scale scores, that "the vast majority of declines are clearly visible already in the first, non-adaptive section of the test", and that the correlation between country-level changes in average percent-correct and changes in mean scores is 0.91 for 2022-2025 and 0.90 for 2018-2025 (p. 312). It also reports that the decline was steeper in the later, adaptive stages than in the Core section over 2018-2025 (Table I.A1.9), which it interprets as declining attention rather than a design artefact. This evidence is item-level and descriptive; the volume does not present an analysis of the adaptive routing itself, nor an equivalent analysis for science or mathematics, beyond the science "endurance" comparison of item-set positions in the linear form (Table I.A1.2).

Mode of delivery and platform

Most systems tested on computer, "as was done in 2015, 2018 and 2022", and six systems, Cambodia, Dushanbe (Tajikistan), the Kurdistan Region (Iraq), Mauritius, Rwanda and Zambia, used "a paper-based version of the assessment that included only trend items for science, reading and mathematics" (p. 50). Paper-based systems have no science competency subscales "because they do not have enough cognitive data" (Table I.2.3 notes). The notes to Figure I.2.3 refer readers to "Annex A5" for the paper-based test, but no Annex A5 is included in the published volume (its annexes run A1-A3, B1-B2, C, D, E), so that description is not available in Volume I.

Three systems changed mode between 2022 and 2025. For Viet Nam the OECD states that it "transitioned to computer-based administration in PISA 2025, whereas in all previous participating cycles it used older paper-based instruments derived from PISA 2012. The uncertainty in the trend comparisons beyond linking errors remains undetermined", and it excludes Viet Nam from all performance-trend analyses while retaining it in 2025 cross-sectional and questionnaire-based comparisons (p. 20). Guatemala and Paraguay moved from the newer PISA-for-Development paper instruments (2017 framework) to computer; their trends are reported but "should be interpreted with caution, as the uncertainty extends beyond what is captured by the linking errors" (p. 20). No mode-effect study is presented in Volume I for any of these cases, and no evidence is given on why the 2012-derived instruments are treated differently from the 2017-derived ones beyond the framework vintage.

Separately, the test-delivery platform was changed for 2025 (the acknowledgements name Open Assessment Technologies as "responsible for the PISA 2025 platform", p. 9). The volume mentions this only in two Annex A1 endnotes: "Due to minor differences in screen layout, and to a change in the test-taking platform in 2025, which might have affected the measurement of time on task, the 'minimum time threshold' is defined separately for each year" and timing categories "based on thresholds defined separately by country and year" (p. 321, notes 2-3). This means that the timing-based indicators used in Annex A1 (hasty responses, reading-fluency categories) are defined relative to within-year distributions rather than on a common clock. The volume presents no evidence on whether the platform change affected scale scores; the treatment of the platform in scaling is deferred to the Technical Report.

The science framework revision

Science was the major domain in 2025, as in 2006 and 2015, and was assessed under a new framework developed by the 2025 Science Expert Group chaired by Jonathan Osborne, with a separate Environmental Sciences Expert Group (p. 9; OECD, *PISA 2025 Assessment and Analytical Framework*, https://doi.org/10.1787/86c36975-en). The framework defines three competencies: “explain phenomena scientifically”; “construct and evaluate designs for scientific enquiry and interpret scientific data and evidence critically”; and “research, evaluate and use scientific information for decision making and action” (p. 77). The 2015 framework’s competencies were “explain phenomena scientifically”, “evaluate and design scientific enquiry” and “interpret data and evidence scientifically”; the 2025 framework thus merges the second and third of these and adds a new information-evaluation competency, and adds an environmental-science subscale (Table I.B1.2a.8) reported only for 2025. Proficiency levels run from 1b to 6, as in 2015, with new descriptors (Tables I.2.2 and I.2.o7).

Volume I does not describe how many 2025 science items were carried over from 2015/2018/2022 and how many are new, nor how the new competency was linked to the 2015 scale. It states that trend comparisons in science rest on the same link-error machinery as in other domains (Annex A3; Table I.A3.1) and that science results for the paper-based systems are based on “trend items” only (p. 50), which implies (my inference) that a trend-item set exists that spans the framework change. The text asserts continuity of the science scale (“science has declined more gradually over the past decade”, p. 26) without discussing the framework revision as a comparability issue anywhere in the Reader’s Guide, Chapter 2 or the annexes. Whether the framework change affects the science trend is therefore unaddressed in Volume I and left to the Technical Report. One observable consequence in the tables is that the science link errors are the largest of the three domains for the recent links (see below).

Linking procedure and link errors

The Reader’s Guide explains that link errors “must be used to ensure comparability” across cycles because they represent “uncertainty around scale values”, give the example that “for comparisons between science results in PISA 2025 and science results in 2022, the link error corresponds to 2.9 score points”, and refer to Annex A3 (p. 24). Annex A3 says that standard errors and p-values for multi-cycle estimates “account for the uncertainty around the reporting of PISA 2025 results on scales that were established in previous cycles (the ‘link error’)”, that the methods for quantifying it are in Chapter 15 of the forthcoming Technical Report, and that the methods for incorporating it in comparisons of proportions and in regression-based trends are those of Annex A7 of PISA 2022 Volume I (p. 327). Table I.A3.1 publishes only the link errors used for the “average decennial trend” regressions (points per decade), not the pairwise link errors:

Cycles in the regression	Science	Reading	Mathematics
2015, 2018, 2022, 2025	3.03	3.23	3.00
2015, 2018, 2025	3.55	3.16	2.87
2015, 2022, 2025	2.91	3.93	3.05
2017, 2022, 2025	3.87	4.87	3.81
2018, 2022, 2025	4.74	2.62	3.65

Source: Table I.A3.1 (https://stat.link/p7oc9v).

The pairwise link errors can be recovered from the annex tables because the SE of each country’s change equals the square root of the sum of the two sampling variances and the squared link error; the implied value is identical across countries to two decimals, which confirms the construction. My back-calculation from Tables I.B1.2a.36-38 gives the following link errors for comparisons with 2025: from 2022, science 3.12, reading 1.09, mathematics 1.22; from 2018, 3.51, 1.83, 2.55; from 2015, 3.41, 3.79, 3.00; from 2012, 6.06, 6.11, 3.78; from 2009, 6.69, 4.79, 4.45; from 2006, 4.82, 8.63, 4.27; from 2003, reading 5.36 and mathematics 5.67; from 2000, reading 6.76. Two things follow. First, the science link error for 2022→2025 embedded in the tables (3.1) is not the 2.9 quoted in the Reader’s Guide; the difference is small but the source of the 2.9 figure is not identifiable from the volume. Second, the science link errors for the two most recent comparisons (3.1 and 3.5) are roughly three times those for reading (1.1, 1.8), so a science change of about 6-7 points is needed for significance at the OECD-average level (OECD-35 science 2022→2025: −2.8, SE 3.2), whereas a reading change of 2.5-3 points suffices. This asymmetry is why the OECD can describe science as “stable” while reporting significant declines of similar or smaller absolute size in earlier science cycles, and it is (my inference) consistent with a science scale re-anchored across a framework revision. Volume I does not comment on the size of the science link errors.

Treatment of the 2022 cycle

The volume notes that the eighth cycle “was originally planned for 2021 … postponed … to 2022 because of the many difficulties education systems faced due to the COVID-19 pandemic” (p. 51, note 1). It treats the 2022 results as a normal baseline: every 2022→2025 comparison in the report uses the published 2022 estimates, with the single exception of the Palestinian Authority, whose “PISA 2022 results were recalculated to match the same coverage as PISA 2025” after conflict-affected schools were excluded in 2025 (p. 21); in the 2025 tables its 2022 science mean is 375.7 against 368.8 in the 2022 tables, reading 355.7 against 349.2 and mathematics 372.1 against 365.8. No other 2022 estimate was revised (the 2025 and 2022 workbooks agree to four decimals on 1,230 overlapping country-cycle-domain means).

Systems whose 2022 samples fell short of standards (the 2022 volume asterisked Australia, Canada, Denmark, Hong Kong (China), Ireland, Jamaica, Latvia, the Netherlands, New Zealand, Panama, the United Kingdom and the United States) are compared with 2025 without adjustment and without a 2022 asterisk in the 2025 tables; the 2025 Reader's Guide mentions the 2022 shortfall only for New Zealand ("when student participation was even lower, at 72%") and judges that the bias "is unlikely to affect comparisons between PISA 2022 and PISA 2025 in meaningful ways" (p. 18), and for the Netherlands and the United States it reports that exclusion rates "remained close" to 2022 (pp. 18-19). The Ukrainian sample changed from 18 to 17 regions (p. 21) and no 2022 value is carried in the 2025 trend tables for Ukraine. The volume does not discuss whether pandemic-era conditions in 2022 affect the 2022→2025 comparison as a measurement matter; its discussion of the pandemic is substantive (school closures, attendance, Box I.2.3 and Annex A1, pp. 96-99 and 316), and it points out that students assessed in 2025 were "around 10 or 11 years old at the onset of the pandemic" (p. 100) and that the correlation of PIRLS 2016→2021 changes with PISA 2022→2025 changes is 0.34 across 23 systems (p. 311). The OECD does not, in Volume I, present any analysis that treats the 2022 samples as anomalous or re-weights them.

Sampling standards, exclusions and response rates in 2025

The Reader's Guide sets out three standards (overall exclusion rate ≤ 5%; weighted school response ≥ 85% before replacement or an acceptable rate after replacement; weighted student response ≥ 80%) and lists the entities that failed at least one (pp. 17-20), with the treatments recommended by the PISA Adjudication Group in February 2026, adding that "there may be a need for subsequent adjustments as new evidence on the quality and comparability of the data emerges" (p. 17). The quantitative statements are: Canada, school response 79%/82% and student 77%, upward bias "unlikely to exceed 15 points in science (±15.0% of a standard deviation)"; the Netherlands, exclusion 9.3%, school response 66%/88%, student 78%, possible student non-response bias "up to ±9.9% of a standard deviation (i.e. about 11 score points in science)"; New Zealand, exclusion 8.1%, school response 45%/52%, student 76%, mean scores "may be over-estimated by about 10 to 15 score points"; Norway, exclusion 10.4%; Albania, school response 76%, small schools not contacted, lower bound for science mean 428; the United States, exclusion 6.9%, school response 45%/54%, student 76%, range-of-bias "up to about ±15.4% of a standard deviation" for schools and "±15.9%" for students with direction undeterminable, plus over 40% of respondents not given the questionnaire. Evidence is described (non-response bias analyses submitted by the national centres, using national test data in New Zealand's case and school-level data in the Netherlands' case; OECD "range-of-bias" simulations) but not tabulated; the volume publishes no table of 2025 exclusion rates or response rates for all systems, referring instead to Chapter 14 of the Technical Report (Annex A2, pp. 323-324). Spain's within-school exclusions rose to 23.2% in Catalonia (not reported) and 12.7% in Murcia (reported with asterisk) (p. 19). Coverage Index 3 for every cycle since 2003 is published (Table I.A2.1) and is in the CSV; the 2025 index falls below 0.60 for Rwanda (0.32), Cambodia (0.43), the Palestinian Authority (0.54) and Guatemala (0.55).

Uzbekistan's mathematics and reading response data showed "inconsistencies … more frequent in some schools and regions" that "undermine the comparability of the scale scores"; only science and questionnaire results are reported (p. 20). The volume does not say what the inconsistencies were.

Test engagement, timing and the reading trend

Annex A1 is the volume's main body of evidence that the 2018-2025 reading decline "is unlikely to be a measurement artefact" (p. 319). It reports that the 2025 reading test reused the 2018 tasks so that comparisons are "not confounded by differences in test design and task content" (p. 312); that item-level percent-correct changes track scale-score changes (r = 0.90-0.91); that declines are larger on longer texts and simple multiple-choice items and, over 2018-2022, on evaluating/reflecting items (Tables I.A1.8-10); that "hasty" incorrect responses rose by over 4 percentage points to about 9% while non-reached items fell by about 2 points (Table I.A1.12); that "hasty" test-takers on the three-minute fluency task rose from 6.6% to 11.4% and accurate fluent readers fell by 7 points, "most of this increase" over 2022-2025 (Table I.A1.15); that self-reported effort on the test fell by 0.5 points on a 10-point scale relative to 2018 (Table I.A1.1); that first-hour versus second-hour differences did not widen (Tables I.A1.3-5); that fatigue within the linear science test is modest (Table I.A1.2); and that questionnaire straightlining fell rather than rose (Table I.A1.17). The OECD's reading of this evidence is that declining "capacity to direct and … sustain attention" is "part of what the 'reading' test in PISA assesses" (p. 319), so it counts the behavioural changes as real changes in the construct rather than as invalidity. That is an interpretive position; the evidence presented supports the narrower claim that the score decline is present in every section and item type of the test and correlates with behavioural indicators, and does not by itself distinguish declining skill from declining willingness to engage with a low-stakes test. Because timing thresholds are set per year and per country after the platform change (p. 321), the hasty-response and fluency trends depend on that normalisation; the volume does not report a sensitivity check.

Other items the OECD flags

The OECD average is not a fixed set: the 38-member "OECD average" is used for 2025 cross-sections, "OECD average-35" for trends since 2015 and "OECD average-23" for trends since 2000 (p. 21); the OECD-38 trend rows are blank for 2018 and 2022. PISA 2000+ (2001/2002), PISA 2009+ (2010) and PISA-for-Development (2017) results are tabulated under 2000, 2009 and 2018 (table notes). The report's "20 score points equals one year of learning" benchmark is re-estimated in Box I.2.2 from 2015-2025 data for ten systems, with country estimates from about 10 (Serbia) to 28 (Scotland) points (Table I.2.7), and the OECD warns against using a single conversion. The database underlying the report is the June 2026 version; updates to the public-use file released on 8 September

2026 are documented only in the Technical Report (p. 22). The indices of curiosity and perseverance are stated to be “not comparable between PISA 2022 and PISA 2025” because of changes in the questionnaire (p. 68), which affects the volume’s engagement trends but not the performance trends.

Summary of what is and is not evidenced in Volume I

Evidenced with tables: link errors for regression trends (Table I.A3.1) and, implicitly, pairwise link errors (recoverable from the SEs); consistency of the reading trend across test sections, item types and behavioural indicators (Annex A1 tables); coverage indices for all cycles (Table I.A2.1); the direction and rough size of sampling bias for the six asterisked systems (Reader’s Guide text, drawing on unpublished bias analyses). Asserted but not evidenced in Volume I: the effect of the science framework revision and of the new adaptive science design on the science scale; the effect of the platform change on scores and timing; the specific comparability problem in Viet Nam; the nature of Uzbekistan’s data inconsistencies; 2025 exclusion and response rates for systems that met the standards. All of these are deferred to the PISA 2025 Technical Report (Chapters 14 and 15 are cited), and the Adjudication Group’s February 2026 assessments are described as provisional.

Appendix C. Country trajectories 2012–2025

> **IN BRIEF** (summary added before publication; see Appendix H)
>
> This section classifies 69 countries and economies by the shape of their mean-score trajectory over 2012-2025, then asks whether documented features distinguish the groups. Eleven systems show a long decline and eight a decline since 2018, almost all OECD members; the eight rising systems are all non-OECD, seven starting below 460 points in 2018. Of sixteen systems that rose in some period after 2018, only the Philippines, Georgia and the United Arab Emirates rose significantly in all three subjects, consistently across periods, with stable or widening coverage of 15-year-olds. None of the features examined, including spending, immigrant share, school closures, governance and phone policy (the most-discussed bans took effect after the 2025 test), significantly separates the groups once income is accounted for, or within the OECD. The one significant group difference, in weeks of full school closure, runs opposite to the usual hypothesis and disappears when GDP per capita is controlled. All of this is observational and cannot establish cause.

This section classifies every country and economy with valid trend data in `pisa_trends.csv` by the shape of its mean-score trajectory over the last three PISA intervals, reports the resulting groups, and then asks whether publicly documented system features discriminate between the groups. All score data are from the StatLink workbooks of *PISA 2025 Results (Volume I)* (OECD, 2026, https://doi.org/10.1787/73451bc5-en; Tables I.B1.2a.36-38 for means and changes, Table I.A2.1 for coverage) and *PISA 2022 Results (Volume I)* (Tables I.B1.5.4-6), as assembled in the dataset described in `data_dictionary.md`. The classification is produced by `classify_trajectories.py`; the feature table by `build_features.py` and `hand_coded_features.csv`; the statistics in section 5 by `analyze_features.py`. Throughout, “significant” means that the change divided by its standard error, with the link error included in the standard error, exceeds 1.96 in absolute value, which is the rule the OECD uses for bold values in its own tables.

1. Rules used for the classification

The rules are stated so that they can be checked or changed.

Entities. Every non-aggregate country or economy for which the OECD reports comparable values in the 2025 trend tables (`trend_comparable_2025 = 1`). This excludes Albania (OECD: no trend reporting) and Viet Nam (OECD: excluded from trend analyses because of the move from 2012-derived paper instruments to computer-based testing). It includes the six entities the OECD asterisks in 2025 (Canada, the Netherlands, New Zealand, Norway, the United States, and Albania which is excluded anyway) and the entities asterisked in 2022, without adjustment, because the OECD compares them without adjustment; the asterisks are carried in the output and discussed where they matter.

Windows. Three consecutive intervals, W1 = 2012 to 2018, W2 = 2018 to 2022 and W3 = 2022 to 2025, plus the long window L = 2018 to 2025 which spans the pandemic and its aftermath.

Change and its standard error. For each domain, change = mean(later cycle) minus mean(earlier cycle), and SE = sqrt(SE(later)^2 + SE(earlier)^2 + link error^2). For W3 and L the OECD’s own change and SE columns are used directly (they embed the link error). For W2 the link error was recovered from the SEs in PISA 2022 Volume I Tables I.B1.5.4-6, which give the 2018-2022 change and its SE for every country; the back-calculated link error is identical across countries to three decimals (reading 1.470, mathematics 2.240, science 1.610), which confirms the construction. For W1 the link errors are those published in PISA 2018 Volume I, Table I.A7.1, “PISA 2012 to 2018” row (reading 3.74, mathematics 3.34, science 4.01; https://doi.org/10.1787/888934028957).

Per-domain verdict. In each window and domain, “rise” if change/SE > 1.96, “fall” if change/SE < -1.96, otherwise “stable”.

Combined verdict for a window. At least two domains must be available. “Fall” if at least two domains fall significantly and none rises; “rise” if at least two rise and none falls; “mixed” if at least one domain rises and one falls; “stable” otherwise, that is, if no domain moves or only one of three moves. The last clause is deliberate: with three tests at the 5 per cent level, a single significant domain is weak evidence of a system-level movement, and it also keeps the science domain, whose 2022-2025 link error is about three times that of reading (see `technical_note.md`), from driving a verdict on its own.

Trajectory group. The sequence (W1, W2, W3) and the long-window verdict L are read in this fixed order, first match wins: (F) W2 fall and W3 rise, “rebound”; (G) W3 rise with W2 not a fall, or W2 rise with W3 not a fall, “rising”, split into G (L is also a rise) and G2 (L is not a rise); (E) a rise in W1 or W2 followed by a fall in W3, “rise then fall”; (A) W1 fall and a fall in W2 or W3 with no rise anywhere, “long decline”; (B) W2 fall and W3 fall, “decline since 2018”; (C) W2 fall and W3 stable or mixed, “fall then stabilised”; (D) W3 fall with W2 not a fall, “recent fall only”; (D2) no fall or rise in W2 or W3 but L is a fall, “gradual decline”; (A2) W1 fall with W2 and W3 both stable or mixed, “early fall then stable”; (S) everything else, “stable”. Entities lacking W2 or W3 are placed in H (“partial”) with the verdict from whichever window exists.

Two consequences follow. Group C contains systems whose 2022-2025 verdict is “stable” although one domain fell significantly (Estonia: reading -12, SE 3.6; mathematics and science unchanged). And because the threshold is a fixed z, systems with large SEs (the United States, about 7 points on 2018-2025 changes) are harder to move out of “stable” than systems with SEs of 3.

2. The groups

Of 89 entities with at least some comparable data, 69 have both W2 and W3 and can be classified; 20 are partial. The distribution of verdicts by window (Table 1) reproduces the counts in `factual_summary.md` for the windows that end in 2025 and adds the 2012-2018 and 2018-2022 windows computed here.

Table 1. Number of systems by combined verdict, by window

Window	n	Fall	Stable	Rise	Mixed
W1 2012-2018	57	18	30	7	2
W2 2018-2022	69	31	23	8	7
W3 2022-2025	75	33	33	8	1
L 2018-2025	74	49	8	9	8

Source: `trajectory_groups.csv`. Per-domain counts: in W3 reading fell in 42 of 75 systems and rose in 8; mathematics fell in 39 and rose in 8; science fell in 11 and rose in 16 (Tables I.B1.2a.36-38). In W2 mathematics fell in 40 of 70 and reading in 34 of 69 (Tables I.B1.5.4-6 of PISA 2022 Volume I).

Table 2. Trajectory groups

Group	Rule (W1, W2, W3; L)	n (OECD)	Members
A. Long decline	fall in 2012-18 and in a later window, no rise	11 (9)	Australia, Bulgaria, Costa Rica, Finland, Germany, Greece, Hong Kong (China), Hungary, Latvia, Netherlands, Switzerland
A2. Early fall, then stable	fall 2012-18, stable/mixed after	2 (2)	Italy, Korea
B. Decline since 2018	stable 2012-18 (or n.a.), fall 2018-22, fall 2022-25	8 (7)	Belgium, Canada, Denmark, France, Iceland, Morocco, Norway, Portugal
C. Fall, then stabilised	fall 2018-22, stable/mixed 2022-25	12 (5)	Cyprus, Estonia, Indonesia, Kosovo, Macao (China), Malaysia, Mexico, Moldova, North Macedonia, Poland, Slovak Republic, United Kingdom
D. Recent fall only	no fall 2018-22, fall 2022-25	10 (5)	Argentina, Austria, Chile, Croatia, Czechia, Ireland, Israel, Malta, Serbia, Singapore
D2. Gradual decline	no window falls, but 2018-25 falls	4 (4)	Colombia, Lithuania, New Zealand, United States
E. Rise, then fall	rise 2012-18 or 2018-22, fall 2022-25	5 (2)	Guatemala, Paraguay, Peru, Slovenia, Sweden
F. Rebound	fall 2018-22, rise 2022-25	2 (0)	Montenegro, Thailand
G. Rising	rise in 2018-22 or 2022-25, no fall, and 2018-25 rise	8 (0)	Brunei Darussalam, Cambodia, Chinese Taipei, Dominican Republic, Georgia, Philippines, Qatar, United Arab Emirates
G2. Rise not sustained	as G but 2018-25 not a rise	2 (2)	Japan, Türkiye
S. Stable	none of the above	5 (0)	Brazil, Kazakhstan, Romania, Saudi Arabia, Uruguay
H. Partial (2018-25 only)		5	fall: Ecuador, Lebanon, Luxembourg; mixed: B-S-J-Z (China); rise: Zambia
H. Partial (2022-25 only)		6	fall: Palestinian Authority, Spain; rise: Jordan; stable: El Salvador, Mongolia, Ukrainian regions (17 of 27)
H. Insufficient		9	Armenia, Azerbaijan, Dushanbe (Tajikistan), Kenya, Kurdistan Region (Iraq), Kyrgyzstan, Mauritius, Rwanda, Uzbekistan

Source: `trajectory_groups.csv`, which carries every per-domain change, SE and verdict behind the table.

Some membership decisions are worth spelling out because they are sensitive to the rules. Poland and the United Kingdom are in C, not A or B, because their 2022-2025 changes are not significant in two domains (Poland reading -7.4, SE 4.1; mathematics -5.4, SE 3.8; science -4.1, SE 4.9; the United Kingdom -0.5, -1.4 and +11.7, the last significant). Estonia is in C for the same reason despite a significant reading fall. Korea is in A2 because its large 2012-2015 losses (reading -19, mathematics -30) were followed by no significant window change since 2018 in more than one domain, although reading fell 15 points (SE 5.1) between 2022 and 2025 and its 2018-2025 reading change is also significant (-13, SE 4.9). Japan is in G2 because it rose in 2018-2022 in reading and science (+12 and +17.5) and then fell back in 2022-2025 in all three domains (-13.1, -10.2, -9.0, of which reading is significant), leaving its 2018-2025 changes at zero. Türkiye is in G2 because its 2022-2025 rise is significant in all three domains but the 2018-2025 change is significant only in science (+26.0, SE 4.7), reading and mathematics having fallen in 2018-2022 before recovering. Sweden is in E because its 2012-2018 recovery (+23 reading, +24 mathematics, +14 science) was followed by two falling windows. The United States is in D2: none of its window changes is significant (2022-2025 reading -13.8 with SE 7.3 is just below the threshold), but its 2018-2025 changes in reading (-15.6, SE 7.1) and mathematics (-15.0, SE 6.8) are.

The OECD membership split is the first pattern to note. Groups A, A2, B, D2 and G2 are almost entirely OECD; groups G, F and S are entirely non-OECD; C and D are mixed. That is partly a product of the OECD's own composition (richer systems, more immigration, longer time series) and partly the substantive finding that the systems that rose after 2018 are, with the exceptions of Chinese Taipei, Qatar and Brunei, low- or middle-performing systems starting from means below 460.

3. The 2000-2025 series for frequently discussed systems

Tables 3-5 give the full series for the systems named in the brief. Values in the 2025 trend tables are shown as published; "m" marks a cycle in which the country participated but the OECD reports no comparable value; a dash marks non-participation; bold marks a significant change. SEs of changes include the link error.

Table 3. Reading

Country	2000	2003	2006	2009	2012	2015	2018	2022	2025	2018→2025 (SE)	2022→2025 (SE)
Finland	546	543	547	536	524	526	520	490	474	**-46** (3.6)	**-16** (3.3)
Sweden	516	514	507	497	483	500	506	487	466	**-39** (4.7)	**-21** (4.2)
Poland	479	497	508	500	518	506	512	489	482	**-30** (4.3)	-7 (4.1)
Estonia	–	–	501	501	516	519	523	511	499	**-24** (3.6)	**-12** (3.6)
Germany	484	491	495	497	508	509	498	480	465	**-33** (5.0)	**-15** (5.2)
Netherlands*	–	513	507	508	511	503	485	459	441	**-43** (4.3)	**-18** (5.3)
United Kingdom	–	m	495	494	499	498	504	494	494	**-10** (3.8)	0 (3.3)
Ireland	527	515	517	496	523	521	518	516	500	**-18** (3.5)	**-16** (3.3)
Singapore	–	–	–	526	542	535	549	543	535	**-15** (3.1)	**-8** (2.9)
Japan	522	498	498	520	538	516	504	516	503	-1 (5.0)	**-13** (5.1)
Korea	525	534	556	539	536	517	514	515	501	**-13** (4.9)	**-15** (5.1)
Chinese Taipei	–	–	496	495	523	497	503	515	508	+5 (4.5)	-7 (4.5)
Australia	528	525	513	515	512	503	503	498	491	**-12** (3.1)	**-7** (2.9)
Canada*	534	528	527	524	523	527	520	507	490	**-30** (3.1)	**-17** (2.9)
United States*	504	495	m	500	498	497	505	504	490	**-16** (7.1)	-14 (7.3)
Türkiye	–	441	447	464	475	428	466	456	472	+6 (3.7)	**+16** (3.2)
B-S-J-Z (China)	–	–	–	–	–	–	555	–	527	**-29** (3.9)	–
Hong Kong (China)	525	510	536	533	545	527	524	500	480	**-44** (4.6)	**-20** (4.4)
Macao (China)	–	498	492	487	509	509	525	510	501	**-24** (2.5)	**-9** (2.1)

Table 4. Mathematics

Country	2000	2003	2006	2009	2012	2015	2018	2022	2025	2018→2025 (SE)	2022→2025 (SE)
Finland	–	544	548	541	519	511	507	484	469	**-39** (3.9)	**-15** (3.1)
Sweden	–	509	502	494	478	494	502	482	464	**-38** (4.5)	**-18** (3.5)
Poland	–	490	495	495	518	504	516	489	484	**-32** (4.6)	-5 (3.8)
Estonia	–	–	515	512	521	520	523	510	508	**-16** (3.8)	-2 (3.3)
Germany	–	503	504	513	514	506	500	475	464	**-36** (4.8)	**-11** (4.5)
Netherlands*	–	538	531	526	523	512	519	493	483	**-36** (4.7)	-9 (4.9)
United Kingdom	–	m	495	492	494	492	502	489	488	**-14** (4.2)	-1 (3.4)
Ireland	–	503	501	487	501	504	500	492	480	**-20** (3.8)	**-12** (3.0)

Country	2000	2003	2006	2009	2012	2015	2018	2022	2025	2018→2025 (SE)	2022→2025 (SE)
Singapore	–	–	–	562	573	564	569	575	563	-6 (3.4)	**-12** (2.3)
Japan	–	534	523	529	536	532	527	536	525	-2 (5.5)	-10 (5.3)
Korea	–	542	547	546	554	524	526	527	522	-4 (5.6)	-5 (5.7)
Chinese Taipei	–	–	549	543	560	542	531	547	546	**+15** (5.0)	-1 (5.0)
Australia	–	524	520	514	504	494	491	487	478	**-14** (3.7)	**-10** (2.9)
Canada*	–	532	527	527	518	516	512	497	485	**-27** (4.0)	**-12** (2.8)
United States*	–	483	474	487	481	470	478	465	463	**-15** (6.8)	-2 (6.8)
Türkiye	–	423	424	445	448	420	454	453	462	+8 (4.2)	**+8** (3.1)
B-S-J-Z (China)	–	–	–	–	–	–	591	–	612	**+21** (4.4)	–
Hong Kong (China)	–	550	547	555	561	548	551	540	522	**-29** (4.9)	**-19** (4.3)
Macao (China)	–	527	525	525	538	544	558	552	549	**-8** (3.2)	-3 (2.0)

Table 5. Science

Country	2006	2009	2012	2015	2018	2022	2025	2018→2025 (SE)	2022→2025 (SE)
Finland	563	554	545	531	522	511	504	**-18** (4.9)	-7 (4.6)
Sweden	503	495	485	493	499	494	485	**-14** (5.5)	-8 (4.9)
Poland	498	508	526	501	511	499	495	**-16** (5.2)	-4 (4.9)
Estonia	531	528	541	534	530	526	527	-3 (4.6)	+1 (4.4)
Germany	516	520	524	509	503	492	486	**-17** (5.8)	-7 (5.9)
Netherlands*	525	522	522	509	503	488	485	**-18** (5.3)	-3 (5.9)
United Kingdom	515	514	514	509	505	500	511	+7 (4.8)	**+12** (4.4)
Ireland	508	508	522	503	496	504	500	+3 (4.7)	-4 (4.4)
Singapore	–	542	551	556	551	561	560	**+9** (4.2)	-2 (3.8)
Japan	531	539	547	538	529	547	538	+8 (5.8)	-9 (5.7)
Korea	522	538	538	516	519	528	526	+7 (5.6)	-2 (5.8)
Chinese Taipei	532	520	523	532	516	537	540	**+24** (5.4)	+2 (5.5)
Australia	527	527	521	510	503	507	509	+6 (4.4)	+2 (4.1)
Canada*	534	529	525	528	518	515	510	-8 (4.5)	-5 (4.1)
United States*	489	502	497	496	502	499	502	-1 (7.6)	+2 (7.9)
Türkiye	424	454	463	425	468	476	494	**+26** (4.7)	**+18** (4.3)
B-S-J-Z (China)	–	–	–	–	590	–	597	+7 (4.9)	–
Hong Kong (China)	542	549	555	523	517	520	496	**-21** (5.3)	**-25** (5.2)
Macao (China)	511	511	521	529	544	543	541	-3 (4.0)	-2 (3.5)

Source for Tables 3-5: Tables I.B1.2a.36-38, PISA 2025 Volume I. Asterisks mark systems the OECD annotates in 2025.

Comparability flags, from the 2025 Reader’s Guide (pp. 17-21), the 2022 trend-table annotations and Table I.A2.1 (all carried in the dataset):

Finland, Sweden, Poland, Estonia, Germany, Singapore, Japan, Korea, Chinese Taipei, Türkiye and Macao carry no OECD annotation in 2025 or 2022. Their Coverage Index 3 in 2025 is 0.93, 0.92, 0.89, 0.89, 0.93, 0.95, 0.86, 0.93, 0.92, 0.72 and 0.93. Two of these deserve a note. Japan’s coverage fell from 0.92 in 2022 to 0.86 in 2025, the lowest in its series, which is not discussed in Volume I; the direction of any effect on Japan’s 2025 means is not determinable from the published tables. Türkiye’s coverage has risen from 0.36 (2003) to 0.72-0.74 (2018-2025), so its long-run gains between 2003 and 2012 were achieved while the tested population was broadening, which usually works against a rising mean; over 2018-2025 coverage is flat (0.73, 0.74, 0.72), so the recent rise is not a coverage artefact. Chinese Taipei’s 2022 school response rate was 82.6 per cent before and 83.8 per cent after replacement, below the 85 per cent standard (Table I.A2.6, PISA 2022 Volume I), which the 2022 volume noted without an asterisk.

The Netherlands is asterisked in both 2022 and 2025: in 2025 the exclusion rate was 9.3 per cent (8.4 in 2022), school response 66 per cent before and 88 per cent after replacement, student response 78 per cent, with possible non-response bias of “up to about 11 score points in science”; its coverage index fell from 0.91 (2018) to 0.79 (2022) and 0.84 (2025). Canada is asterisked in both years (2025: school response 79/82 per cent, student 77 per cent, upward bias “unlikely to exceed 15 points”), while its coverage index rose from 0.86 (2018) to 0.96 (2025). The United States is asterisked with limited reporting in 2025 (school response 45/54 per cent, student 76 per cent; the OECD cannot determine the direction of bias) and was asterisked in 2022 (school response 51/63 per cent); its coverage is stable at 0.85-0.86. The United Kingdom was asterisked in 2022 (student response 75.2 per cent, school response 81.8 per cent after replacement) but not in 2025; its coverage index moved from 0.85 (2018) to 0.97 (2022) and back to 0.85 (2025), an unusually large swing that the volume does not comment on. Ireland and Australia were asterisked in 2022 (student response 76.8 and 76.1 per cent) but not in 2025; Australia’s coverage fell to 0.83 in 2025 from 0.90. Hong Kong was asterisked in 2022 (school response 59.6/79.9 per cent, student 75.3 per cent) and not in 2025; its coverage index moved from 0.98 (2018) to 0.81 (2022) and 0.90 (2025). B-S-J-Z (China) has only 2018 and 2025 data in the trend tables (no 2022 participation), coverage 0.81 and 0.83, no annotation; its 2018-2025 change is a 29-point significant fall in reading, a 21-point significant rise in mathematics, and no change in science, a divergence the OECD does not explain in Volume I and which this section cannot resolve.

The earlier cycles reported as “m” (the United Kingdom 2003, the United States reading 2006, Spain reading 2018) are excluded by the OECD from comparisons; Volume I 2025 does not restate the reasons.

4. Systems that rose while most fell, and whether the rises are real

Sixteen systems rose significantly in at least two domains in some window after 2018 (groups F, G, G2, plus Jordan and Zambia). Because rises against a general decline attract attention, each was checked for three things: significance in more than one domain, consistency across windows, and stability of the coverage index (Table I.A2.1) and sampling notes over the cycles concerned.

Table 6. Rising systems: changes (SE) and coverage index

System	Group	Cov. 2018 / 2022 / 2025	2018→2025 R / M / S	2022→2025 R / M / S	Assessment
Cambodia	G	0.28 / 0.36 / 0.43	+26 / +41 / +52 (all sig.)	+18 / +29 / +35 (all sig.)	Rise while coverage widened; PISA-D 2017 paper baseline; 2025 on paper trend items; means 347-382. Real within its (narrow) covered population, but the covered population is less than half of 15-year-olds.
Philippines	G	0.68 / 0.83 / 0.82	+27 / +18 / +16 (all sig.)	+20 / +16 / +17 (all sig.)	Rise while coverage widened by 14 points; consistent in all domains and both windows. The most robust rise in the set.
Georgia	G	0.83 / 0.86 / 0.93	+4 / +19 / +39	+10 / +26 / +38 (all sig.)	Rise while coverage rose; concentrated in 2022-2025, largest in science (where the link error is largest).
United Arab Emirates	G	0.92 / 0.94 / 0.98	+1 / +18 / +24	+16 / +22 / +26 (all sig.)	Coverage rose; 2018-2022 fell in reading (-14), so the reading level in 2025 is back to 2018. Immigrant share 57 per cent.
Chinese Taipei	G	0.92 / 0.93 / 0.92	+5 / +15 / +24	-7 / -1 / +2	All of the rise occurred 2018-2022; 2022-2025 flat. Coverage stable. 2022 school response below standard (83.8 per cent).
Qatar	G	0.92 / 0.94 / 0.94	+11 / +2 / +15	-1 / +2 / +2	Rise occurred 2018-2022 (reading +12, science +13); 2022-2025 flat. Coverage stable.
Brunei Darussalam	G	0.97 / 0.98 / 0.99	+18 / +5 / +8	-4 / -7 / -6 (M sig.; R z = -1.93)	Rose 2018-2022 in all domains; gave back part of the gain 2022-2025 (mathematics significantly, reading and science just short of the threshold), so W3 is “stable” by the rule. Coverage stable.
Dominican Republic	G	0.73 / 0.64 / 0.68	+4 / +14 / +25	-6 / 0 / 0	Rise occurred 2018-2022 while coverage fell 9 points; 2022-2025 flat. The 2018-2022 rise coincides with a narrower tested population and should not be read as a gain.
Türkiye	G2	0.73 / 0.74 / 0.72	+6 / +8 / +26	+16 / +8 / +18 (all sig.)	Coverage flat; 2022-2025 rise in all domains; 2018-2025 significant only in science. Real but recent.
Japan	G2	0.91 / 0.92 / 0.86	-1 / -2 / +8	-13 / -10 / -9 (R sig.)	2018-2022 rise reversed in 2022-2025; coverage fell to 0.86. Not a rising system over 2018-2025.
Montenegro	F	0.95 / 0.93 / 0.94	-3 / -19 / +20	+13 / +5 / +32 (all sig.)	Rebound after a 2018-2022 fall (-16, -24, -12); mathematics still 19 below 2018. Coverage stable.
Thailand	F	0.72 / 0.75 / 0.69	-1 / -12 / +6	+13 / +13 / +23 (all sig.)	Rebound after a 2018-2022 fall (-14, -25, -17); coverage fell 6 points in 2025, which could contribute. Mathematics still 12 below 2018.
Jordan	H	0.88 / 0.94 / 0.83	n.a. (reading, science 2018 “m”)	+20 / +26 / +32 (all sig.)	Coverage fell 11 points between 2022 and 2025; mathematics fell 38 points 2018-2022. The 2022-2025 rise coincides with a narrower population; treat with caution.
Zambia	H	0.36 / – / 1.00	+50 / +55 / +28 (all sig.)	–	PISA-D 2017 baseline; coverage index 0.36 in 2018 and 1.002 in 2025 (as published), which implies a change in the population definition or the denominator; the comparison is not interpretable without the Technical Report.
Uzbekistan	H	–	science only	+83 (SE 4.6)	Reading and mathematics withheld for data inconsistencies; the science rise is from a 2022 baseline only and is not corroborated by other domains.

Source: `trajectory_groups.csv`; coverage from Table I.A2.1; annotations from the Reader's Guide.

Brunei illustrates the sensitivity of the rules to the threshold: its 2022-2025 reading change (-3.55, SE 1.84, z = -1.93) and science change (-6.40, SE 3.48, z = -1.84) fall just short of significance, so it stays in G; with a z of 1.9 it would be in E. Readers should treat it as a 2018-2022 riser that has since given back part of the gain.

The summary is that only three of the sixteen (the Philippines, Georgia, the United Arab Emirates) show rises that are significant in all three domains, consistent across windows, and accompanied by stable or widening coverage; Türkiye and Montenegro show real but recent (2022-2025) rises against flat coverage; Cambodia's rise is internally consistent but its covered population is 43 per cent of the cohort; Chinese Taipei and Qatar rose in 2018-2022 and have since been flat; the Dominican Republic's, Thailand's and Jordan's rises coincide with falls in coverage that the OECD does not adjust for; Japan's 2018-2022 rise has reversed; Zambia and Uzbekistan cannot be assessed. B-S-J-Z (China) is not a riser: its mathematics gain is significant but its reading loss is larger.

5. What the groups have in common, and what discriminates between them

`system_features.csv` assembles, for every classified entity, GDP per capita (PISA 2022 Volume I Table B3.2.1 for 2021 and World Bank WDI NY.GDP.PCAP.PP.CD for 2023), cumulative expenditure per student from age 6 to 15 and annual expenditure per lower-secondary student (Table B3.2.2, 2019 USD PPP), the share of students with an immigrant background in 2018, 2022 and 2025 (Table I.B1.2d.1, PISA 2025 Volume I), the age of first stratification in the system (Table B3.1.4; "not applicable" coded as 16), weeks of full and partial school closure between February 2020 and March 2022 (UNESCO global monitoring of school closures database, March 2022 release, https://covid19.uis.unesco.org/global-monitoring-school-closures-covid19/; full weeks are days with status "Closed due to COVID-19" divided by seven, partial weeks are days "Partially open" divided by seven; academic breaks excluded), the percentage of decisions taken at the school level and at the central level in public lower secondary education in 2017 (OECD, *Education at a Glance 2018*, Table D6.1, https://doi.org/10.1787/888933805876; 35 systems only), and a coding of national school phone policy as of the PISA 2025 test window. The last is the weakest feature: the UNESCO GEM Report and PEER country profiles could not be retrieved programmatically (the site returns a JavaScript challenge), so the coding rests on the GEM Report team's January 2025 update (https://world-education-blog.org/2025/01/23/to-ban-or-not-to-ban-monitoring-countries-regulations-on-smartphone-use-in-school/, which reports 79 systems with bans by the end of 2024), the European School Education Platform's 2025 country table (https://school-education.ec.europa.eu/en/discover/news/exploring-mobile-phone-use-schools), the House of Lords Library's international review (https://lordslibrary.parliament.uk/smartphones-in-schools-practice-policy-and-international-perspectives/), and national sources for individual countries; every coding carries its source in the file, and 23 of the 69 classified entities are "not coded" (24 counting Mexico, coded only as "partial"). The UNESCO closure database does not cover Hong Kong, Macao, Chinese Taipei or Kosovo.

Table 7. Group medians of system features (number of entities with data in brackets)

Group	n	GDP pc 2023	Cum. spend 6-15	Immigrant % 2025	Tracking age	Full closure wks	Partial wks	% decisions at school
A. Long decline	11	63,467 [11]	119,584 [9]	14 [10]	14 [11]	15 [10]	24 [10]	23 [8]
A2. Early fall	2	58,890 [2]	125,120 [2]	7 [2]	14 [2]	12 [2]	46 [2]	22 [2]
B. Decline since 2018	8	68,796 [8]	126,011 [8]	18 [8]	16 [7]	9 [8]	20 [8]	15 [7]
C. Fall then stabilised	12	41,181 [12]	62,999 [10]	3 [12]	15 [10]	18 [10]	23 [10]	51 [4]
D. Recent fall only	10	55,682 [10]	94,260 [10]	15 [10]	15 [10]	17 [10]	19 [10]	46 [5]
D2. Gradual decline	4	53,835 [4]	77,188 [4]	14 [4]	16 [4]	9 [4]	41 [4]	42 [3]
E. Rise then fall	5	17,585 [5]	63,806 [4]	4 [5]	14 [4]	32 [5]	42 [5]	42 [2]
F. Rebound	2	27,731 [2]	–	4 [2]	15 [2]	18 [2]	45 [2]	–
G. Rising	8	25,861 [7]	46,517 [7]	4 [8]	15 [8]	25 [7]	22 [7]	–
G2. Rise not sustained	2	47,961 [2]	74,054 [2]	2 [2]	14 [2]	16 [2]	15 [2]	14 [2]
S. Stable	5	38,547 [5]	38,884 [3]	3 [5]	15 [5]	22 [5]	31 [5]	–

Source: `system_features.csv`; `analyze_features.py`.

Group A (long decline). The members share high income (median GDP per capita 63,000 international dollars) and high cumulative spending; six of the ten with data had immigrant shares above 10 per cent in 2025 (Australia 32, Switzerland 38, Hong Kong 42, Germany 29, Greece 14, the Netherlands 13), but Bulgaria (3), Hungary (3), Latvia (5) and Finland (8) did not. Tracking age is spread from 10 (Germany) and 12 (the Netherlands, Switzerland, Costa Rica) to 16 (Australia, Finland, Latvia). Closure length ranges from zero full weeks (Australia) to 43 (Costa Rica). Governance ranges from the most school-autonomous system in the OECD table (the Netherlands, 92 per cent of decisions at school level) to among the most centralised (Greece 8 per cent, Switzerland 8 per cent at school level with 48 per cent at state level). Phone policy at the time of the 2025 test ranged from a national rule in force (the Netherlands from January 2024, Greece and Hungary from the 2024/25 year) to none (Finland's law took effect in August 2025, Latvia's in May 2025 for grades 1-6 only). Nothing in the feature set is shared by all members of A except wealth, and wealth is shared with B and D.

Group B (decline since 2018). Seven of eight are OECD members and the median GDP per capita is the highest of any group (69,000); the immigrant share is also the highest (median 18 per cent, Canada 39). Tracking is late (Canada, Denmark, Iceland and Norway 16; France and Portugal 15; Morocco 12; Belgium not coded in Table B3.1.4). Closures were short: median 9 full weeks, and Iceland had none. Governance is mixed (Iceland 60 per cent at school level, Norway 15, France 10). Phone policy: France's 2018 law is the only national ban that pre-dates 2024 in this group; Norway had national guidance from February 2024; Denmark, Belgium and Portugal legislated after the test. Group B is therefore the group whose members combine wealth, short closures and late tracking, which is the profile that a "closures caused the decline" account would not predict; it is also the group with the highest immigrant share, but the OECD's own decomposition attributes about 1.5 points of the 25-point OECD-29 reading decline to changing composition (Annex A1, p. 311; about 3 of Portugal's 30 points), which is far too small to explain a group whose median 2018-2025 three-domain change is -26 points.

Group C (fall then stabilised). This is a mixed group: five OECD members (Estonia, Mexico, Poland, the Slovak Republic, the United Kingdom) and seven partners, median GDP per capita 41,000, low immigrant shares except Cyprus (26), the United Kingdom (28) and Macao (52). Its distinguishing feature is that the 2022-2025 change is not significant in at least two domains, and its members' full-closure weeks span 10 (Slovak Republic) to 53 (Mexico), with Poland at 26 and the United Kingdom at 16. The four members with governance data have high school-level decision shares (Estonia 58, Slovak Republic 44, England 65, Mexico 17), which is the highest group median (51) but rests on four observations. Phone policy is mostly absent: Estonia and Poland have no national rule, the United Kingdom has non-statutory guidance from February 2024, and six members are not coded.

Group D (recent fall only). Five OECD members and five partners; wealth ranges from Argentina (30,000) to Singapore (145,000) and Ireland (133,000). Immigrant shares are high in Austria (27), Singapore (30), Ireland (23) and Malta (17), low in Czechia (8) and Argentina (6). Tracking age spans 10 (Austria) and 11 (Czechia) to 16 (Chile, Malta). Closures span 4 full weeks (Singapore) to 28 (Serbia). Governance data exist for five (Czechia 68, Chile 48, Austria 46, Ireland 46, Israel 19 per cent at school level). No feature is shared.

Group G (rising). All eight are non-OECD; six have GDP per capita below 26,000, but Qatar (129,000) and Brunei (85,000) do not, and the United Arab Emirates (78,000) is between. Immigrant shares are very low (1-9 per cent) in six and very high in Qatar (61) and the United Arab Emirates (57). Tracking age is 15 in five, 12 in Brunei, 14 in the United Arab Emirates and 16 in the Philippines. Closures were long: median 25 full weeks, with the Philippines at 75, the highest of any classified system. Only one phone-policy coding is a pre-2024 national ban (the United Arab Emirates, 2018 regulation), and five are not coded. What the G members share is a low or moderate starting level: seven of the eight had a 2018 three-domain mean below 460 (Chinese Taipei is the exception at 517). Across all 75 systems with 2018-2025 data, the correlation between the 2018 level and the 2018-2025 change is -0.46 ($p < 0.001$); within the 37 OECD members it is -0.23 ($p = 0.18$).

Group S (stable) and the small groups. The five stable systems (Brazil, Kazakhstan, Romania, Saudi Arabia, Uruguay) are all non-OECD with low immigrant shares and closures between 9 and 50 full weeks; Saudi Arabia is the only system coded as having reversed a phone ban. Group E pairs three Latin American systems that rose in 2018-2022 with long closures (32-34 full weeks) with two European systems that had short or no closures (Slovenia 21, Sweden 0) and rose in 2012-2018; the group is defined by timing and does not share anything else. Groups A2, F and G2 have two members each and are not summarised.

Does any feature discriminate? Table 8 tests each feature across the seven groups with five or more members (A, B, C, D, E, G, S), using a Kruskal-Wallis test, and Table 9 gives correlations of each feature with the average three-domain change across all classified systems.

Table 8. Kruskal-Wallis tests of feature differences across groups A, B, C, D, E, G, S

Feature	H	p	groups with n >= 3
GDP per capita 2023	8.89	0.18	7
Cumulative spending 6-15	7.26	0.30	7
Spending per lower-secondary student	8.60	0.13	6
Immigrant share 2025	10.68	0.10	7
Age of first tracking	6.03	0.42	7
Full closure weeks	17.21	0.009	7
Partial closure weeks	6.35	0.39	7
Total closure weeks	11.39	0.08	7
% decisions at school level	3.61	0.31	4
% decisions at central level	4.18	0.24	4

Table 9. Correlations between features and the average change across reading, mathematics and science (Spearman, with n and p; Pearson in brackets)

Feature	2018→2025, all	2018→2025, OECD only	2018→2022, all	2022→2025, all
GDP per capita 2023	-0.25, n=74, p=0.03 (-0.24)	-0.29, n=37, p=0.08	-0.06, n=69, p=0.60	-0.27, n=74, p=0.02

Feature	2018→2025, all	2018→2025, OECD only	2018→2022, all	2022→2025, all
Cumulative spending 6-15	-0.28, n=60, p=0.03 (-0.31)	–	-0.13, n=60, p=0.32	-0.32, n=64, p=0.01
Spending per lower-secondary student	-0.45, n=52, p=0.001 (-0.49)	–	-0.32, n=52, p=0.02	-0.38, n=56, p=0.003
Immigrant share 2025	-0.32, n=74, p=0.006 (-0.13, p=0.26)	-0.17, n=36, p=0.32	-0.20, n=69, p=0.09	-0.25, n=74, p=0.03
Change in immigrant share 2018-25	-0.26, n=68, p=0.03	–	–	–
Age of first tracking	-0.20, n=66, p=0.11 (-0.07)	-0.35, n=34, p=0.045 (-0.22, p=0.20)	-0.17, n=66, p=0.17	+0.01, n=70, p=0.91
Full closure weeks	+0.22, n=71, p=0.07 (+0.32, p=0.007)	+0.11, n=37, p=0.53 (+0.21, p=0.21)	+0.18, n=66, p=0.14	+0.33, n=71, p=0.005 (+0.40)
Total closure weeks	+0.28, n=71, p=0.02 (+0.27)	+0.25, n=37, p=0.14 (+0.29, p=0.09)	+0.18, n=66, p=0.16	+0.36, n=71, p=0.002
% decisions at school level	-0.03, n=34, p=0.85	(same set)	-0.02, n=33, p=0.91	+0.13, n=34, p=0.48
% decisions at central level	+0.09, n=34, p=0.63	(same set)	+0.13, n=33, p=0.47	+0.02, n=34, p=0.89

Source: `analyze_features.py`. A positive sign means that a higher value of the feature goes with a more favourable (less negative) change. All of this is observational, n is small, the features are measured at one point in time, and the outcome mixes three domains with different link errors.

Three results stand out and each needs qualification.

First, the one feature that differs significantly across groups, weeks of full school closure, differs in the direction opposite to the one usually hypothesised: the groups that declined most (A and B) had the shortest closures (medians 15 and 9 full weeks) and the rising group had the longest (25). The correlation between full-closure weeks and the 2018-2025 change is positive (Spearman +0.22, Pearson +0.32) across all 71 systems with data, and positive but not significant within the 37 OECD members (+0.11 / +0.21). Comparing the sustained decliners (A and B, median 12 full weeks, n = 18) with the other classified systems that did not rise (C, D, D2, S; median 16 weeks, n = 29) gives a Mann-Whitney p of 0.025, again in the “shorter closures, larger decline” direction. This is not evidence that closures helped. It is a composition effect: long closures were concentrated in Latin America and South and Southeast Asia (the Philippines 75 full weeks, Mexico 53, Costa Rica 43, Malaysia 42, Cambodia 40, Brazil 38, Peru 34), where several low-scoring systems rose, while the wealthy European and East Asian systems that declined most had short closures (Sweden 0, Iceland 0, Norway 5, Switzerland 6, France 7, Finland 8, Denmark 8). Once log GDP per capita is controlled, the closure coefficient falls to +0.20 points per full week (SE 0.16), that is, indistinguishable from zero, while log GDP itself has a coefficient of -7.2 points (SE 3.1). Within the OECD, where closures ranged from 0 to 53 full weeks, nothing in these data links closure length to the size of the 2018-2025 decline. That is the plain finding: closure length, as recorded by UNESCO, does not discriminate between trajectories once income is taken into account, and does not discriminate within the OECD at all. It says nothing about within-country effects of closures on the students who experienced them, which the OECD discusses separately (Box I.2.3) and which are outside the scope of this section. Two caveats cut in different directions: the UNESCO measure counts national status by calendar day and ignores regional variation (the United States is recorded as zero full weeks and 77 partial weeks; Germany 14 full weeks) and the intensity of remote instruction, so it is a coarse proxy; and the 2025 cohort was aged 10-11 at the onset of the pandemic, so closures at primary level are what matters, which the national-status series does not separate.

Second, the features that do correlate with the outcome across all systems, GDP per capita, spending per student and immigrant share, all point the same way: richer, higher-spending, higher-immigration systems declined more. But these are also the systems that started higher, and the 2018 level alone correlates at -0.46 with the change. Within the OECD the correlations weaken to non-significance (GDP -0.29, p = 0.08; immigrant share -0.17, p = 0.32). The Spearman and Pearson correlations for immigrant share differ (-0.32 against -0.13) because Qatar and the United Arab Emirates, with immigrant shares near 60 per cent, rose; without them the relationship is monotone but modest. The OECD’s decomposition puts the compositional contribution of rising immigrant shares at about 1.5 points of a 25-point OECD-29 reading decline (Annex A1, p. 311). A comparison of groups A and B with groups C, D, D2 and S shows higher medians for GDP (65,000 against 48,000, p = 0.055), immigrant share (16 against 8 per cent, p = 0.076) and cumulative spending (121,000 against 75,000, p = 0.078) among the sustained decliners; none reaches the 5 per cent level.

Third, governance and tracking do not discriminate at all. The share of decisions taken at school level ranges from 8 per cent (Greece, Switzerland, Türkiye) to 92 per cent (the Netherlands) among the 34 systems in the OECD table, and its correlation with the 2018-2025 change is -0.03. The sustained decliners include the most centralised (Greece, Portugal at 77 per cent central, Luxembourg at 83 per cent, France at 55 per cent) and the most decentralised (the Netherlands, Latvia 64 per cent, Iceland 60 per cent) systems. Age of first tracking is 10 in Austria and Germany, 11 in Czechia and the Slovak Republic, 12 in the Netherlands, Switzerland and Singapore, and 15-16 in most other systems; it has no relationship with the change across all systems (Spearman -0.20, p = 0.11; Pearson -0.07) and a weak one within the OECD (-0.35, p = 0.045; Pearson -0.22, p = 0.20), where the sign means later-tracking systems declined slightly more, which is again the Nordic and Anglophone pattern rather than anything about tracking. Early-tracking systems are found in A (Germany, the Netherlands, Switzerland), D (Austria, Czechia, Singapore) and C (the Slovak Republic), and late-tracking systems in every group.

Phone policy cannot discriminate between trajectories that end in spring 2025, for a reason of timing rather than of evidence. Of the 46 classified systems with a coding, only three had a national ban in force before 2024 (France 2018, Italy 2007, the United Arab Emirates 2018; China’s 2021 notice applies to B-S-J-Z, which is unclassified), and they sit in groups B, A2 and G; a further seven had national rules or guidance that took effect in 2024 (the Netherlands, Greece, Hungary, New Zealand, Norway, Kazakhstan, and England’s non-statutory guidance), spread across A, B, C, D2 and S. The bans that dominate current discussion took effect after the PISA 2025 test window: Austria and Latvia (May 2025, and for younger grades only), Finland (August 2025), Türkiye, Belgium, Luxembourg, Portugal and Chinese Taipei (September 2025), Korea (March 2026), Denmark and Sweden (2026). The 15-year-olds tested in 2025 were, in almost every system, schooled under whatever their school decided. The only claim the data support is negative: the two OECD systems with the longest-standing national bans, France and Italy, are in groups B and A2, one a sustained decliner and one flat, and the rising group contains one banning system and five uncoded ones. Whether the post-2024 wave of bans has any effect will first be observable in PISA 2029.

6. Where no discriminating feature is found, and what does discriminate

Stated plainly: among the features examined, none separates the trajectory groups at conventional significance once income is accounted for, and none separates them within the OECD. The single significant group difference, in closure length, runs opposite to the usual hypothesis and disappears when GDP per capita is controlled. Governance structure, tracking age and phone policy show no relationship to the trajectory in any specification. GDP per capita, spending and immigrant share correlate with the size of the decline across all systems, but this is largely the same fact as “higher-scoring systems fell more”, and the OECD’s own decomposition limits the compositional contribution of immigration to a few points.

What does discriminate is where a system started and whether it is an OECD member. Every system in the rising group is a non-OECD system, and seven of eight began below 460 points in 2018; every system in the long-decline and decline-since-2018 groups except Bulgaria, Costa Rica and Morocco is an OECD member or Hong Kong, and all but Greece and Hungary among those began above 480. The three non-OECD exceptions (Bulgaria, Costa Rica, Morocco) are low-scoring, which shows that a low starting level is not sufficient for a rise. Whether the association between starting level and decline reflects regression to the mean, a ceiling in what low-stakes testing can measure among high-performing adolescents, the OECD’s engagement evidence (Annex A1) applying more to some populations than others, or changes in the tested constructs that affected high-scoring systems differently, cannot be decided from these tables. The technical note in this report sets out why the science trend in particular, with a link error of 3.1-3.5 points and a revised framework, is the least secure of the three, and readers should weigh the science-driven rises (Georgia, Türkiye, Montenegro) accordingly.

Two further limitations should be recorded. The classification treats the six asterisked 2025 samples and the thirteen asterisked 2022 samples as the OECD does, without adjustment; if the Netherlands’ 2025 mean is biased upward by up to 11 points and Canada’s by up to 15, their declines are, if anything, understated, which would not change their group. And the features are national averages measured at one date; sub-national variation in closures, phone rules and governance (large in Australia, Canada, Germany, Spain, Switzerland and the United States) is not captured.

Files

`trajectory_groups.csv`: one row per entity with every window change, SE and verdict, coverage indices and group. `system_features.csv`: one row per entity with the features and a source column for each. `hand_coded_features.csv`: the governance and phone-policy codings with sources. `classify_trajectories.py`, `build_features.py`, `analyze_features.py`: the scripts; `derived/` holds the aggregated UNESCO closure weeks, the World Bank download and the aggregation script.

Appendix D. What PISA measures and how far it can be trusted

IN BRIEF (summary added before publication; see Appendix H)

This section reviews PISA as a measuring instrument, against OECD-average declines of 25 points in reading since 2018 and 22 in mathematics since 2015. No single issue can account for them. Scoring-model uncertainty is a few points and largely cancels in trends. Sampling bias, where quantified, is upward: 7 to 15 points for asterisked systems but about 1 to 2 on the OECD average (this section's calculation). The reading decline appears in the non-adaptive section of a test unchanged since 2018, a genuine but unreplicated check. A material part of what PISA reading now measures, and of what has changed, is students' willingness and capacity to engage with a low-stakes test, which the data cannot separate from skill; this section puts it at a small fraction to roughly a third of the reading decline. Several systems with large reported declines have the least secure samples, and the 2025 science comparison spans three design changes the OECD has yet to analyse.

Scope and how to read this section

This section reviews the validity evidence on PISA as an instrument: the framework and how it differs from curriculum-based studies; the psychometric model and its critics; sampling, exclusion and coverage; test-taking effort; mode and design changes; the linking of scales across cycles; the broader critiques and the OECD's responses; and the evidence that PISA scores predict later outcomes. For each issue it states what is established, what is contested, and how large the effect could plausibly be against the changes reported in *PISA 2025 Results (Volume I)* (OECD, 2026, https://doi.org/10.1787/73451bc5-en; "Volume I"): an OECD-average reading decline of 25 points between 2018 and 2025 (OECD average-35, Table I.B1.2a.37, Annex A1 p. 309) and a mathematics decline of 22 points between 2015 and 2025 (485 to 463; Preface, note 2). Where a magnitude is my own calculation rather than a published figure, it is labelled as such. The section draws on the technical note in `A/technical_note.md`, whose quotations from Volume I I checked against the text and found accurate, with one minor exception noted below.

The assessment framework and the contrast with TIMSS

PISA's own description of what it measures is deliberately not a curriculum. Volume I states that the tests "do not just measure whether students near the end of their compulsory education can reproduce what they have learned; they also assess how well students can extrapolate from what they have learned, both in and outside of school, and apply it to real-life problems" (Volume I, "What is PISA?", p. 47). The target population is age-based (15 years 3 months to 16 years 2 months, enrolled in grade 7 or above) rather than grade-based, on the ground that "differences in education systems – such as entry age, early childhood provision and structure – make grades difficult to compare internationally" (Annex A2, note 1, p. 325). The constructs are defined in the *PISA 2025 Assessment and Analytical Framework* (OECD, 2026, https://doi.org/10.1787/86c36975-en); for 2025 the science framework was revised, with a new competency ("research, evaluate and use scientific information for decision making and action") and a new environmental-science subscale (Volume I p. 77; Table I.B1.2a.8).

TIMSS is built on the opposite premise. Its framework is organised around the "intended, implemented and attained" curriculum, samples by grade (4 and 8), and specifies content domains (number, algebra, geometry, data; biology, chemistry, physics, earth science) negotiated with participating countries to reflect what is taught (TIMSS 2023 Assessment Frameworks, https://timssandpirls.bc.edu/timss2023/frameworks/pdf/T23_Frameworks.pdf; https://timss2023.org/wp-content/uploads/insights/T23_Insights_CurriculumAlignment.pdf). Wu's comparison of PISA 2003 and TIMSS 2003 mathematics found that Western countries did relatively better on PISA and Asian and Eastern European countries relatively better on TIMSS, and attributed the gap mainly to content balance (PISA has more data-handling and "real-world" items, TIMSS more number and algebra) and to the age-versus-grade sampling definition (Wu, 2010, OECD Education Working Paper 32, https://www.oecd.org/en/publications/comparing-the-similarities-and-differences-of-pisa-2003-and-timss_5km4psnm13nx-en.html). Volume I itself acknowledges that PISA mathematics "emphasises students' capacity to apply mathematical knowledge and reasoning to real-world situations, which are presented to them in the form of written texts," and uses this to explain why reading trends correlate at 0.79 to 0.82 with science and mathematics trends (Annex A1, p. 309).

What is established is that PISA and TIMSS measure related but distinct things and that a country's rank can differ between them by design rather than by error. What is contested is whether "literacy" as PISA defines it is the more policy-relevant construct. Critics from science education (Sjøberg, 2015, https://www.ejmste.com/article/pisa-and-global-educational-governance-a-critique-of-the-project-its-uses-and-implications-4353) argue that a test detached from national curricula and translated into dozens of languages measures a construct nobody teaches and rewards particular reading and cultural conventions. The practical implication is that PISA declines are declines on PISA's construct; the OECD's finding that the reading decline is concentrated on long texts and on "hasty" responding (Annex A1) is as much a statement about the construct as about the students.

The psychometric model and its critics

PISA scores come from an item response theory (IRT) model with plausible values imputed from test responses conditioned on questionnaire data. From 2000 to 2012 the model was a one-parameter (Rasch) model with international item parameters; from 2015 the contractor (ETS) moved to two-parameter and generalised partial credit models, calibrated concurrently on several cycles, with item-by-country and item-by-cycle deviations permitted for a subset of items ("partial invariance") (https://link.springer.com/chapter/10.1007/978-3-031-90951-1_14).

The best-known critique is Kreiner and Christensen's reanalysis of the 2006 reading data (Psychometrika, 2014, https://link.springer.com/article/10.1007/s11336-013-9347-z; working paper https://ncm.gu.se/media/ncm/dokument/pisa_kreiner_.pdf). Using 20 items administered in all 56 countries, they found "strong evidence of misfit of the PISA scaling model" and "very strong evidence" of differential item functioning (DIF) across countries. Ranking countries on different subsets of items, Denmark ranked 17th on all 20 but 3rd on one subset and 42nd on another; the United Kingdom 23rd overall but 8th and 36th on two subsets. Kreiner concluded in interviews that "the best we can say about PISA rankings is that they are useless"; Goldstein argued that removing items that show DIF "smooths out" the differences that are educationally interesting (Tes, 2013, https://www.tes.com/magazine/archive/pisa-fundamentally-flawed); Spiegelhalter judged that with a misspecified imputation model "the rankings would be even less reliable than claimed" (https://understandinguncertainty.org/pisa-statistical-methods-more-detailed-comments). The OECD's reply, from Ray Adams, was that Kreiner's subsets were "small groups selected to show most variation" and that the full item pool evens out the variation (Tes, 2013).

That reply has been tested independently, and the results are the most important thing a reader needs to know about this debate. Jerrim and colleagues re-scaled PISA 2015 for 35 OECD countries under Rasch and 2PL models, with and without item-by-country interactions, with 2006 rather than 2015 item parameters, and with not-reached items scored wrong. Rank correlations with the official results were about 0.99 in every condition and the largest changes in country means were 3 to 4 points; "even when multiple alterations are made to the scaling model, this only has a trivial impact upon cross-country comparisons within a given PISA cycle" (Educational Measurement: Issues and Practice, 2018, https://onlinelibrary.wiley.com/doi/10.1111/emip.12211; manuscript https://discovery-pp.ucl.ac.uk/10049242/). Robitzsch's comparison of eleven IRT models on PISA 2009 found the range of a country's mean across models averaged 2.4 points in reading (0.3 for Sweden to 7.7 for Japan) and 2.8 in mathematics and science, with model error for means about a quarter of the sampling standard error; for standard deviations and percentiles, model error was of the same order as sampling error or larger (https://pmc.ncbi.nlm.nih.gov/articles/PMC9223051/). Rutkowski and Rutkowski (Educational Researcher, 2016, https://journals.sagepub.com/doi/10.3102/0013189X16649961) accept that means are fairly robust but stress that the equal-difficulty assumption "frequently fails," that violations "have consequences for ranking especially middle-performing" systems, and that this error is not carried in PISA's confidence intervals.

In summary: it is established that a fully invariant Rasch model does not fit PISA data and that country DIF is pervasive, which the OECD conceded in practice by adopting partial invariance in 2015. It is established that with the full item pool the scaling model moves country means by a few points, not tens, so the Kreiner rank ranges describe 20 items chosen to maximise disagreement rather than the sensitivity of the published scale. It is contested how much residual DIF error remains among closely spaced mid-table countries, and established that model choice matters more for spread, percentiles and shares below Level 2 than for means. Against a 22- or 25-point average change, scaling uncertainty of 3 to 8 points at the country level is not of the same order, and because one model is applied to all cycles it largely cancels in trends.

Sampling, exclusions and coverage

PISA's technical standards require overall exclusions not to exceed 5% of the desired target population, a weighted school response rate of at least 85% before replacement (with a sliding after-replacement target), and a weighted student response rate of at least 80% (Volume I, Reader's Guide, p. 17; Annex A2 pp. 322 to 324). In 2025 the Adjudication Group flagged Canada (school response 79%/82% after replacement, student response 77%, upward bias "unlikely to exceed 15 points in science"), the Netherlands (exclusions 9.3%, school response 66%/88%, student response 78%, possible student non-response bias "up to ±9.9% of a standard deviation (i.e. about 11 score points in science)"), New Zealand (exclusions 8.1%, school response 45%/52%, student response 76%, mean scores "may be over-estimated by about 10 to 15 score points"), Norway (exclusions 10.4%), Albania (76% school response, small schools not contacted, science lower bound 428) and the United States (exclusions 6.9%, school response 45%/54%, student response 76%, range-of-bias "up to about ±15.4%" of a standard deviation for schools and "±15.9%" for students, direction undeterminable), plus seven Canadian provinces and Murcia; Catalonia was suppressed after within-school exclusions rose from 5.6% to 23.2% (pp. 17 to 19). The OECD adds that "there may be a need for subsequent adjustments as new evidence on the quality and comparability of the data emerges" (p. 17). No table of 2025 exclusion and response rates for all systems is published; they are deferred to Chapter 14 of the Technical Report (Annex A2, p. 322).

The 2022 cycle, the baseline for the 2022-to-2025 comparisons, had more extensive problems. The England national report records that 13 systems failed sampling standards in 2022 and that the OECD "estimates that this may translate to an upwards bias of between 7 and 10 points in 5 of these education systems but the data required to calculate the size of the bias were not available in the remaining 8 education systems, and the OECD concluded that bias in these systems' data could not be ruled out"; for England the estimated upward bias was "approximately 7 or 8 points," with an initial school response of 66% and a pupil response of 75% (Department for Education, pp. 8 and 21, https://assets.publishing.service.gov.uk/media/656dc3321104cf0013fa742f/PISA_2022_England_National_Report.pdf). The same report lists Australia, Canada, Hong Kong, Ireland, the Netherlands, Northern Ireland, New Zealand, Scotland, the United States and Wales among the systems below threshold. New Zealand's own 2022 analysis found 72% participation, under-representation of chronically absent students, and a likely upward bias of "about 10 PISA points" (https://www.educationcounts.govt.nz/publications/schooling/pisa-2022-non-response-bias-analysis); Scotland's respondents had significantly higher prior attainment and lower free-school-meal eligibility than non-respondents (NFER, https://www.nfer.ac.uk/media/xweb1h0f/report_on_the_pisa_2022_student_sample_in_scotland.pdf); the US school response was 63.3% after replacement (NCES, https://nces.ed.gov/surveys/pisa/pisa2022/technical-notes/index.asp). I could not retrieve the OECD's 2022 Reader's Guide directly (the site returned a bot-check page), so the 2022 list is taken from national reports. Volume I compares these systems with 2025 without adjustment and without carrying a 2022 asterisk, noting only that New Zealand's 2022 participation "was even lower, at 72%," and judging that the bias "is unlikely to affect comparisons between PISA 2022 and PISA 2025 in meaningful ways" (p. 18).

The direction of non-response bias is consistent: non-responding pupils are lower-achieving. Durrant and Schnepf, linking English PISA samples to administrative records, found that low-achieving pupils had “considerably lower response probabilities” and that schools at both extremes responded less often, so non-response understates dispersion as well as inflating the mean (Journal of the Royal Statistical Society A, 2018, https://academic.oup.com/jrsssa/article/181/4/1057/7072033). Jerrim found total non-participation in parts of the UK in 2018 “of around 40% (amongst the highest anywhere in the world)” (Review of Education, 2021, https://discovery.ucl.ac.uk/id/eprint/10141316/), and the UK Office for Statistics Regulation agreed the national reports “were not sufficiently transparent about the limitations of the data and the potential sources of bias” (https://osr.statisticsauthority.gov.uk/correspondence/ed-humpherson-response-to-professor-john-jerrim-request-to-undertake-an-independent-review-of-the-pisa-data-across-the-uk).

Coverage is a distinct issue. Coverage Index 3, the weighted number of participating students divided by the total population of 15-year-olds, “ranges from 32% in Rwanda and 43% in Cambodia to 90% or more in 34 countries and economies” in 2025, and the OECD warns that results “cannot be readily generalised to the entire population of 15-year-olds in countries where many young people of that age are not enrolled” (Volume I, p. 111; Table I.A2.1). Volume I addresses this for low performance by computing a lower bound under the assumption that every uncovered 15-year-old scores below Level 2 (Figure I.2.11, p. 112); it does not do so for means or trends, and where coverage has expanded (Cambodia, the Dominican Republic and Brazil are discussed on p. 108) an unchanged mean is consistent with improvement among the previously covered group.

How large are these effects relative to the reported changes? For an individual flagged system, a 7- to 15-point upward bias is of the same order as the country’s reported 2022-to-2025 change; that is why the OECD asterisks them. For the trend of a flagged system the question is whether the bias changed between cycles; the OECD’s argument for New Zealand is that it did not, which is plausible but not demonstrated, and for the United States the direction is unknown in both years. For the OECD average the arithmetic is more forgiving: if six of the 35 trend countries carried an upward bias of 10 points in one cycle and not the other, the OECD-35 average would be biased by about 1.7 points (my calculation), small against 22 or 25 points. Exclusion-rate creep (Norway 7.3% to 10.4%, the Netherlands 8.4% to 9.3%, Spain’s regions) works against the reported declines, since excluded students are typically low-performing, so those systems’ declines are if anything understated. The established point is that sampling bias is real, quantified for a minority of systems, upward where quantified, and concentrated in a familiar set of English-speaking and Dutch systems; the contested point is whether reporting rather than adjusting, and dropping the asterisk from baseline cycles in trend tables, gives readers an adequate warning.

Test-taking effort

PISA is low-stakes for students: they receive no feedback and face no consequences (Volume I, Annex A1, p. 316). The question is whether effort differs between countries and over time enough to distort comparisons.

The cross-country evidence is the stronger part. Gneezy, List, Livingston, Qin, Sadoff and Xu offered US and Shanghai students a financial incentive immediately before a 25-question PISA-style mathematics test. US students improved by about 0.2 to 0.23 standard deviations, with the effect concentrated in the second half of the test; Shanghai students did not respond. The authors calculate that raising US effort would add 22 to 24 points and move the United States from 36th to roughly 19th in the 2012 mathematics ranking (AER: Insights, 2019, https://www.aeaweb.org/articles?id=10.1257%2Faeri.20180633; working paper http://fieldexperiments-papers2.s3.amazonaws.com/papers/00614.pdf). Borghans and Schils found that the probability of a correct answer declines roughly linearly through the test, by 4% in Finland and 15% in Greece, that the country-level decline is stable across cycles (correlations of 0.84 to 0.89 between 2003 and 2006), and that this “motivation effect” explains 19% of the between-country variance in mean scores (https://www.sole-jole.org/assets/docs/13260.pdf). Zamarro, Hitt and Mendez, using item non-response, careless answering and survey-based proxies, report that their effort measures “explain between 32 and 38 percent of the variation in test scores across countries” but only a minor share within countries (Journal of Human Capital, 2019, https://www.journals.uchicago.edu/doi/10.1086/705799). Two more recent studies are more reassuring. Michaelides and Ivanova, using response times in PISA 2015 for 59 systems, found rapid guessing at about 3% of responses on average, mostly between 1% and 6%, highest in Qatar and lowest in Estonia and Macao, and positively correlated with country mean scores ($r_s = 0.27$ in mathematics, 0.37 in reading) (https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-3/PTAM__3-2022_5_kor.pdf). A 2026 multilevel analysis of self-reported effort in PISA 2022 across 36 OECD countries found effort associated with achievement at the student ($\beta = 0.18$) and school ($\beta = 0.18$) levels but not at the country level ($\beta = -0.01$, $p = 0.93$) (Large-scale Assessments in Education, https://link.springer.com/article/10.1186/s40536-026-00288-9). The discrepancy between the experimental and observational literatures is partly a matter of what is measured (self-reports correlate with both ability and conscientiousness; response-time indicators catch only the crudest disengagement) and is unresolved.

A concrete demonstration is Spain in 2018, whose reading results the OECD initially withheld. In regions that had moved high-stakes tenth-grade examinations earlier in the year, the share of students giving patterned answers to the reading-fluency items rose by 16 percentage points in the final testing weeks, self-reported effort fell, and reading scores of late-testing schools were 14 points lower and mathematics 11 points lower; the OECD judged that “the overall impact on the country’s mean performance did not exceed a handful of PISA points” but was larger for the five regions concerned (PISA 2018 Results Volume I, Annex A9, https://www.oecd.org/content/dam/oecd/en/publications/support-materials/2019/12/pisa-2018-results-volume-i_947e3529/pisa-2018-volume-I-annex-nine-2020.pdf).

The evidence on effort over time is thinner and comes mainly from the OECD, because the effort thermometer was only introduced in 2018. Volume I reports that self-reported effort fell by 0.5 points on the 10-point scale between 2018 and 2025 across 35 OECD countries (0.3 since 2022), that the share reporting very little effort rose from 3.5% to 5.4%, that “hasty” (fast and incorrect) responses on reading items rose by over 4 percentage points to about 9% while non-reached items fell by about 2 points, that “hasty” test-takers on the reading-fluency task rose from 6.6% to 11.4% with most of the increase after 2022, and that mathematics also shows more hasty responses between 2022 and 2025 (Annex A1, pp. 314 to 316; Tables I.A1.1, I.A1.12, I.A1.15). Against this, questionnaire straightlining fell rather than rose (Table I.A1.17), first-hour versus second-hour differences did not widen (Tables I.A1.3 to I.A1.5), and fatigue within the linear science test is modest (Table I.A1.2). The OECD's interpretation is that declining effort is "part of what the 'reading' test in PISA assesses" (p. 319). That is a defensible reading of the framework, but it is an interpretive choice; a reader who regards willingness to try on a low-stakes test as separate from reading skill will see the same tables as evidence that part of the decline reflects less effort on the test rather than a loss of reading skill. The technical note is right that the timing thresholds behind the "hasty" indicators are defined separately by year because of the 2025 platform change (Annex A1 notes 2 and 3, p. 321), so the hasty-response trend rests on a within-year normalisation for which no sensitivity check is published.

How much of the 25-point reading decline could effort account for? The following is my own bounding exercise. The Gneezy experiment shows that effort can move a mean by about 20 points when students who are not trying are induced to try; that is an upper bound for a cross-sectional gap, not an estimate of a change. Taking the 0.5-point fall in self-reported effort (about 0.23 SD on a scale with SD 2.15) and the student-level association of β = 0.18 from the 2022 study gives about 0.04 SD, or 4 points, almost certainly a lower bound because self-reports are noisy. The doubling of hasty fluency responses and the 4-point rise in hasty item responses point to something larger; if the additional hasty responses were treated as invalid rather than wrong, the effect on percent-correct is a few percentage points, of the order of 5 to 10 scale points. A plausible range is therefore that between a small fraction and roughly a third of the OECD-average reading decline since 2018 reflects changed test-taking behaviour rather than changed skill; the data cannot separate the two, and because effort fell in nearly all countries the issue affects levels more than the cross-country pattern of change.

Mode effects and adaptive testing

PISA moved from paper to computer in 2015. Jerrim and colleagues analysed the OECD's randomised 2014 field trial for Germany, Ireland and Sweden. Computer-based scores were lower by 0.10 to 0.22 SD in mathematics, 0.15 to 0.20 SD in reading, and 0.07 (Sweden) to 0.25 (Germany) SD in science; two-thirds to four-fifths of items were harder on screen. ETS estimated a separate mode effect per item and anchored the link to previous cycles on items found invariant across modes (61 science, 51 mathematics, 65 reading); restricted to those items the residual effect largely disappeared in mathematics but −0.19 SD remained in science for Germany (Oxford Review of Education, 2018, https://www.tandfonline.com/doi/abs/10.1080/03054985.2018.1430025; manuscript https://discovery.ucl.ac.uk/id/eprint/10039757/). Jerrim's summary was that the raw gap was about 20 points and "a 10-point gap remained common even after the OECD applied adjustments" (https://ffteducationdatalab.org.uk/2018/01/is-pisa-still-a-fair-basis-for-comparison-some-serious-questions-have-emerged/). Robitzsch, Lüdtke, Goldhammer, Kroehne and Köller re-estimated German trends with and without the field-trial adjustment: 2012-to-2015 science was −15 to −21 points without and +4 to +7 with it, mathematics −8 to −16 without and +1 to +2 with, while 1PL versus 2PL scaling made little difference (Frontiers in Psychology, 2020, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.00884/full). It is established that the transition introduced a country-varying mode effect of 10 to 25 points, partly but not wholly removed, making the 2012-to-2015 link the weakest in the series. For the changes at issue here this matters less than it might seem: the 2015-to-2025 mathematics and 2018-to-2025 reading comparisons are computer-to-computer for nearly all systems. The exceptions are Viet Nam, excluded from all trends because the uncertainty "beyond linking errors remains undetermined," and Guatemala and Paraguay, reported with a caution that the uncertainty "extends beyond what is captured by the linking errors" (Reader's Guide, p. 20); six systems remained on paper with trend items only (p. 50). Volume I presents no mode-effect study for any of these. A further undocumented change in 2025 is the test-delivery platform, rebuilt by a new contractor (p. 9), mentioned only in the Annex A1 endnotes on timing thresholds; there is no evidence in Volume I either way on whether it affected scores.

Adaptive testing was introduced for reading in 2018, mathematics in 2022 and, on Volume I's description, science in 2025 (with a linear form for 25% of test-takers, Annex A1 p. 318). The reading design routes students after a Core stage into easier or harder blocks with probabilities of 0.9/0.5/0.1 depending on prior performance (NCES, https://nces.ed.gov/surveys/pisa/pisa2018/pdf/MSAT-ReadingWebReport.pdf; OECD Education Working Paper 209, https://www.oecd.org/en/publications/introduction-of-multistage-adaptive-testing-design-in-pisa-2018_b9435d4b-en.html). Frey and colleagues find the design raised test information by only 4 to 7% and is "more restrictive than necessary," but identify no comparability failure (https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1446799/full). The worry that adaptivity could manufacture a trend is addressed directly in Volume I: percent-correct on the trend items, weighted by inverse routing probabilities, declines in the non-adaptive Core section as well as in Stages 1 and 2, country-level changes in percent-correct correlate 0.90 to 0.91 with changes in mean scores, and "the vast majority of declines are clearly visible already in the first, non-adaptive section of the test" (Annex A1, p. 312; Table I.A1.7; Annex A3, p. 329). This is a genuine check and the best available evidence that the reading decline is not an artefact of the adaptive design; it is descriptive, unreplicated, and has no counterpart for the newly adaptive science test or for mathematics. Adaptive designs also change precision at the tails, so shares below Level 2 or above Level 5 are more design-sensitive than means.

Linking scales across cycles and the link error

Each cycle's results are placed on the scale of the cycle in which the domain was last major, using common "trend items," and a "link error" is added to the sampling variance of every cross-cycle comparison. Volume I says the link error "represents uncertainty around scale values" and gives 2.9 points as the science 2022-to-2025 example (Reader's Guide, p. 24); Table I.A3.1 publishes the link errors for the regression-based "average decennial trend," from 2.6 (reading, 2018 to 2025) to 4.9 (reading, 2017 to 2025) points per decade. The technical note's back-calculation of pairwise link errors from the standard errors in Tables I.B1.2a.36 to 38 (for comparisons with 2025: from 2022, science 3.1, reading 1.1, mathematics 1.2; from 2018, 3.5, 1.8, 2.6; from 2015, 3.4, 3.8, 3.0; from 2012, 6.1, 6.1, 3.8) is a legitimate reconstruction, since the identical implied value across countries confirms the additive structure; it does not reproduce the 2.9 quoted in the Reader's Guide, a small unexplained discrepancy. Two features stand out. Science link errors for the two most recent comparisons are about three times those for reading, consistent with a scale re-anchored across the 2025 framework revision, which Volume I nowhere discusses as a comparability issue. And links to 2012 and earlier carry errors of 6 points or more, so a decade-scale change of 10 points in one country is at the edge of significance on the OECD's own accounting.

The academic literature holds that these figures understate the uncertainty. Gebhardt and Adams showed in 2007 that reported trends are sensitive to the equating method and choice of link items (Journal of Applied Measurement, https://pubmed.ncbi.nlm.nih.gov/17804896/). Robitzsch and colleagues argue that "the uncertainty in the original trend estimate is underestimated by the officially reported link error because country DIF is ignored" (Frontiers in Psychology, 2020), and Robitzsch and Lüdtke conclude that the link error used since 2015 "provides no generalization of the old PISA linking error but rather reflects different aspects of uncertainties and model error" (Measurement, 2024, https://eric.ed.gov/?q=PISA&id=EJ1413302); no magnitude is given in the record I could access. Sachse, Mahler and Pohl add that changes in the non-response mechanism across cycles can themselves bias trends (https://doi.org/10.1177/0013164419829196). Volume I's own cross-check is the Annex A1 item-level analysis, which uses the same 2018 reading tasks in all three cycles, so the reading trend since 2018 rests on a genuinely common instrument; no such check exists for mathematics or science.

For magnitude: link errors of 1 to 4 points for recent comparisons are small against 22 or 25 points; if the critics are right that country-level trend uncertainty is understated by a factor of about two, a single country's 2018-to-2025 reading change would carry perhaps 5 to 8 points of linking uncertainty (my inference), still well below the average decline though enough to make many individual three-year changes non-significant. The science trend deserves the most reserve, because its link runs across a framework revision, a new adaptive design and a new platform simultaneously, none of which is analysed in Volume I.

The broader critiques and the OECD's responses

The scholarly critique runs along three lines that call for different kinds of evidence. The first is technical and has been covered above; its most prominent authors are Goldstein (Assessment in Education, 2004, https://eric.ed.gov/?id=EJ680318), whose concerns were the restricted modelling, the interpretations drawn from it, features that "raise questions about the adequacy of the data," and the absence of a longitudinal component; Kreiner and Christensen; Rutkowski and Rutkowski; and Jerrim, whose work on sampling has been the most consequential in practice. The second is about construct and culture: Sjøberg (2015) argues that curriculum-free, translated science items privilege a particular genre of reading and reward test-taking over interest in science. The third is about governance and effects: Meyer and Benavot's *PISA, Power, and Policy* (Symposium Books, 2013) frames PISA as an instrument of "global educational governance" through which an economic organisation sets the terms of national debate; the 2014 open letter to Andreas Schleicher signed by Meyer, Zahedi, Andrews and about 80 others argued that the three-year cycle "has caused a shift of attention to short-term fixes," narrows curricula, and relies on commercial testing firms (https://ojs.didaktik-der-mathematik.de/index.php/mgdm/article/download/339/335/628); Zhao's synthesis (Journal of Educational Change, 2020, https://link.springer.com/article/10.1007/s10833-019-09367-x) groups the criticisms into a narrow conception of education, implementation flaws and misinterpretation by policymakers, and argues they "have been largely ignored." Hopfenbeck and colleagues' systematic review of peer-reviewed PISA articles from 2000 to 2015 (Scandinavian Journal of Educational Research, 2018, https://eric.ed.gov/?id=EJ1174062) concludes that the data have "led to progress in educational research" while stressing "the need for caution when using this research to inform educational policy."

The OECD's responses have been of two kinds. On technical matters it has changed practice: 2PL scaling and partial invariance in 2015, range-of-ranks tables computed by simulation with a multiple-testing correction (Annex A3, pp. 327 to 328), link errors in all cross-cycle significance tests, release of Spain's anomalous 2018 data with a full analysis, an effort thermometer from 2018, and the Annex A1 item-level checks in 2025, which directly answer the objection that a trend could be a design artefact. It annotates rather than suppresses sub-standard samples, which critics regard as too lenient. On construct and governance it has not conceded ground: Volume I reasserts that the reading construct includes sustained attention and that the choice of what to measure "reflects a consensus among the countries that are represented in its Governing Board" (p. 319), and the Preface moves directly from results to policy prescriptions on curriculum focus, teacher skills and device use (pp. 4 to 5), the practice the governance critics object to. The OECD's own text is careful in places the Preface is not: Annex A1 says that correlations between attitudes and reading trends "do not constitute proof of causality" (p. 316), and Box I.2.2 warns against a single "20 points equals one year" conversion, reporting country estimates from about 10 points (Serbia) to about 25 (United Kingdom) (p. 90; the technical note gives 28 for Scotland from Table I.2.7, which I could not check in the text file). The Preface nonetheless describes the 22-point mathematics decline as "equivalent to just over one year of learning."

Does PISA predict later outcomes?

The predictive-validity evidence is the strongest part of the case for PISA at the individual level and the most contested at the national level.

Individual follow-ups exist for Canada, Australia, Denmark and Switzerland (OECD, *Learning beyond Fifteen*, 2012, p. 20, https://www.oecd.org/content/dam/oecd/en/publications/reports/2012/04/learning-beyond-fifteen_g1g1b72a/9789264172104-en.pdf). Canada's Youth in Transition Survey followed the PISA 2000 cohort to 25. Students at reading Level 5 at 15 "were 20 times more likely to access university than those at or below Level 1," and high scores contributed to secondary completion and post-secondary participation "even after taking other student background characteristics into account" (OECD, *Pathways to Success*, 2010, pp. 3 and 8, https://www.oecd.org/content/dam/oecd/en/publications/reports/2010/03/pathways-to-success_g1ghc4dd/9789264081925-en.pdf). By 25, 46.0% of those below Level 3 had completed post-secondary education against 77.3% at Level 4 or 5; 9.4% against 45.7% had a university degree; among full-time workers the earnings gap between the lowest and highest groups was about 10% (Statistics Canada, https://www150.statcan.gc.ca/n1/pub/81-004-x/2011002/article/11490-eng.htm). The Danish follow-up of 1,881 PISA 2000 students re-tested in PIAAC in 2011/12 found "a clear relationship between reading skills at age 15 and how well the young people fare later on," in reading at 27, tertiary completion and risk of long-term benefit receipt (Rosdahl, SFI 14:13, https://uvm.dk/media/pexpdkix/140617-fra-15-27-aar.pdf, my translation). Borgonovi, Ferrara and Piacentini, using LSAY in Australia (n = 3,110), Danish register-linked data (n = 1,130) and Switzerland's TREE (n = 1,962), found that the decline in performance during the PISA test predicts tertiary completion at 25 independently of the score, a one-SD steeper decline lowering the probability by 6 to 7 percentage points, whereas self-reported motivation had no predictive value (Journal of Adolescence, 2021, https://discovery-pp.ucl.ac.uk/10136598/1/Borgonovi_RPS.pdf). This cuts both ways: PISA scores carry a non-cognitive component, and that component is itself predictive, so "contamination" by effort is not simply noise. All of these studies are correlational with partial controls, and attrition is selective (the Canadian follow-up sample had a PISA mean of 541 against 534 for the full cohort; *Learning beyond Fifteen*, note 1, p. 46). Reading level at 15 sorts strongly on educational pathways and weakly on earnings by 25.

The national-level claim is Hanushek and Woessmann's: that a country's mean score on international tests, pooled from 1964, predicts its subsequent growth in GDP per capita far better than years of schooling, with support for causality from instrumental variables, immigrant earnings and within-country changes (Journal of Economic Growth, 2012, https://www.nber.org/papers/w14633). Critics have concentrated on timing. Komatsu and Rappleye show that the test scores (1964 to 2003) overlap the growth period (1960 to 2000) and argue that when scores are related to strictly subsequent growth (1995 to 2014) the variance explained by score changes falls to about 10% (Comparative Education, 2017, https://eric.ed.gov/?id=EJ1135711; https://freshedpodcast.com/komatsurappleye/). Kamens, examining 1960 to 2012, finds "little evidence of macro-level effects of test score performance on subsequent economic growth" (Comparative Education Review, 2015, https://www.journals.uchicago.edu/doi/abs/10.1086/681989). On the other side, Heller-Sahlgren and Jordahl extend the data to 2016 and find that "the positive relationship between growth and test scores holds," driven more by top performers than by basic proficiency (https://ideas.repec.org/p/hhs/iuiwop/1408.html), and Balart, Oosterveen and Webbink find both the starting level and the within-test decline positively and about equally associated with growth (Economics of Education Review, 2018, https://ideas.repec.org/a/eee/ecoedu/v63y2018icp134-153.html). The correlation is robust to updating; its causal interpretation rests on a small number of countries over overlapping periods and is contested. The growth literature uses means averaged across tests and decades; nothing in it speaks to whether a three- or seven-year change on PISA predicts anything.

Summary of components

"Plausible size" is the range the cited evidence supports for a single country's mean or trend unless stated; OECD-average effects are smaller because country-specific errors partly cancel.

Issue	Established	Contested	Plausible size against reading −25 (2018-25) and mathematics −22 (2015-25)
Framework	PISA and TIMSS rank countries differently by design (Wu 2010)	Whether "literacy" is the right construct; whether engagement belongs in it	A definitional choice, not an error; part of the reading decline is, on the OECD's account, a decline in sustained engagement
Scaling model and DIF	Invariant Rasch misfits; full-pool re-scaling moves means 3-4 points, ranks r≈0.99	Residual error for mid-table comparisons; distributional statistics	3-8 points for a country mean within a cycle; largely cancels in trends
Sampling and response	Upward bias of 7-15 points for flagged systems in 2022 and 2025; low achievers under-respond	Whether bias changed between cycles; annotate-not-adjust policy	Comparable to a flagged country's own change; about 1-2 points on the OECD average (my calculation)
Exclusions and coverage	Exclusions above 5% and rising in several systems; CI3 as low as 0.32	Net effect on trends	Rising exclusions understate declines; low coverage limits generalisation, not the within-population estimate
Test-taking effort	Differs across countries (20+ points for US in Gneezy et al.; 19-38% of between-country variance observationally); fell 2018-25 on all OECD indicators	Construct or contamination; observational studies disagree at country level	Cross-sectionally up to ~20 points for some countries; for the 2018-25 reading trend, a few points to roughly a third (my bounding)
Mode (paper to computer, 2015)	Raw effects 10-25 points, partly adjusted; residual ~10 in some cases	Adequacy of adjustment	Affects 2012-15 links; not the 2015-25 or 2018-25 comparisons except Viet Nam, Guatemala, Paraguay
Adaptive testing (2018, 2022, 2025)	Reading decline visible in non-adaptive Core; percent-correct and scale changes correlate 0.90-0.91	Unreplicated; no check for science or mathematics	Unlikely to create the reading trend; unknown for science 2025
Platform change 2025	Occurred; timing thresholds re-normalised	No evidence either way	Unknown
Link error	1-4 points for recent links, 6+ to 2012 and earlier; critics say country DIF is ignored	Size of understatement	Perhaps 5-8 points for a country's 2018-25 change; small for the average; science link least secure
Science framework revision 2025	Science link errors ~3x reading	Not analysed in Volume I	Unknown
Individual predictive validity	Level at 15 predicts attainment to 25 strongly, earnings weakly, net of background	Selective attrition; partial controls	Supports the construct's relevance, not any cross-country comparison
National predictive validity	Score-growth correlation robust to updating	Timing and causality	Does not bear on short-run trends

Two things follow. First, no single issue reviewed here is large enough to account for a 22- or 25-point OECD-average decline, and the Annex A1 item-level checks are a genuine, if unreplicated, demonstration that the reading decline appears on every section and item type of a test unchanged since 2018. Second, the same evidence shows that a material part of what PISA reading now measures, and of what has changed, is students' willingness and capacity to engage with a two-hour low-stakes test; that several systems with large reported declines also have the least secure samples; and that the science trend runs across three simultaneous design changes that Volume I does not analyse. The reader must decide how much of “engagement” belongs in “reading,” and should wait for the Technical Report before treating the science comparisons, the 2025 platform, or the flagged samples as settled.

Appendix E. Candidate explanations: predictions, checks and assessment

IN BRIEF (summary added before publication; see Appendix H)

This section tests explanations against the patterns they predict, using checks that can rule them out, not confirm them. Four are consistent with the overall pattern: digital distraction, declining reading habits, falling test effort and, as a residual, the OECD's multi-factor account. They overlap, and their combined size can only be bounded: from about a fifth to most of the reading decline. Four are contradicted as general accounts (the pandemic as main cause, demographic change, spending and cognitive decline), though each survives as a minor contributor or for particular systems. Five fit some countries only. Two cannot be settled: mental health versus digital distraction (though PISA's well-being indicators did not worsen), and whether test changes affected 2025 science results. Digital displacement of reading fits best but has the weakest direct causal evidence at scale; the data cannot separate it from the view that PISA now tests something rich-country 15-year-olds no longer do for free, and the section does not choose.

Scope, method and the pattern to be explained

This section takes each explanation offered for the decline in PISA performance, including every distinct one in the claims inventory (`C/claims_inventory.md`) and several from the research literature, and treats it in three steps. Step one states the explanation and writes down, before consulting the data, what pattern it predicts. Step two checks the predictions against the trend dataset (`A/pisa_trends.csv`), the trajectory groups (`B/trajectory_groups.csv`), the PISA 2025 questionnaire tables, and other assessment series. Step three summarises evidence independent of PISA and assesses how much of the change the explanation could account for. The computations are in `analysis.py`, `timss_compare.py` and `pisa_tables.py`; outputs are in `computed/` (`analysis_log.txt`, `correlations.csv`, `questionnaire_changes.csv`, `oecd_questionnaire_by_country.csv`, `composition_by_country.csv`, `timss_vs_pisa.csv`, `timss_g8_vs_pisa_by_country.csv`, `extra_checks.txt`). "Significant" means the OECD's rule, |estimate/SE| > 1.96 with the link error in the SE. Correlations are Spearman rank correlations across countries unless stated, and are ecological associations.

The pattern to be fitted is the one established in the factual summary, the trajectory analysis and the validity review. It has eleven features.

(F1) Timing. On the OECD-23 aggregate, reading was flat from 2000 to 2012 (within 6 points of 500), then fell every cycle: −4.1 (2012-15), −4.4 (2015-18), −10.9 (2018-22), −16.1 (2022-25). Mathematics was flat from 2003 to 2018, then fell 16.1 to 2022 and 10.9 to 2025. Science peaked in 2009 and fell gradually; its 2022-25 change (−2.8 on OECD-35, SE 3.2) is not significant (Tables I.B1.2a.36-38).

(F2) Domain order. Reading fell most, mathematics next, science least (2022-25 declines in 42, 38 and 11 of 74 systems; Table I.2.6). The science comparison is the least secure: its recent link errors (3.1, 3.5) are three times reading's, and the 2025 science test combined a revised framework, a newly adaptive design and a new platform (technical note; validity review).

(F3) Who fell. The largest declines are in rich, previously high-scoring systems (the Nordic countries, the Netherlands, Germany, Canada, Hong Kong, Latvia, Slovenia, Israel, Malta). Every riser is non-OECD and seven of eight began below 460 in 2018; the 2018 level correlates at −0.40 (reading) and −0.42 (mathematics) with the 2018-25 change across all systems, but only −0.17 and −0.28 within the OECD (`analysis_log.txt`, section 10).

(F4) Socio-economic status. Over 2018-22 the OECD-35 ESCS gap widened (reading 87.9→93.4; mathematics 86.7→93.7); over 2022-25 it narrowed because advantaged students fell more (reading top quarter −20.4, bottom −3.5; mathematics −13.8 and −2.6 n.s.). Over 2015-25 both quarters fell about 20 points in reading and mathematics (Tables I.B1.2b.22-24).

(F5) Distribution. In reading the whole distribution moved down over 2022-25 (p10 −15.5, p90 −12.8); in mathematics the bottom fell more (−13.6 against −5.5); in science only the bottom moved over 2018-25 (p10 −12.6, p90 +2.4 n.s.) (Tables I.B1.2a.39-41).

(F6) Gender. Boys fell more in reading (OECD-35 2022-25: −17.1 against −11.7) and girls more in mathematics (−11.2 against −7.4) (Tables I.B1.2c.26-30).

(F7) Test behaviour. Self-reported effort fell from 7.67 (2018) to 7.47 (2022) and 7.16 (2025) on a 10-point scale (Table I.A1.1, OECD-35); hasty incorrect responses rose over 4 points to about 9% while non-reached items fell; hasty readers on the fluency task rose from 6.6% to 11.4%; declines were larger on long texts and in later test sections (Annex A1, Tables I.A1.7-15).

(F8) No pandemic signature. Closure length does not discriminate between trajectories once income is controlled, nor within the OECD; the shortest closures (Sweden 0 weeks, Iceland 0, Norway 5, Finland 8, Denmark 8) are among the largest decliners (trajectories, section 5). The 2018-22 and 2022-25 changes are uncorrelated (reading rho = 0.00, n = 69; mathematics 0.18; science −0.11), so systems that fell most in the pandemic window neither rebounded nor fell further than others.

(F9) Composition. The OECD attributes about 1.5 of the 25-point OECD-29 reading decline over 2018-25 to the rising immigrant share (Annex A1, p. 311).

(F10) Other assessments. TIMSS 2023 grade 8 (the PISA 2025 birth cohort) shows significant declines since 2019 in 14 of 37 systems and rises in 5, median −4 in mathematics and −8 in science; PIRLS 2021 showed declines in 21 of 32 countries since 2016; NAEP 2024 is below 2019 in every grade and subject; PIAAC

2023 found adult literacy stable or declining in most of 27 countries, with 16-24-year-olds' literacy up only in Norway, Finland and England and down in eight. Sources are given below.

(F11) The questionnaire trends (OECD average, or unweighted mean of OECD members where the OECD row is missing; `analysis_log.txt`, sections 2-6 and 11). More than an hour a day on devices for leisure at school: 34.5% (2022) to 30.7% (2025); for learning at school: 55.8% to 50.0%; more than an hour on social networks on a weekday: 75.1% to 80.1% (Tables I.B1.4.30, I.B1.4.39). Students in schools banning cell phones: 33.6% to 49.4% (Table I.B1.4.49). Sense-of-belonging index: 0.00 (2018), −0.02 (2022), +0.09 (2025); "I feel like I belong at school": 72.6% (2015), 71.0% (2018), 76.2% (2025) (Tables I.B1.3.63-64). Bullying index: 0.00 (2018), −0.30 (2022), −0.07 (2025) (Table I.B1.4.142). Skipped a whole day in the past two weeks: 21.7%, 19.8%, 22.4%; skipped classes: 27.7%, 22.5%, 24.3% (Table I.B1.4.161). Staff-shortage index: 0.03, 0.19, 0.30 (Table I.B1.4.77). Disciplinary climate in science: 0.00 (2015) to 0.14 (2025), an improvement (Table I.B1.4.134). Family-support index: −0.18 over 2022-25 (Table I.B1.4.122).

A caution applies to every questionnaire check. PISA samples a new cohort each cycle. A cross-sectional association between an indicator and a score cannot show that the indicator caused the score: both may be caused by a third thing (socio-economic status, prior achievement, motivation), and the direction can run from achievement to behaviour. The OECD's "after accounting for ESCS" coefficients remove one confounder. A correlation between the change in an indicator and the change in scores across countries holds constant what does not change within a country, but it is an ecological association across 35-75 units, sensitive to a few influential cases, and cannot separate a common cause from an effect. What these checks can do is falsify: if an explanation predicts that an indicator moved in a direction or in a set of countries and it did not, the explanation loses; if it moved as predicted, the explanation survives, which is much weaker.

1. Pandemic school closures and learning loss

Step one. The explanation is that school closures and remote instruction in 2020-21 caused a loss of learning that the students tested in 2022 and 2025 have not recovered. Predictions: declines should appear first and largest in the 2018-22 window and not continue at the same rate afterwards; systems with longer closures should have fallen more; mathematics should have fallen more than reading, as the meta-analytic evidence finds; disadvantaged students should have fallen more, because remote learning depended on home resources; the same cohorts on other assessments (TIMSS 2023 grade 8, which sampled the PISA 2025 birth cohort; PIRLS 2021; NAEP) should show declines of comparable size; and systems that recovered on national assessments should have recovered on PISA.

Step two. The timing prediction is half met. The largest single-cycle decline in OECD-23 mathematics is in 2018-22 (−16.1), but 2022-25 (−10.9) is not much smaller, and in reading the 2022-25 decline (−16.1) exceeds the pandemic window (−10.9). A one-off unrecovered loss would show as a step, not an accelerating slide. The closure-length prediction fails: across systems with UNESCO data, full-closure weeks correlate positively with the 2018-22 change (reading rho = +0.24, n = 65; mathematics +0.25), near zero within the OECD (+0.09, +0.14), and the coefficient is +0.14 points per week (SE 0.13) with log GDP controlled, −0.05 (SE 0.18) within the OECD (`analysis_log.txt`, section 8). The 2018-22 and 2022-25 changes are uncorrelated (F8): no system that fell in 2018-22 recovered in proportion. The domain prediction fails (reading fell more than mathematics in every window since 2018). The socio-economic prediction holds for 2018-22 (gap widened 5-7 points) and fails for 2022-25 (advantaged students fell more, F4).

The cross-assessment check is the most informative. TIMSS 2023 grade 8 tested the 2009 birth cohort, the same cohort PISA tested in 2025. Its 2019-23 changes (computed from Exhibits 1.2.10 and 2.2.10, https://timss2023.org/results/grade-8-math-achievement-trends/; `computed/timss_trends.csv`) were significantly negative in 14 of 37 systems in mathematics and 15 in science, positive in 5 and 4, with median changes of −4 and −8. Across 29 systems in both studies the TIMSS mathematics change correlates with the PISA mathematics change 2018-25 at rho = 0.26 (p = 0.17); within the 16 OECD members at 0.00. The divergences are largest for the systems at the centre of the PISA story: Sweden's grade-8 mathematics rose 14 points on TIMSS while its PISA mathematics fell 38 over 2018-25; Finland −5 against −39; France −4 against −38; Norway (grade 9) −2 against −49; Japan +1 against −2 and Israel −32 against −30 (consistent); the United States −27 against −15. Over 2015-23 the median TIMSS grade-8 mathematics change in the 21 systems common to both studies was +3 against a median PISA change of −8 over 2015-25 (`computed/extra_checks.txt`). TIMSS grade 4 (born about 2013) relates more strongly to the PISA 2022-25 change (mathematics rho = 0.42, n = 39, p = 0.008; science 0.46, p = 0.003), which matches the OECD's finding that PIRLS 2016-21 changes correlate at 0.34 with PISA 2022-25 changes (Annex A1, p. 311). Something happened to fourth-graders around 2019-23 that is visible in PIRLS, TIMSS and later PISA; but the eighth-grade cohort's curricular mathematics did not fall in most of the European systems whose PISA mathematics fell 30-50 points.

Step three. The best evidence on pandemic learning loss is Betthäuser, Bach-Mortensen and Engzell's meta-analysis (Nature Human Behaviour, 2023, https://www.nature.com/articles/s41562-022-01506-4): 42 studies, 15 countries, pooled deficit d = −0.14 (about 14 PISA points), larger in mathematics than reading by 0.07 SD, larger in middle-income countries and among disadvantaged children, with no recovery to mid-2022. NAEP 2024 remains below 2019 in every grade and subject (https://www.nagb.gov/powered-by-naep/the-2024-nations-report-card/10-takeaways-from-2024-naep-results.html). Germany's IQB-Bildungstrend 2024 found ninth-grade mathematics down 24 points since 2018, with the share below the intermediate-certificate minimum standard up from 22% to 34%, and names the pandemic a "possible reason" alongside composition (https://deutsches-schulportal.de/bildungswesen/iqb-bildungstrend-die-wichtigsten-ergebnisse/). Volume I states that "there is no simple relationship between the duration of closures and subsequent reading trends" (p. 99) and suggests the pandemic "may have acted as a catalyst" for shifts in habits and values.

Assessment. It is established that the pandemic produced deficits of 0.1-0.2 SD, larger in mathematics and among the disadvantaged, incompletely recovered. It is established that the PISA pattern lacks the pandemic's signature: the decline predates 2020 in reading, accelerated after 2022, is larger in reading, is concentrated among advantaged students in 2022-25, is unrelated to closure length, and is not mirrored by TIMSS grade 8 where PISA fell most. A pandemic deficit of 10-15 points, incompletely recovered, is plausibly present in the 2018-22 mathematics change (−16) and consistent with the ESCS widening in that window; that would make the pandemic responsible for perhaps a third to a half of the OECD-average mathematics decline since 2018 and less of reading. It cannot explain the pre-2018 decline, the 2022-25 acceleration, the concentration among advantaged students, or the cross-country distribution. The “catalyst” formulation is speculation, not testable here. The absence of a closure gradient shows only that the marginal effect of additional weeks is invisible at country level with a coarse measure.

2. Smartphones, social media and digital distraction

Step one. The explanation is that the diffusion of personal smartphones and attention-maximising social media among adolescents from about 2012 has reduced sustained attention, displaced reading and homework, disrupted sleep and lowered performance. Predictions: declines should begin in the early-to-mid 2010s where adolescent smartphone ownership saturated first (the Nordic countries, the Netherlands, Korea, Japan, the United States, Australia) and later in poorer countries; reading should be affected most, science least; heavier users should score lower after controlling for socio-economic status; countries and cycles with more use should have fallen more; declines should be at least as large among advantaged students, whose access is universal; test-taking behaviour should show more hasty responding and less persistence on long texts; the effect should appear on other reading-intensive assessments and in adult cohorts that grew up with smartphones (PIAAC 16-24-year-olds); and school phone bans should be followed by gains if in-school use is the mechanism.

Step two. The timing prediction holds in outline: OECD-23 reading turned down after 2012, when adolescent smartphone ownership passed half in wealthy countries; mathematics after 2015-18. The country-order prediction holds partly: the earliest and largest reading declines are in wealthy early adopters (F3) and the risers are mostly poorer late adopters; but Japan and Korea, early adopters with near-universal ownership, have the smallest declines among wealthy systems (reading −1 and −13 over 2018-25; mathematics −2 and −4), and Singapore and Chinese Taipei declined little or rose. The domain order (F2), the behavioural signature (F7) and the 2022-25 concentration among advantaged students (F4) all hold, and the last is predicted by few other accounts. The within-country prediction holds cross-sectionally: an extra hour a day of leisure device use at school is associated with −9.7 science points before and −8.3 after accounting for socio-economic profile (Table I.B1.4.34), negative and significant after adjustment in 88 of 89 systems, non-linear (up to one hour +5.3 relative to none; each further step −9 to −25). This is a universal association that cannot separate distraction from selection of weaker or less motivated students into heavier use.

The predictions about change fail or are untestable. Between 2022 and 2025 the OECD-average share reporting more than an hour of leisure device use at school fell from 34.5% to 30.7% and the share in phone-banning schools rose from 33.6% to 49.4%, while reading fell 14 points. Across systems, the change in leisure use at school correlates with the reading change at rho = +0.21 (p = 0.09), the wrong sign for the hypothesis; the change in the banning share at −0.03 (reading) and +0.08 (science); the 2025 level of bans at 0.00-0.10 with every change; the within-country difference for students in banning schools is −6.0 before and −4.2 after ESCS adjustment (Table I.B1.4.50), significant in 21 systems and positive in 10, most plausibly because schools with problems adopt bans. The change in the share spending more than an hour on social networks on a weekday (33-36 systems) correlates positively with the reading change (rho = +0.26; OECD +0.22). The quantities the hypothesis is about, hours of out-of-school smartphone use and age at first smartphone, are not in the PISA 2025 trend tables. Other assessments are mixed: TIMSS grade-8 mathematics did not fall in Sweden, Finland or France (section 1), which a reading-focused distraction account allows and a general cognitive-displacement account does not; PIAAC 2023 found the literacy of 16-24-year-olds rising only in Norway, Finland and England and falling in eight countries (https://ilsa-gateway.org/studies/results/1595), consistent with a smartphone-cohort effect and with several other accounts. The gender pattern (boys fell more in reading, girls in mathematics) does not map onto gendered device use (girls social media, boys games) without two separate mechanisms.

Step three. The causal literature is smaller than the debate suggests. Beland and Murphy (Labour Economics, 2016, https://www.lpbeland.com/uploads/7/8/7/5/7875420/lpblabour_1-s2.0-s0927537116300136-main.pdf), using staggered school phone bans in four English cities, found GCSE scores 6.4% of an SD higher after a ban, 0.14 SD in the lowest prior-achievement quintile and nothing at the top. Kessel, Hardardottir and Tyrefors (Economics of Education Review, 2020, https://www.ifn.se/en/publications/scientific-articles-in-english/2020/2020-56/) replicated the design in Sweden and found no effect, rejecting even small gains. Abrahamsson (SSRN 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4735240) found Norwegian middle-school bans improved girls’ GPA and mental health and reduced bullying. Figlio and Özek (NBER WP 34388, https://www.nber.org/system/files/working_papers/w34388/w34388.pdf), using Florida’s 2023 ban, found second-year test gains of 0.6-1.1 percentile points (about 0.02-0.03 SD), a 5-10% fall in unexcused absences and a first-year rise in suspensions. The ban evidence supports in-school effects of 0.02-0.14 SD, largest for low achievers, the wrong end of the distribution for 2022-25 and an order of magnitude short of a 25-point average decline; bans address only in-school use. Out-of-school evidence is observational: Orben and Przybylski (Nature Human Behaviour, 2019, https://www.nature.com/articles/s41562-018-0506-1) found technology use explaining at most 0.4% of variance in adolescent well-being; Odgers and Jensen (Journal of Child Psychology and Psychiatry, 2020) reached similar conclusions; WHO/HBSC found problematic social-media use up from 7% (2018) to 11% (2022) (https://www.who.int/europe/news/item/25-09-2024-teens--screens-and-mental-health). None of the digital-resource items in the OECD’s own table of system-level correlations among changes (Table I.B1.4.189) is significant for reading or mathematics.

Assessment. The explanation is consistent with more of the descriptive pattern than any other single account and is contradicted by nothing decisive in PISA, because PISA does not measure what it is about. Against treating that as strong support: the direct causal estimates (bans) are small and concentrated among low achievers; the within-PISA change measures moved in the benign direction between 2022 and 2025 while scores fell, and do not correlate with score changes across countries; and the exceptions (Japan, Korea, Singapore, Chinese Taipei) are saturated smartphone societies with small or no declines, so the mechanism, if real, is conditional on something else, plausibly out-of-school study or what adolescents do with the devices. Displacement of reading and sleep could account for a material share of the reading decline, perhaps 5-15 of 25 points if the largest ban estimates applied across the distribution, but that is an extrapolation; the cross-country evidence is compatible with anything from near zero to most of the decline. The claim that screens are “the poison” and the claim that Singapore refutes the screen hypothesis (claims inventory B1.2, B1.12) are both overstated.

3. Declining reading for pleasure and changing reading habits

Step one. The explanation, distinct from the previous one in mechanism, is that adolescents read fewer and shorter texts, especially fiction and print, and that comprehension develops through volume of sustained reading. Predictions: reading should fall more than mathematics and science; the decline should be visible on long texts and higher-order items more than on locating information; the 2009-18 fall in reading enjoyment reported in PISA 2018 should be followed by score declines in the same countries, and systems where enjoyment fell most should have fallen most; the effect should be larger for boys, who read less for pleasure; and it should carry through to adult reading (PIAAC) as cohorts age.

Step two. The domain and item-type predictions hold (F2, F7): long-text items fell most, and over 2018-22 evaluating and reflecting items fell most while locating-information items fell least (Annex A1, Tables I.A1.8-10). The gender prediction holds in reading (F6) and the account does not predict the mathematics gender pattern. The check that matters most, the correlation between the change in reading enjoyment and the change in scores, cannot be done in the 2025 data: the reading-habits questions were asked in 2009 and 2018, and the 2025 questionnaire carries no enjoyment-of-reading index (the only enjoyment measure is enjoyment of science, Tables I.B1.3.53-54). Volume I reports that the enjoyment index "declined over the decade to 2018" across OECD countries, that students read fewer fiction books, magazines and newspapers, and that parents' reading also declined (p. 98, citing OECD 2021, *21st-Century Readers*); I could not retrieve the country-level 2009-18 changes (the OECD PDF is blocked to automated fetching), so the country-order prediction is untested here. The PIAAC prediction is consistent with what is reported.

Step three. The longitudinal evidence that reading volume causes reading skill is moderate. Torppa and colleagues (Child Development, 2020, https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/cdev.13241), following about 2,500 Finnish children from grade 1 to 9, found reciprocal effects between leisure reading and comprehension, but only for book reading, not digital or magazine reading. Sullivan and Brown (British Educational Research Journal, 2015, https://bera-journals.onlinelibrary.wiley.com/doi/10.1002/berj.3180), using the 1970 British Cohort Study, found reading for pleasure at 10 and 16 predicted vocabulary and mathematics progress net of background. Mol and Bus (Psychological Bulletin, 2011) found print exposure explaining 12-13% of comprehension variance in adolescence. On trends, the American Time Use Survey shows daily reading for pleasure falling about 40% over two decades (https://pmc.ncbi.nlm.nih.gov/articles/PMC12496190/) and PIRLS 2021 found fewer fourth-graders who "very much like reading" than in 2016 in most countries (https://pirls2021.org/results). Taking the within-country PISA 2018 difference of 30-40 points between frequent and non-readers after ESCS adjustment and a shift of about 10 percentage points in shares gives 3-4 points from composition alone.

Assessment. The mechanism is well supported and fits reading; the sizes the independent evidence supports are modest, perhaps 3-8 of the 25 reading points as a direct composition effect, more if fragmented digital reading lowers the return to reading time, which Torppa's result suggests but does not quantify. The explanation says nothing about mathematics, where the 2015-25 decline (−22) is nearly as large, unless PISA mathematics items are treated as reading-intensive; that is supported by the 0.79-0.82 correlation of reading and mathematics trends (Annex A1, p. 309) and by the TIMSS divergence, but it turns a reading-habits explanation into a claim about the instrument (section 10).

4. Changes in the composition of the student population

Step one. The explanation is that rising immigration, more students with a home language other than the test language, and other demographic shifts have lowered average scores without any change in what schools do. Predictions: the decline should be concentrated in systems where the immigrant share rose most (Portugal, Germany, Norway, Ireland, Slovenia, Iceland, Austria, Switzerland, France, on the OECD's list, p. 311); non-immigrant students' scores should be stable or should fall much less than the all-student mean; the bottom of the distribution should fall more than the top, because immigrant students are over-represented among low performers; the ESCS gap should widen rather than narrow; systems with negligible immigration (Japan, Korea, Poland, Finland at 8%, Latvia at 5%) should be stable; and other assessments in the same countries should show the same composition-driven pattern.

Step two. The first prediction holds weakly across all systems and not within the OECD: the 2025 immigrant share correlates at −0.32 with the 2018-25 change across 74 systems (Pearson −0.13) and −0.17 within the OECD (p = 0.32); the change in the share at −0.26 (trajectories, Table 9). The second prediction fails. Using the OECD's means by immigrant background (Tables I.B1.2d.7-9; `nonimmigrant_mean` in the CSV), non-immigrant students' reading fell 13.0 points on average across 73 systems over 2018-25 against 17.9 for all students, and mathematics 11.0 against 14.6; the two changes correlate at 0.97-0.98 (`computed/composition_nonimmigrant_vs_all.csv`). On the OECD-29 aggregate non-immigrant reading fell from 498.6 to 476.2 (−22.4) against −24.6 for all students on OECD-35. Native-born students of native-born parents account for roughly nine-tenths of the average decline. The country table (`computed/composition_by_country.csv`) shows where the difference is larger: Germany (all students −33.0 in reading, non-immigrant −20.4; immigrant share 22%→29%), Israel (−34.6 against −19.4), Iceland (−51.7 against −41.0), Latvia (−49.0 against −38.8), France (−36.3 against −28.1), Portugal (−29.8 against −22.2; share 7%→18%), Slovenia (−50.6 against −42.5). In Sweden the difference is 4.6 of 39.4 points, the Netherlands 1.0 of 43.4, Finland 5.3 of 45.8, Norway 6.5 of 46.6, Denmark 3.4 of 41.1; in the United Kingdom, Hong Kong and Luxembourg natives fell more than all students. Two cautions: the difference is not a pure composition effect, since it includes any larger decline among immigrant students themselves (substantial in Germany), and the background means use the sub-sample with valid data. The distributional and gap predictions fail for 2022-25 (F4, F5). The low-immigration prediction fails: Latvia (5%) and Finland (8%) are among the largest decliners.

Step three. The OECD's decomposition (Annex A1, p. 311) puts the mechanical effect of a 3.8-point rise in the immigrant share at a 40-point gap at 1.5 of the 25-point OECD-29 reading decline and 3 of Portugal's 30. Sweden has the most detailed evidence: Skolverket attributed 15-25% of the 2000-12 PISA decline to composition (https://skolvarlden.se/artiklar/skolverket-invandring-paverkar-skolresultaten), and Heller Sahlgren's IFN re-analysis (Policy Paper 71, https://www.ifn.se/media/bn3m5vo3/pp71.pdf) put it at 29% for 2000-12 and 43% for 2009-12, of which 11 points came from the falling share of Swedishheritage students and 18 from larger declines among the others. Germany's IQB 2024 report notes the first-generation share of ninth-graders rose from 6% to 13% over 2018-24, that foreign-born students score 78 points lower in mathematics, and attributes part of the decline to composition without a share (https://deutsches-schulportal.de/bildungswesen/iqb-bildungstrend-die-wichtigsten-ergebnisse/). The German PISA team's statement that the dividing line runs between social classes rather than migration backgrounds (B9.2) concerns levels, not trends.

Assessment. Composition explains a small share of the OECD-average decline (1.5-3 points) and a larger but minority share in a few systems with rapid recent immigration: roughly a third of Germany's reading decline and two-fifths of Israel's on the native/all-student comparison, a fifth to a quarter in Portugal, France and Iceland, a tenth or less in Sweden, Norway, Denmark, Finland and the Netherlands. The claim that the link with asylum migration is "unmistakable" (B9.1) is contradicted by the native-student series wherever it has been made, and the claim that composition is irrelevant is contradicted by Germany and Israel. The explanation is consistent with part of the pattern in some countries and inconsistent with the overall pattern (F3-F5).

5. Changes in curriculum, standards and pedagogy, country by country

Step one. The generic claim is that changes in what is taught and how (skills-based or competence-based curricula, reduced content, phenomenon-based or inquiry learning, less explicit instruction, less homework, "inclusion" without support, grade inflation) have lowered achievement, and the converse claim is that knowledge-rich curricula, phonics and explicit instruction have protected some systems. These claims are only testable country by country, because the reforms differ, and the predictions have to respect timing: a reform affects the PISA cohort that passed through the reformed grades, so a curriculum introduced in year t at grade g reaches PISA at about year t + (10 − g) at the earliest and shows fully only for cohorts schooled entirely under it. The generic prediction is that systems that moved towards skills-based curricula in the 2010s should have fallen when the affected cohorts reached 15, and systems that moved the other way should have held; that the decline should be visible on curriculum-based tests (TIMSS, national assessments) as well as PISA; and that it should be concentrated in the domains the reform touched.

Step two. The country checks use the trajectories in `trajectories.md`, Tables 3-5, the TIMSS comparison and the national assessments cited.

England. The 2010-14 reforms (synthetic phonics with a check from 2012, a knowledge-based national curriculum from 2014, mathematics mastery, the EBacc, Progress 8) reached the PISA cohort fully from about 2022. The United Kingdom's reading was 495 (2006), 504 (2018), 494 (2022), 494 (2025); mathematics 495, 502, 489, 488; science 515, 505, 500, 511. Its 2018-25 changes (−10, −14, +7) are among the smallest in the OECD, and TIMSS grade-8 mathematics for England rose 10 points over 2019-23. This is consistent with the reforms having protected England. The sampling record gives a different shape. England's 2022 sample was estimated to be biased upward by 7-8 points (validity review), so its true 2018-22 fall was larger than published (about 17 points in reading and 20 in mathematics on the UK figures, against 10.9 and 16.1 on the OECD-23 average) and its flat 2022-25 figures conceal a recovery of about 6-8 points while the OECD-23 average fell 16.1 and 10.9; on that reading England was not protected through the pandemic but recovered afterward. The 2018-25 changes do not involve the 2022 sample. The coverage index (0.85 in 2018, 0.97 in 2022, 0.85 in 2025) marks 2022 as the anomalous sample; if the 2022 rise in coverage were real it would have brought in more low scorers and pulled that year's mean down, the opposite direction to the non-response bias. Machin, McNally and Viarengo's evaluation of synthetic phonics (AEJ: Economic Policy, 2018, https://www.aeaweb.org/articles?id=10.1257%2Fpol.20160514) found initial effects that faded on average by age 11 but persisted for children prone to struggle, which predicts narrowing at the bottom rather than a general lift. Scotland (Curriculum for Excellence from 2010) and Wales (new curriculum reaching PISA only after 2025) declined more over 2012-22; Scotland's 2022-25 change is flat. The within-UK contrast is the strongest natural experiment in the set, confounded by sampling (Scotland's 2022 respondents had higher prior attainment than non-respondents) and by differences in immigration, funding and poverty.

Sweden. The 2011 curriculum and grading reform (Lgr11) was followed by a 2012-18 recovery (reading +23, mathematics +24) and then falls of 39 and 38 over 2018-25. TIMSS grade-8 mathematics rose 14 over 2019-23 and 16 over 2015-23. No curriculum change coincides with 2018, and curricular mathematics shows no decline.

Finland. The 2016 core curriculum ("phenomenon-based" modules, more self-direction) schooled the 2025 cohort entirely. But Finland's decline began with the 2009 cohort (reading 547→536), continued through 2012, 2015 and 2018 (524, 526, 520) before the curriculum could affect any PISA cohort, and then accelerated (490, 474). TIMSS grade-8 mathematics fell 5 over 2019-23. The curriculum can account at most for the post-2018 acceleration; Finland's PISA director attributes the trend to reading habits and perseverance (claims inventory B4.4), and Finland has no national assessment that could corroborate the acceleration (B13.5).

Norway. The 2020 curriculum (LK20) reached the 2025 cohort for its last five years; digitalisation is treated in section 6. Reading was flat at 500-513 from 2000 to 2015, then 499, 477, 453; TIMSS grade-9 mathematics fell 2 over 2019-23. LK20 is too recent to explain the 2018 fall, and the 2025 exclusion rate of 10.4% works against the reported decline.

Poland. The 2017-19 abolition of gimnazja reached the 2022 cohort in transition and the 2025 cohort fully. Mathematics fell 27 over 2018-22 and 5 (n.s.) over 2022-25; reading −23 and −7. The timing fits reform disruption (B12.13) and equally the pandemic (long closures); TIMSS grade-4 mathematics rose 26 over 2019-23, the largest rise in the set, which does not fit general disruption.

Japan and Korea. Japan's reduced-content "yutori" curriculum coincided with its 2003-06 dip and the post-2011 content restoration with its 2009-12 recovery; the 2020-22 "active learning" courses reached the 2025 cohort late. Japan rose in 2018-22 (reading +12, science +17) and fell back in 2022-25 (−13, −10, −9; reading significant) while its coverage index fell to 0.86. Korea fell sharply in 2012-15 (reading −19, mathematics −30) and has been flat since. Neither maps onto a curriculum change as the "active learning" critics claim (B12.1).

France. The 2016 collège reform reached the 2025 cohort; the 2017-19 halving of CP/CE1 classes in priority areas did not. Reading was 505 (2012), 499, 493, 474, 456; TIMSS grade-8 mathematics fell 4 over 2019-23. Nothing in the reform timing fits the post-2018 acceleration. Estonia has had no major curriculum change and a gradual reading decline (523 to 499), science flat; the 2024 Russian-school transition is too recent to matter.

Australia and New Zealand. Australia's declines (reading 528 in 2000 to 491; mathematics 524 in 2003 to 478) predate the 2012 national curriculum and continued through it; TIMSS grade-8 mathematics fell 8 over 2019-23. New Zealand's structured-literacy policies date from 2024 and cannot have affected the 2025 cohort; its 2022-25 changes are non-significant on a sample the OECD says may be over-estimated by 10-15 points.

Türkiye. The alignment of the high-school entrance examination (LGS) with PISA-style items from 2018 (B12.14) preceded a 2022-25 rise of 8-18 points in all domains on flat coverage (0.72-0.74). This is the clearest case of a national policy that predicts its PISA trend, and it is a policy about testing format rather than curriculum content.

Step three. On phonics and explicit instruction, the quasi-experimental evidence (Machin, McNally and Viarengo 2018; the Education Endowment Foundation toolkit) supports early decoding gains and persistent gains for struggling readers, not a general lift in adolescent comprehension. The PISA 2015 finding that inquiry-based science teaching is negatively associated with science scores after controls (PISA 2015 Results Volume II, ch. 2) is cross-sectional. No comparative study codes curriculum "knowledge-richness" across the 2010s against PISA changes; the trajectory analysis found tracking age and governance structure do not discriminate between groups.

Assessment. The reform-timing test gives one case where a reform predicts the trend at the right time (Türkiye), one consistent with reforms protecting a system but confounded by sampling (England), one where a reform and the pandemic are equally consistent (Poland), and cases where the decline began before the blamed reform (Finland), continued through the credited reform (Australia), or has no reform at the right time (Sweden, Norway, France, Estonia). The TIMSS divergence is the most awkward fact for the curriculum account: a curriculum that had stopped teaching mathematics would show on the curricular test, and in Sweden, Finland and France it did not. Reform and pedagogy plausibly explain a few points in particular systems and the direction of England's and Türkiye's departures from the average; they do not explain a synchronised decline across systems with opposite reform histories. That is not evidence that curriculum is unimportant for levels, only that it does not fit the trend.

6. Digitalisation of schooling itself

Step one. The explanation is that the replacement of textbooks, handwriting and teacher-led instruction by tablets, laptops and platforms, most fully in the Nordic countries from about 2014, has reduced learning; the medium is instructional rather than recreational. Predictions: the largest declines should be in the earliest and fullest one-to-one adopters (Denmark, Norway, Sweden, Iceland, Finland, the Netherlands, Estonia, Australia, New Zealand, the United States) and the smallest where print was kept (Japan, Korea, Germany, Italy); heavier learning-device users should score lower after controls; systems that reduced instructional device use over 2022-25 should have improved; mathematics and science should fall as much as reading; and the effect should show on curriculum-based tests.

Step two. The country-order prediction holds for the Nordic countries and the Netherlands, which lead the declines, but fails for Japan (small decline, 13% of students with more than an hour of leisure device use at school), Germany (large decline, the lowest learning-device use in the OECD in 2022 at 36%), and Estonia and Australia (high use, moderate decline). The 2025 share with more than an hour a day of learning-device use at school correlates with the 2018-25 reading change at rho = 0.00 across 73 systems and −0.07 within the OECD (Pearson −0.17, −0.23), and with mathematics at −0.15 and −0.10 (Pearson −0.27, p = 0.02 across all; −0.23 within the OECD). The school digital-preparedness index correlates with the 2018-25 reading change at −0.30 within the OECD (p = 0.07; Pearson −0.33, p = 0.045). The 2022-25 change in learning-device use correlates with the 2018-25 mathematics change at −0.23 (p = 0.07) and with the others between −0.03 and −0.10. These are weak negative associations that mostly do not reach significance. The within-country coefficient for learning-device use (Table I.B1.4.32) is +1.0 points per hour before and −0.4 after ESCS adjustment, significantly negative in 51 systems and positive in 15, non-monotonic (up to one hour +13.0 relative to none; more than five hours −11.4). The domain prediction fails (F2). Sweden, Finland and Norway, the most digitalised systems, did not fall on TIMSS grade-8 mathematics.

Step three. The best causal evidence comes from the Nordic one-to-one programmes themselves. Hall and Lundin's staggered difference-in-differences on 57,000 Swedish primary students in 270 schools adopting one-to-one computers over 2009-20 (IFAU WP 2023:3, https://www.ifau.se/globalassets/pdf/se/2023/wp-2023-3-teaching-technology-and-test-scores.pdf; Economics of Education Review, 2024) found no average effect on national tests in mathematics, Swedish or English, small positive effects for high-SES students and none for low-SES students; Hall, Lundin and Sibbmark (Labour Economics, 2021, https://ideas.repec.org/a/eee/labeco/v69y2021ics0927537120301615.html) found the same for secondary programmes. The wider literature (Bulman and Fairlie's 2016 Handbook review; Cristia et al. on Peru, AEJ: Applied, 2017) finds device-provision effects clustered around zero. The OECD's *Students, Computers and Learning* (2015) found no PISA improvement in heavy ICT investors and a negative association at high use. Sweden's 2023 pivot to print (https://eurydice.eacea.ec.europa.eu/news/swedish-digitalisation-strategy-school-systems-led-results) and the Norwegian minister's "catastrophe" (B2.1) are policy positions, the latter citing no study.

Assessment. The causal evidence supports the claim that one-to-one programmes did not raise achievement and may have widened gaps, not that they lowered average achievement by tens of points; the Swedish estimates are null on the same national tests that show no TIMSS decline. The PISA associations are weakly negative and not robust. The explanation fits the Nordic geography and not Germany, Japan, the domain ordering or TIMSS. A bound from the largest credible negative estimates (−0.05 SD for heavy unsupported use) is a few points in the most digitalised systems and about zero elsewhere. A larger effect through screen reading of long texts (Delgado et al.'s meta-analysis, Educational Research Review, 2018, finds screen inferiority of about 0.2 SD for expository comprehension under time pressure) is possible, and is the reading-habits explanation applied to school.

7. Teacher supply, qualifications and turnover

Step one. The explanation is that teacher shortages, a growing share of unqualified or out-of-field teachers, and higher turnover have lowered instructional quality. Predictions: the staff-shortage index should have risen, and risen most in the systems that fell most; the share of fully certified teachers should have fallen in the decliners; student-teacher ratios should have risen; the effect should be largest for disadvantaged students, who are concentrated in the schools that shortages hit; it should be present in all domains; and it should be visible on curriculum-based tests.

Step two. The OECD-average staff-shortage index rose from 0.03 (2018) to 0.19 (2022) and 0.30 (2025) (Table I.B1.4.77). Its 2018-25 change correlates with the reading change at −0.25 ($n = 69$, $p = 0.04$) and mathematics at −0.27 ($p = 0.02$) across all systems, in the predicted direction; within the OECD −0.08 and 0.00. The share of fully certified teachers was 82.7% (2015), 81.7% (2018), 83.3% (2025), and its change correlates at +0.12 to +0.23 with score changes (predicted sign, not significant; Table I.B1.4.112). The student-teacher ratio fell from 13.3 to 12.7 (Table I.B1.4.82), against the prediction. The distributional prediction fails for 2022-25. The OECD's own correlations of changes (Table I.B1.4.189) give the staff-shortage change −0.10 ($p = 0.56$) with science and −0.08 with mathematics among OECD countries. The TIMSS divergence applies again.

Step three. Uncertified, out-of-field or novice teachers are associated with lower achievement growth of 0.03-0.10 SD per year in North Carolina administrative data (Clotfelter, Ladd and Vigdor, Journal of Human Resources, 2010). The OECD reports 40% of students in 2025 in schools whose principals say shortages hinder instruction (claims inventory B11.1) and TALIS 2024 reports class time keeping order up from 13% to 16% (Volume I, p. 97). If a third of students had a materially less qualified teacher for a third of their instruction at 0.05 SD per year, the cumulative effect over three years is 0.02-0.05 SD, or 2-5 points.

Assessment. Shortages rose where scores fell across all systems but not within the OECD, and the plausible aggregate effect is a few points. The explanation cannot be distinguished from a common cause (a system under stress reports shortages and scores lower); the falling student-teacher ratio and flat certification argue against a large supply effect; and the largest decliners include systems with the highest teacher pay and qualification standards (Finland, the Netherlands, Denmark).

8. Changes in school accountability and testing regimes

Step one. The explanation has two opposite versions. One holds that the weakening of external accountability (fewer standardised tests, less use of data, the abolition of league tables, grade inflation) removed pressure to maintain standards; the other holds that high-stakes accountability narrowed teaching and harmed deeper learning. Predictions for the first: systems that reduced external testing or its consequences in the 2010s should have fallen and systems that maintained or strengthened it (England, Türkiye's examination alignment) should have held; principals' use of external evaluation data should have declined in the decliners; the effect should be broad across domains. Predictions for the second: systems with the strongest high-stakes regimes (the United States under NCLB/ESSA, England, Korea, Singapore) should have fallen most.

Step two. The trajectory analysis found no relationship between the share of decisions taken at school or central level and any change (rho −0.03, +0.09); the sustained decliners include the most centralised (Greece, France, Portugal) and the most decentralised (the Netherlands, Iceland, Latvia) systems. Volume I reports a decline in principals' use of evaluation data (Table I.B1.4.178), notes that Norway reduced it, but that Türkiye did too and the pattern "does not appear to be consistently associated with more negative trends" (p. 98). The strong-accountability version is contradicted by England, Singapore and Korea (small declines); the weak-accountability version by Sweden (expanded national testing and central grading in the 2010s, still −39) and Denmark (national tests from 2010, large declines). Türkiye's rise after aligning its high-stakes examination with PISA-format items fits the first version, and equally the narrower reading that students practised PISA-style items.

Step three. Bergbauer, Hanushek and Woessmann's panel of 59 countries over PISA 2000-15 (Journal of Human Resources, 2024, https://jhr.uwpress.org/content/59/2/349) found that expanding standardised testing with external comparison raised achievement, internal testing without comparison had little effect, and standardised monitoring helped only low-performing systems. Dee and Jacob (Journal of Policy Analysis and Management, 2011) find NCLB gains in mathematics, not reading. Burgess, Wilson and Worth (Journal of Public Economics, 2013) found that Wales's 2001 abolition of league tables lowered GCSE performance by about 0.09 SD, concentrated in the lower half. Large regime changes move scores by 0.05-0.10 SD.

Assessment. Accountability changes of the size that occurred within the OECD in the 2010s could account for a few points in particular systems in either direction; no coding of regime change fits the cross-country pattern, the two versions cancel across the decliners, and neither predicts the 2022-25 concentration among advantaged students. Consistent with some countries, not with the overall pattern.

9. Declining student effort and motivation on a low-stakes test

Step one. The explanation is that students try less hard on PISA than they did, so that part of the measured decline is test-taking behaviour rather than skill. It is close to, but not the same as, the OECD's account, which treats declining engagement as part of the construct. Predictions: self-reported effort should have fallen, most where scores fell most; hasty responses should have risen; the decline should be larger in later test sections and on items requiring sustained work; it should be larger on PISA than on tests with consequences or shorter curriculum tests (TIMSS); it should be present in all domains sharing the session; and it should be larger among students with least to gain from trying.

Step two. Most predictions hold. Effort fell in 66 of 71 systems (Table I.A1.1); the 2018-25 effort change correlates with the 2018-25 reading change at rho = 0.32 across 69 systems ($p = 0.008$) and 0.46 within the OECD ($p = 0.004$), with mathematics at 0.35 and 0.52, and with science at 0.20 and 0.44; the 2022-25 effort change correlates with the 2022-25 reading change at 0.49 ($p < 0.001$) (`analysis_log.txt` , section 1). Levels are uninformative (effort and reading levels correlate at −0.51, because the highest-scoring systems report the lowest effort). An OLS regression of the reading change on the effort change gives 17 points per effort point (SE 7.3), which at the OECD-35 effort change of −0.52 attributes 8.8 of the 24.6-point decline to effort; this ecological slope is an upper bound on what the self-report can carry and is inflated by any common cause. The behavioural predictions (F7) hold. The cross-test prediction holds strongly for the Nordic and French cases and not for the United States or Israel: TIMSS grade-8 mathematics, a 90-minute curricular test of the same cohort, did not fall in Sweden, Finland, France or Norway while PISA mathematics fell 38-49 points; it fell in Israel (−32) and the United States (−27) where PISA fell 30 and 15 (section 1). All three domains show the same behavioural shift. The OECD reports that “a group of students from relatively affluent families, but who struggle academically, are contributing most to the overall decline” (Annex A1, p. 316), which fits disengaged students with nothing at stake, and the 2022-25 concentration among advantaged students fits this account better than most alternatives.

Step three. The validity review covers the literature: Gneezy and colleagues’ experiment (AER: Insights, 2019) in which a financial incentive raised US students’ scores on a PISA-style test by about 0.2 SD with no effect in Shanghai; Borghans and Schils’ finding that the within-test decline explains 19% of between-country variance; Zamarro, Hitt and Mendez’s 32-38%; Michaelides and Ivanova’s more reassuring response-time evidence; and a 2026 multilevel study finding no country-level effort-achievement association in 2022. Borgonovi, Ferrara and Piacentini (Journal of Adolescence, 2021) show that the within-test decline predicts tertiary completion independently of the score, so what looks like effort is partly a stable trait with consequences. Volume I’s position is that attention and persistence are part of what reading is (p. 319). The new piece of evidence is the TIMSS divergence: two low-stakes tests of the same cohort, one two hours and reading-heavy, the other 90 minutes and curricular, diverged by 30-50 points in the systems with the largest PISA declines.

Assessment. It is established that test-taking behaviour on PISA changed in a way that lowers scores, that it changed most where scores fell most, and that the change is visible in every section and domain. What is contested is whether this is measurement or construct. The validity review’s bound (a few points to roughly a third of the reading decline since 2018) stands; the effort regression gives about 9 points from the self-report; the TIMSS divergence suggests the PISA-specific component is larger than a third in the Nordic and French cases, though TIMSS has its own sampling issues that I did not verify. The explanation does not say why students stopped trying, which pushes the causal question to sections 2, 3, 13 and 14, and it does not fit systems where PISA and TIMSS fell together (Israel, the United States, Hungary, Chile, Hong Kong). It is compatible with every feature F1-F7, it has the strongest within-dataset correlations, and its size depends entirely on a definitional choice.

10. Changes in PISA’s own instrument, mode and sampling

Step one. The explanation is that some part of the measured decline is an artefact of the assessment: the 2015 move to computer, adaptive testing (reading 2018, mathematics 2022, science 2025), the 2025 platform change, the 2025 science framework revision, linking error, changes in exclusions and response rates, and the treatment of sub-standard samples. Predictions: an artefact of a design change should appear as a step in the cycle when the change was introduced, in the domain affected, and should be similar across countries; a sampling artefact should be concentrated in the systems whose samples changed; a linking artefact should be shared by all countries in a given cycle-domain; an artefact should not be visible on other assessments; and, if the OECD’s Annex A1 checks are sound, the reading decline should be visible in the non-adaptive Core section and on the same items.

Step two. The validity review supplies the inputs. Against a 25-point reading decline: scaling-model uncertainty is 3-8 points for a country mean and largely cancels in trends; sampling bias is 7-15 points upward for the flagged systems (the Netherlands, Canada, New Zealand, the United States, England in 2022) and 1-2 points on the OECD average; link errors are 1-4 points for recent links and may be understated by about half; the 2015 mode effect does not touch the 2015-25 or 2018-25 comparisons except for Viet Nam, Guatemala and Paraguay; the reading decline is present in the non-adaptive Core section and correlates 0.90-0.91 with percent-correct changes on the same 2018 tasks; there is no equivalent check for mathematics or science; and the 2025 science test combined a framework revision, a new adaptive design and a new platform, none analysed in Volume I. The step-pattern prediction is not met: the reading decline is progressive across four cycles with different designs. The flagged systems are among the decliners, but their bias is upward (under-representation of low achievers), so their declines are if anything understated, as are Norway’s and the Netherlands’ given rising exclusions. The uniformity prediction fails: declines run from −52 (Iceland) to +26 (UAE). The other-assessment prediction is where the instrument story gains ground: PISA mathematics fell where TIMSS grade-8 mathematics did not, and the OECD attributes the 0.79-0.82 correlation of reading and mathematics trends to the reading load of PISA mathematics items (Annex A1, p. 309). That is not an error; it is a statement that PISA mathematics measures something that includes reading and persistence, and that this fell.

Step three. The literature is in the validity review (Kreiner and Christensen 2014; Jerrim et al. 2018; Robitzsch et al. 2020; Rutkowski and Rutkowski 2016; Annexes A1 and A3). Nothing in it supports an artefact of the size of the average decline; several things support caution about individual three-year changes and about science 2025.

Assessment. No single instrument issue can account for the average decline; the reading trend since 2018 rests on a common instrument; several of the most-discussed national declines (the Netherlands, Canada, Norway, the United States, England’s 2022 baseline) carry sampling uncertainty of the order of their three-year changes; and the science 2025 comparison is the least secure, which matters because “science held up” is part of the pattern other explanations are asked to fit: if the science link is off by 5 points either way, the domain ordering (F2) changes. The explanation is consistent with a few points of the average and a larger share of particular countries' changes, inconsistent with the cross-country variation and the progressive timing, and should be held open for science until the Technical Report appears.

11. Educational spending and austerity

Step one. The explanation is that real cuts in education budgets after 2010 (in the United Kingdom, Spain, Italy, Greece, Ireland, Portugal, Finland after 2015, and parts of the United States) or failure to keep spending in line with needs reduced achievement. Predictions: systems that cut spending per student in the 2010s should have fallen more than those that increased it; the effect should lag by several years and be larger for disadvantaged students; systems with the highest spending should have held up; and material-shortage indicators should have risen in the decliners.

Step two. Cumulative spending per student (age 6-15, 2019) correlates with the 2018-25 change at −0.28 across 60 systems and spending per lower-secondary student at −0.45 (trajectories, Table 9): the highest spenders fell most, the opposite of the prediction and the same fact as F3. The material-shortage index fell from 0.02 (2018) to −0.17 (2022) and −0.06 (2025), and its change is uncorrelated with score changes (rho 0.02 to −0.05; Table I.B1.4.89). The distributional prediction fails for 2022-25. The austerity systems do not stand out: the United Kingdom, Italy and Spain, whose real spending per student fell for most of the 2010s (the Institute for Fiscal Studies documents a 9% real fall per pupil in England between 2009-10 and 2019-20), are among the smallest decliners; Finland, which cut after 2015, and Norway and Denmark, which did not, are among the largest. I did not construct a change-in-spending variable (Education at a Glance indicator C1 gives it; the OECD pages could not be fetched programmatically), so the lagged prediction is untested beyond these cases.

Step three. The causal evidence is that money matters: Jackson, Johnson and Persico (Quarterly Journal of Economics, 2016) found a 10% spending increase sustained over twelve years raised completed schooling by 0.3 years; Lafortune, Rothstein and Schanzenbach (AEJ: Applied, 2018) found finance reforms raised achievement in low-income districts by about 0.1 SD over ten years; Jackson and Mackevicius's meta-analysis (AEJ: Applied, 2024, https://ideas.repec.org/a/aea/aejapp/v16y2024i1p412-46.html) gives 0.032 SD on test scores for $1,000 per pupil over four years; Jackson, Wigger and Xiong (AEJ: Economic Policy, 2021) found Great Recession cuts lowered scores by about 0.03 SD per $1,000. A real cut of 10% sustained through a cohort's schooling would lower scores by 0.03-0.05 SD, or 3-5 points.

Assessment. Spending has causal effects that would produce a few points of change for the cuts that occurred in some systems; the cross-country pattern is the reverse of what a spending explanation predicts, and the systems that cut are not the systems that fell. The OECD's statement that "wise spending choices are far stronger predictors than how much countries spend" (A1.8) concerns levels, not the trend.

12. School absenteeism

Step one. The explanation is that persistent absence rose after the pandemic and stayed high, so that the 2025 cohort had less instruction over several years. Predictions: absence at 15 and, more importantly, cumulative absence since 2020 should have risen in the decliners; the within-country association between absence and scores should be strong; the change in absence across systems should correlate with the change in scores; the effect should be larger in mathematics (more school-dependent) and among disadvantaged students; and systems that reduced absence (New Zealand's attendance drive) should have recovered.

Step two. On the OECD average the share who skipped a whole day of school in the two weeks before the test was 21.7% (2018), 19.8% (2022) and 22.4% (2025); skipped classes 27.7%, 22.5%, 24.3%; arrived late 47.7%, 45.3%, 49.7% (Table I.B1.4.161). Self-reported absence at 15 is essentially unchanged since 2018. The within-country association is large: students who skipped a class or day score 38 points lower in science before and 29 after ESCS adjustment (Table I.B1.4.162), so the compositional bound is tiny: 0.6 points of share times 29 is 0.2 points. Across systems, the 2018-25 change in skipping a day correlates with the reading change at −0.14 ($p = 0.23$) and +0.08 within the OECD; skipping classes at −0.40 ($p < 0.001$); lateness at −0.18. Whole-day absence rose most in the United States (+16 points to 40%), Costa Rica (+15), Australia (+11 to 44%), Ireland (+10), Canada (+9), Denmark (+8), the United Kingdom and Poland (+7), and fell in Portugal (−21), Colombia (−13), Latvia (−12), Lithuania (−10), Mexico and the Slovak Republic (−9) (`computed/oecd_questionnaire_by_country.csv`). Latvia, the largest 2022-25 decliner, is a system where absence fell; Australia and the United States, where it rose most, are among the smaller decliners. The domain and distributional predictions fail. Cumulative absence since 2020, which the OECD rightly says matters more (p. 99), cannot be tested with a two-week question.

Step three. Administrative data show large rises in some countries. In England persistent absence (10% or more of sessions) rose from 10.9% in 2018/19 to 22.5% in 2021/22 and 22.3% in 2022/23, 28.3% in secondary schools (https://fullfact.org/education/persistent-absence-rates/). In the United States chronic absenteeism rose 13.5 percentage points between 2018-19 and 2021-22, a 91% increase (Dee, PNAS, 2024, https://www.pnas.org/doi/10.1073/pnas.2312249121), and had only partly receded by 2024. Volume I cites TIMSS fourth-graders absent at least fortnightly rising from 10.7% to 14.8% over 2019-23 (p. 99). The causal effect is well estimated: Aucejo and Romano (Economics of Education Review, 2016) find ten additional days of absence lower mathematics by about 0.055 SD and reading by 0.03 SD; Gershenson, Jacknowitz and Brannegan (Education Finance and Policy, 2017) find similar sizes. Applying 0.05 SD per ten days to a rise of five to ten days a year over three or four years gives 0.05-0.15 SD, or 5-15 points, in the systems with the largest documented rises, and much less elsewhere.

Assessment. Absenteeism is a well-identified cause with effects of the right order in a few countries. It fails as an explanation of the PISA pattern because the PISA absence measure did not rise on average, the systems with the largest documented rises (England, the United States, Australia) are among the smaller decliners, and the largest decliners (Latvia, Iceland, Finland, Norway) have no documented surges of that size. It could account for 5-15 points in the English-speaking systems and is not consistent with the overall pattern.

13. Mental health and well-being

Step one. The explanation is that adolescent mental health worsened from about 2012, especially among girls, and that anxiety, depression, sleep loss and disengagement reduce learning and test performance. Predictions: indicators of well-being in PISA (sense of belonging, life satisfaction, bullying, feeling safe) should have worsened in step with scores; systems where they worsened most should have fallen most; girls should have fallen more than boys where the mental-health trend is gendered; the effect should be broad across domains; and it should coincide with the post-2012 timing documented in the epidemiology.

Step two. The PISA well-being indicators did not worsen between 2018 and 2025. Across OECD members the belonging index averaged 0.00 (2018), −0.02 (2022), +0.09 (2025); the share agreeing "I feel like I belong at school" rose from 71.0% to 76.2% and the share feeling lonely fell from 15.6% to 13.3%; the bullying index went from 0.00 (2018) to −0.30 (2022) and −0.07 (2025), still below 2018 (Tables I.B1.3.63-64, I.B1.4.142). The 2018-25 belonging change correlates with score changes at 0.03 (reading) to 0.19 (science), non-significant; the 2015-25 change in the belonging share at 0.08 (reading) and 0.30 (science, $p = 0.02$). The 2018-25 bullying change correlates with the reading change at −0.43 ($n = 65$, $p < 0.001$) and mathematics at −0.38, in the predicted direction; the OECD reports $r = −0.63$ for the share threatened by peers (p. 97). Life satisfaction is cross-sectional only in the 2025 tables (Table I.B1.3.79); PISA 2022 Volume II reported a 2015-22 decline. The gender prediction fails for reading (boys fell more) and holds for mathematics; the broad-domain prediction fails (science held up).

Step three. The epidemiological evidence that adolescent mental health worsened from about 2012 in many high-income countries, more among girls, is strong (WHO/HBSC 2022 report, https://www.hbsc.org/new-who-hbsc-international-report-a-focus-on-adolescent-mental-health-and-well-being/; national surveys in the United States, the United Kingdom and the Nordic countries; Twenge and Odgers disagree about the role of social media). The evidence that poor mental health lowers achievement is observational and moderate: adolescent depression is associated with lower later attainment after controls (Fletcher, Health Economics, 2008; McLeod and Kaiser, American Sociological Review, 2004), with effects of a tenth to a quarter of an SD for clinical-range symptoms, which affect a minority. Bullying victimisation is associated with lower achievement (Nakamoto and Schwartz, Social Development, 2010, $r \approx −0.10$).

Assessment. The PISA well-being indicators show no deterioration that could carry the decline; belonging and loneliness improved and bullying is below 2018. The one indicator that moves with scores, bullying, is as plausibly a marker of school-climate disruption (TALIS's rising time spent keeping order) as a cause. The epidemiological trend is real, but its likely contribution to the average is a few points at most; the explanation cannot be distinguished from digital distraction, with which it shares a timing and a proposed cause, and it is inconsistent with the PISA belonging trends.

14. Secular trends in cognitive test performance (the reversal of the Flynn effect)

Step one. The explanation is that the twentieth-century rise in scores on cognitive tests (about 3 IQ points per decade) has reversed in high-income countries for cohorts born after about 1975, for environmental reasons, and that PISA is picking up a general cognitive decline. Predictions: the decline should be gradual and roughly linear across birth cohorts from the mid-1970s on, so PISA (cohorts born 1984-2009) should have declined steadily from 2000, at a rate of 1-3 IQ points per decade, which is 7-20 PISA points per decade; it should be uniform across domains, or largest on the most g-loaded (reasoning-heavy) content; it should be present in all high-income countries with saturated Flynn gains and absent or reversed in developing countries still on the rising part of the curve; it should appear equally in TIMSS, PIRLS, national tests and adult tests (PIAAC across ages); and it should not be concentrated in any socio-economic group.

Step two. The timing prediction fails: the OECD-23 aggregate was flat in reading from 2000 to 2012 (cohorts born 1985-97), in mathematics from 2003 to 2018 (born 1988-2003), and science rose to 2009. A decline of even 1 IQ point per decade (7 PISA points) across those cohorts would have produced a fall of about 12 points; none is visible. Norway's PISA reading was 505 (2000), 484, 503, 504, 513 (2015), 499, 477, 453 (`computed/extra_checks.txt`); the Norwegian conscript series on which the reversal claim rests fell from 102.3 for the 1975 birth cohort to 99.4 for the 1989 cohort (Bratsberg and Rogeberg, PNAS, 2018, https://www.pnas.org/doi/10.1073/pnas.1718793115), about 0.2 IQ points per birth year, which would predict roughly 20 PISA points of decline between the 2000 and 2015 cohorts (born 1984 and 1999); the observed change was +8. The domain prediction fails (reading most, science least). The country prediction holds in outline (F3) but is the same fact as regression on starting level. The cross-test prediction fails for TIMSS grade 8 in the Nordic countries and France and holds partially for PIAAC (adult literacy falling across ages, consistent with a period effect rather than a cohort effect). The no-concentration prediction fails for 2018-22.

Step three. The reversal is established for some countries and tests. Dutton, van der Linden and Lynn's review (Intelligence, 2016, https://gwern.net/doc/iq/2016-dutton.pdf) lists declines in Norway, Denmark, Britain, the Netherlands, Finland, France and Estonia averaging 2.4-3.2 IQ points per decade over test years from the 1990s to the 2000s; Pietschnig and Voracek (Perspectives on Psychological Science, 2015) find Flynn gains slowing after the 1970s; Bratsberg and Rogeberg's within-family analysis of 736,808 Norwegian conscripts born 1962-91 reproduces the rise, the 1975 turning point and the fall inside families, ruling out immigration and dysgenic fertility; Dworak, Revelle and Condon (Intelligence, 2023) report declines on reasoning subtests in a large US sample between 2006 and 2018. The magnitudes would suffice if they applied to PISA cohorts and the PISA construct; the timing and shape do not: the conscript decline began with cohorts born in the late 1970s, and PISA shows nothing until cohorts born after about 1997-2000.

Assessment. The Flynn reversal is real on conscription and psychometric tests in several of the countries at the centre of the PISA decline, and it is environmental. As an explanation of the PISA trend it fails the timing test: PISA was flat through the cohorts for which the conscript decline is best documented, fell only later, with an accelerating rather than linear profile and unevenly across domains. What the literature contributes is that the Norwegian decline was environmental, family-invariant and operating on adolescents at testing, the profile of a period effect on test-taking rather than of a change in schooling, which fits section 9. It could account for a slow background decline of a few points per decade and not for the post-2018 acceleration.

15. Other explanations in the inventory: AI chatbots, classroom discipline, family support and early childhood

These are treated briefly because they are too recent or too diffuse to predict the pattern.

AI chatbots (B3). Generative AI reached students in late 2022, so it can bear at most on the 2022-25 window. Volume I reports that students who use AI to summarise texts score lower in science (Table I.B1.4.60), a cross-sectional association with the same selection problem as device use. Across systems the 2025 AI-use index correlates positively with the 2022-25 reading change (rho = 0.24, p = 0.04) and with mathematics and science (0.15, 0.19): systems whose students use AI more did not fall more. There is no causal evidence yet; the claim that AI explains the 2025 results (B3.3) is contradicted by the timing.

Classroom discipline (B7). The PISA disciplinary-climate index in science improved from 0.00 (2015) to 0.14 (2025) (Table I.B1.4.134) and its change is uncorrelated with score changes (rho 0.02 to 0.19); TALIS 2024 reports more time keeping order. The two sources disagree in direction; the bullying correlations in section 13 are the evidence for a climate channel.

Family support and early childhood (B14). The family-support index fell 0.18 over 2022-25 (Table I.B1.4.122) and its change is uncorrelated with score changes (rho −0.08 to +0.02), as Volume I says (p. 100). The early-childhood claim is supported by the OECD's PIRLS correlation and the TIMSS grade-4 correlations here (0.42-0.46 with the PISA 2022-25 change): whatever happened to fourth-graders in 2019-23 is visible in PISA three years later; these do not say what it was.

16. Comparison across explanations

The table summarises, for each explanation, the features it fits and fails, the most informative check, the size the independent evidence supports, and the verdict: consistent with the overall pattern, contradicted by it, consistent with some countries only, or not distinguishable with existing evidence.

Explanation	Fits	Fails	Most informative check	Plausible size (reading −25 since 2018; mathematics −22 since 2015)	Verdict
1. Pandemic	2018-22 mathematics fall and ESCS widening; grade-4 cohort signal	Pre-2018 decline; 2022-25 acceleration; reading > mathematics; advantaged fell most 2022-25; no closure gradient; TIMSS g8 flat in Nordics/France	Closure weeks vs change (≈0 with GDP); W2 vs W3 rho ≈ 0	A third to a half of mathematics since 2018; less of reading; none pre-2018	Part of 2018-22; contradicted as main cause
2. Smartphones / social media	Post-2012 timing; reading most; hasty responding; advantaged fell most 2022-25; early-adopter geography	Japan, Korea, Singapore, Chinese Taipei; in-school use and bans moved benignly 2022-25; no cross-country change correlation	Leisure-use change vs score change (rho +0.2, wrong sign); within-country −8/hour after ESCS in 88/89 systems (cross-sectional)	Ban studies 0.02-0.14 SD among low achievers; out-of-school effect unmeasured; 5-15 reading points is an extrapolation	Consistent; not testable with PISA's measures; size unestablished
3. Reading habits	Reading most; long texts; boys; parents' reading	Mathematics decline of similar size; enjoyment-change test impossible in 2025 data	None available (enjoyment asked 2009/2018 only)	3-8 reading points as composition; more if kind of reading matters	Consistent with reading; silent on mathematics; not distinguishable from 2
4. Composition	Germany, Israel, Portugal, France, Iceland partly	Natives fall ~90% as much as all students; Latvia, Finland; gap narrowed 2022-25	Non-immigrant vs all-student change (r = 0.97; OECD-29 natives −22.4 vs all −24.6)	1.5-3 points on average; a third of Germany's, two-fifths of Israel's reading decline	Some countries; contradicted as general explanation
5. Curriculum / pedagogy	Türkiye; England (confounded)	Finland (decline predates reform); Sweden, Norway, France (no reform at right time); Australia; TIMSS g8 flat where PISA fell	Reform timing vs cohort; TIMSS divergence	A few points in particular systems, either direction	Some countries; not the synchronised decline
6. Digitalised schooling	Nordic and Dutch lead	Germany, Japan; reading > mathematics; TIMSS g8; causal estimates null	Learning-device level vs change (rho 0 to −0.3, n.s. within OECD)	0 to a few points in the most digitalised systems	Some countries; contradicted as a large effect
7. Teacher supply	Shortage index rose; rho −0.25 across all systems	≈0 within OECD; student-teacher ratio fell; certification flat; Finland, Netherlands, Denmark	Staff-shortage change vs change	2-5 points	Part; not distinguishable from common cause
8. Accountability	Türkiye; England (strong version)	Sweden, Denmark (weak version); Korea, Singapore (strong version); governance uncorrelated	Decision-locus features (rho ≈ 0)	A few points, either direction	Some countries; the two versions cancel
9. Test effort	Fell in 66/71 systems; rho 0.32-0.52 with changes; hasty responding; advantaged strugglers; TIMSS divergence	Israel, US, Hungary, Chile, Hong Kong fell on TIMSS too; does not say why	Effort change vs score change; TIMSS g8 vs PISA (rho 0.26; 0.00 in OECD)	A few points to a third of reading (self-report bound ≈9); more in Nordic/French cases	Consistent; size depends on construct vs contamination
10. Instrument, sampling	Flagged systems among decliners (bias upward, so understated); science link least secure	Progressive not stepwise; large cross-country variation; Core-section check	Link errors; Annex A1 percent-correct (r 0.90-0.91)	1-2 points on average; 7-15 for flagged systems; science 2025 unknown	Contradicted as main cause; open for science 2025 and individual countries

Explanation	Fits	Fails	Most informative check	Plausible size (reading −25 since 2018; mathematics −22 since 2015)	Verdict
11. Spending	None	Highest spenders fell most; austerity systems fell least; material shortages fell	Spending vs change (rho −0.28 to −0.45, wrong sign)	3-5 points where real cuts occurred	Contradicted
12. Absenteeism	England, US, Australia surges	PISA two-week absence unchanged (21.7%→22.4%); largest decliners not high-absence; reading > mathematics	Compositional bound 0.2 points; change vs change ≈ 0 within OECD	5-15 points in English-speaking systems; ≈0 elsewhere	Some countries; contradicted as general explanation
13. Mental health	Post-2012 timing; bullying change rho −0.43	Belonging and loneliness improved; bullying below 2018; gender inconsistent	Belonging change vs change (rho ≈ 0)	A few points at most	Contradicted by PISA well-being indicators; not distinguishable from 2 on timing
14. Flynn reversal	Rich systems fell most; environmental, period-like	PISA flat for cohorts born 1985-2000; accelerating not linear; reading > science; TIMSS g8	OECD-23 series 2000-12 flat	A few points per decade of background trend	Contradicted on timing

Four explanations are consistent with the overall pattern in the sense that none of F1-F11 contradicts them: digital distraction (2), reading habits (3), declining effort on the test (9) and, as a residual, the OECD's own composite. They are not independent: 2 and 3 propose reasons why 9 happened, and 9 is the only one with strong within-dataset correlations because it is the only one PISA measures directly. Their combined size cannot be estimated from these data; the bounds run from about a fifth of the reading decline (effort self-report alone) to most of it (if the Nordic and French TIMSS divergence is read as showing a large PISA-specific component).

Four are contradicted by the pattern as general accounts: the pandemic as main cause (1), composition (4), spending (11) and the Flynn reversal (14), each on a different feature, and each surviving as a contributor of a few points or as an explanation for particular systems (the pandemic for the 2018-22 mathematics fall and the grade-4 cohort; composition for Germany and Israel; the Flynn reversal for a slow background trend; spending for nothing identifiable).

Five are consistent with some countries and not others: curriculum and pedagogy (5), digitalised schooling (6), teacher supply (7), accountability (8) and absenteeism (12). For each a country can be named where the timing or geography fits (Türkiye, the Nordic countries, England, the United States) and one where it does not, and none fits the synchronisation across systems with opposite histories.

Two cannot be distinguished with existing evidence: mental health (13) from digital distraction (2), which share a timing and a proposed cause while PISA's well-being indicators move the wrong way for 13; and the instrument explanation (10) for the 2025 science comparison, until the Technical Report documents the framework link and the platform.

Three facts constrain any combination. Native-born students in high-income systems with stable schooling, short or no closures, high spending and no major reform (Finland, Norway, Denmark, Sweden, the Netherlands) fell 35-45 points in reading and mathematics between 2018 and 2025. The same cohort in several of these systems did not fall on TIMSS grade-8 mathematics. Students in the most advantaged quarter fell most in the most recent window, while belonging, discipline, absence at 15 and in-school device use moved in the benign direction. An explanation that fits all three has to operate outside school, across social classes, on the willingness or capacity to sustain effort on long, reading-heavy tasks, and more on PISA than on TIMSS. Digital displacement of reading and attention fits that description best, and it is also the candidate with the weakest direct causal evidence at the scale required. The alternative reading, that PISA has become a test of something 15-year-olds in rich countries no longer do for free, is not a rival hypothesis but the same one seen from the instrument's side. The data assembled here cannot separate them, and this section does not choose.

Appendix F. Inventory of publicly offered explanations

IN BRIEF (summary added before publication; see Appendix H)

This section is an inventory of explanations offered for the PISA 2025 results, compiled on 9 September 2026 from web sources and the OECD's report; it records who said what, where, and with what evidence, without evaluating any claim. It keeps the OECD's own explanations separate from third-party ones, covers mostly European, anglophone and East Asian countries, and counts distinct speakers per explanation type. The explanation offered most often is smartphones, screens and social media, found in ministerial statements in several countries and in OECD materials. Socio-economic inequality and funding is second, offered mostly by unions, opposition parties and researchers. Curriculum and pedagogy claims cluster in anglophone countries and Finland. Migration as a main cause appears chiefly in German-language and Nordic right-of-centre commentary and was contested in Germany and Finland. Sampling caveats were raised in every covered country the OECD asterisked. Some outlets blocked automated retrieval, some attributions are unchecked against a primary source, and counts are approximate.

Compiled 9 September 2026 from web sources published between the run-up to the release and 9 September 2026, and from the text of *PISA 2025 Results (Volume I)* (OECD, released 8 September 2026). This is an inventory, not an evaluation: each entry records who said what, where, and what evidence (if any) the speaker offered. Where a quotation is given in a language other than English, the original is followed by a translation. Where the source was only available to us through a summarising fetch rather than the full page, the claim is marked "(paraphrase per source)". Duplicate or near-identical claims are grouped under a single entry with multiple sources.

Coverage: OECD (Volume I and statements by Mathias Cormann, Andreas Schleicher, Francesco Avvisati, Tue Halgreen); United States; United Kingdom (England, Scotland, Wales, Northern Ireland); Germany; France; Netherlands; Belgium (Flanders); Sweden; Norway; Denmark; Finland; Iceland; Australia; New Zealand; Canada; Japan; Korea; Singapore; Estonia; Ireland; Switzerland; Austria; Italy; Spain; Poland; Türkiye.

Two caveats on the inventory itself. First, several major outlets (NRK, TV2 Norway, DR, The Globe and Mail, NZ Herald, Nikkei, Yahoo Japan, The Hill, US News) blocked automated retrieval, so their coverage is represented indirectly where another source quotes the same speaker. Second, some speaker attributions come from national press summaries and could not be checked against a primary transcript; those are flagged.

Part A. The OECD's own explanations (kept distinct from third-party claims)

A1. Volume I text

A1.1 — The decline is real and predates the pandemic. Preface (Mathias Cormann, Secretary-General): "These developments are part of a longer-term trend. The weakening of performance predates the COVID-19 pandemic in many countries, pointing to deeper and more persistent challenges within education systems. In reading, increased digitalisation, longer screen time and falling rates of reading for enjoyment have coincided with weaker literacy outcomes." Source: Volume I, Preface, p. 4 (vol1.txt lines 78–82). Evidence offered: OECD-average trend series 2015–2025 (science 489→482; mathematics 485→463; reading 489→461), Preface notes 1–3.

A1.2 — Foreword (Andreas Schleicher): COVID is an insufficient explanation; reading habits and digital reading are the leading hypothesis. "People quickly turn to explanations such as the impact of the COVID pandemic, but a decline in reading was already visible well before the pandemic hit. We cannot say for sure, but declining levels of reading for enjoyment and the shift towards digital reading – skimming social media feeds and rapidly processing information – may be compromising the ability of students to engage with complex texts and data." Also: "This raises questions about whether devices such as smartphones – designed to maximise a user's attention – are compatible with sustained academic attention." Source: Volume I, Foreword, pp. 6–7 (lines 194–201). Evidence offered: the "hasty readers" share almost doubled 2018–2025; leisure use of devices at school associated with lower scores.

A1.3 — Foreword: AI. "Students who are using AI chatbots for specific tasks such as summarising texts get lower scores in science, on average, than those who don't. The patterns are complex, and depend a lot on how GenAI is used … Where technology short-circuits the productive struggle of learning, it will undercut students' development." Source: Volume I, Foreword, p. 7 (lines 202–209). Evidence: Chapter 4 associations between AI-chatbot use for specific tasks and science scores (executive summary, lines 1793–1797).

A1.4 — Box I.2.3 "Why has reading performance declined over time? Four underlying forces." The Box explicitly frames its four levels as "complementary rather than competing" hypotheses and states "The evidence does not point to one simple explanation" (lines 14698–14701, 14914–14928). The four forces:

(a) Students — reduced attention and effort rather than uniform skill loss. Success rates fell most on long-text items; the share of "hasty readers" rose from about 7% in 2018 to 11% in 2025; hasty responses rose across the reading test while non-reached items fell; declines were steeper in later test sections; self-reported test effort fell; changes in curiosity correlate with reading trends ($r = 0.54$) and "school is a waste of time" beliefs with reading trends ($r = 0.53$). Source: lines 14703–14765; Annex A1, lines 37174–37270, 37301–37332.

(b) Schools and classrooms — more disruption, weaker perceived support. Bullying rose again between 2022 and 2025; changes in the share of students threatened by peers correlate with reading declines ($r = −0.63$); TALIS 2024 shows class time spent keeping order rose from 13% (2018) to 16% (2024). Source:

lines 14767–14802.

(c) System-level — more diverse classrooms and "feeble governance." Principals report more refugee, non-native-speaker and special-needs students; "PISA data show that changes in student body composition do not explain the overall decline in performance (Annex A1)"; declining use of data for school improvement in some systems (e.g. Norway*), but "this pattern does not appear to be consistently associated with more negative trends." Source: lines 14804–14833.

(d) Societal — reading habits, digital lives, the pandemic, absenteeism, family support. More students report >1 hour/day on social networks or games; the 2025 cohort (born c. 2009) "grew up during a period of rapid change and diffusion of personal electronic devices"; on the pandemic, "there is no simple relationship between the duration of closures and subsequent reading trends" (Sweden, no closures, −21 points; Costa Rica, ~20 weeks, no significant change); post-pandemic absenteeism (TIMSS fourth-graders absent at least fortnightly rose from 10.7% in 2019 to 14.8% in 2023); fewer students report parents asking about school, but this trend does not differ between high-decline and low-decline systems. Source: lines 14835–14908.

A1.5 — Measurement artefacts ruled out (OECD's own position). "Results in Annex A1 show that the decline is unlikely to reflect measurement issues or PISA-specific anomalies" (line 14691). Annex A1 adds that trends are visible in the first, non-adaptive section of the test; test fatigue effects have not widened; questionnaire straightlining fell; PISA 2025 means correlate with PIRLS 2021 (r = 0.74) and PIRLS 2016→2021 changes anticipate PISA 2022→2025 changes (r = 0.34, n = 23), suggesting causes "could … be plausibly found in the earlier childhood years." Source: lines 37018–37041, 37068–37077, 37351–37433.

A1.6 — Demographic change explains only a few points. Immigrant/non-immigrant gaps were stable in most countries; the composition effect is roughly 1.5 of the 25-point OECD reading decline (0.038 × 40), about 3 of 30 points in Portugal; the share of immigrant students rose by more than 3 points in Portugal, Germany, Norway, Ireland, Slovenia, Iceland, Austria, Switzerland and France. "Such demographic changes are never the main driving force behind the large trends observed." Source: lines 36976–37016.

A1.7 — The decline is concentrated among advantaged students. "the largest drops were observed among the most advantaged quarter of socio-economic status … a group of students from relatively affluent families, but who struggle academically, are contributing most to the overall decline." Source: lines 36957–36964; executive summary line 1763.

A1.8 — Spending is not the explanation. "good policies and wise spending choices are far stronger predictors for performance on PISA than just how much countries spend per student: At USD 288 521 per student, Luxembourg is by far the biggest spender … but it posted lower average science scores than Ireland, France and Italy." Source: Executive summary, lines 1687–1692.

A1.9 — Data-quality caveats issued by the OECD itself. Full reporting with asterisk: Canada (and seven provinces), the Netherlands, New Zealand, Norway, Murcia (Spain); limited reporting with asterisk: Albania and the United States. For the United States, school response 45% before replacement (54% after), student response 76%; "it is not possible to rule out bias in PISA performance results, nor to determine its most likely direction"; over 40% of US respondents did not receive the student questionnaire. New Zealand: mean scores "may be over-estimated by about 10 to 15 score points." Norway: exclusion rate 10.4% "may bias performance results upwards." Catalonia's results are not reported separately (exclusions 23.2%). Source: Reader's Guide, lines 920–1080; US country note, p. 2.

A2. OECD officials' public statements around the launch

A2.1 — Mathias Cormann, press launch, 8 Sept 2026 (Reuters). "In reading, more screen time and less reading for pleasure have gone hand in hand with weaker results." "We're also seeing more hasty reading, students rushing through a text and giving a quick wrong answer." Source: Reuters via Yahoo Finance, https://finance.yahoo.com/news/teen-reading-slumps-worst-century-111013754.html (8 Sept 2026). Also quoted by Al Jazeera: "The most successful education systems focus on fewer areas in greater depth, invest in teachers, engage parents and provide targeted support for the students and schools that need it the most." https://www.aljazeera.com/news/2026/9/9/us-reading-scores-hit-near-25-year-low-in-latest-global-assessment (9 Sept 2026).

A2.2 — Andreas Schleicher, OECD Education and Skills Today blog, 8 Sept 2026. Names as drivers of the reading decline "the changing nature of reading through digitalisation and screen use, which may be displacing time previously devoted to deeper reading," "falling rates of reading for enjoyment," and "changing attitudes towards learning and school since the COVID-19 pandemic, possibly reflected in increasing levels of student absence in many countries, as some studies have suggested"; says the pandemic "cannot be ignored" but "educational performance was already trending downward in many countries long before the pandemic struck." Source: https://oecdedutoday.com/the-state-of-global-education-according-to-pisa/.

A2.3 — Schleicher, US-facing remarks (K-12 Dive, 8 Sept 2026). "People quickly turn to explanations, such as the COVID pandemic … if you're honest, this decline started much earlier and it continues." "Students spent 1.7 hours per day on devices for learning and then another hour for leisure while in school." "The biggest increase in the U.S. is being in browsing social networks such as Instagram and TikTok … a quite troubling shift from creating to consuming." Source: https://www.k12dive.com/news/from-science-to-screen-time-4-key-takeaways-from-the-latest-pisa-results/829867/.

A2.4 — Schleicher on the US: class size versus teacher quality (NCEE webinar, as reported by Joanne Jacobs, 9 Sept 2026). "In the U.S. if you make a choice between a better teacher and a smaller class, you typically choose a smaller class." Also (paraphrase per source): Asian systems use AI to challenge students, Western schools use it to simplify learning. Source: https://www.joannejacobs.com/post/us-reading-scores-fall-but-others-do-even-worse.

A2.5 — Schleicher on Germany (Tagesspiegel, 8 Sept 2026; FAZ interview, 8 Sept 2026). Germany shows pronounced between-school inequality: about 50% of performance variation lies between schools versus under 10% in Iceland, Cambodia and Finland (paraphrase per source). FAZ interview headline quote: "Es ist zu einfach, sich durchs deutsche Bildungssystem zu schlängeln" ("It is too easy to wriggle through the German education system"); "Jeder sollte mal eine asiatische Schule besuchen" ("Everyone should visit an Asian school once"). Sources: https://www.tagesspiegel.de/wissen/neue-pisa-studie-neue-negativrekorde-die-bildungsrepublik-deutschland-steht-so-schlecht-wie-nie-da-16028742.html; FAZ via https://newzs.de/2026/09/08/pisa-verlierer-und-gewinner-es-ist-zu-einfach-sich-durchs-deutsche-bildungssystem-zu-schlaengeln/ (full FAZ text paywalled; only headline quotes verified).

A2.6 — Schleicher on the UK nations (IBTimes UK, 8 Sept 2026). "pointed to England's consistent education policies when discussing the differences in performance between the UK nations" (paraphrase per source). Source: https://www.ibtimes.co.uk/englands-students-excel-reading-global-decline-1818609.

A2.7 — Francesco Avvisati (OECD, Volume I co-author), Tagesspiegel, 8 Sept 2026. "Junge Leute lassen sich leichter in die Irre führen, weil sie zu hastig lesen" ("Young people are more easily misled because they read too hastily"). Evidence: rushed answers, fewer students wanting to understand how things work. Source: Tagesspiegel URL above.

A2.8 — Tue Halgreen (OECD senior analyst, Spain coordinator), Madrid presentation, 8 Sept 2026. Four lines of evidence for a smartphone link: the drop came "between 2015 and 2025, a decade when smartphones became an indispensable part of our lives"; it happened "in numerous countries that have implemented vastly different educational policies"; "students who consume fewer screens show better concentration levels and also improve their results"; performance fell "much worse" on long texts (>400 words) while, in the article's summary, locating-information skills have remained more stable. Halgreen also cautioned, in reported speech, that "no se puede establecer una relación causal entre ambas cuestiones" ("a causal relationship cannot be established between the two issues") (Público). Sources: https://www.todoalicante.es/english/oecd-links-academic-decline-20260908090938-nt.html; https://www.publico.es/sociedad/educacion/espana-toca-fondo-informe-pisa-minimos-historicos-lectura-matematicas-ciencias.html; https://www.infobae.com/espana/2026/09/08/el-gobierno-admite-que-los-datos-de-pisa-2025-no-son-buenos-pero-rechaza-la-autocritica-con-la-lomloe-completamente-implantada/.

Part B. Third-party explanations, by type and country

B1. Smartphones, social media, screen time and digital distraction (leisure use)

B1.1 — United States. Linda McMahon, Secretary of Education: the 14-point reading drop "is equivalent to a seven-month loss in learning"; her main framing was structural (see B13.1), but the K-12 Dive report pairs her remarks with the OECD screen findings. Source: K-12 Dive, 8 Sept 2026 (URL at A2.3). Education Week (Stephen Sawchuk, 8 Sept 2026) reports the share of US students spending >1 hour/day on social networks rose 6 points and a negative correlation between AI writing help and reading scores: https://www.edweek.org/leadership/reading-ability-craters-on-international-exam-both-in-the-united-states-and-abroad/2026/09.

B1.2 — Sweden (government). Simona Mohamsson (L), Minister for Schools: "Vi är mitt uppe i en skärmpandemi" ("We are in the middle of a screen pandemic"); "Det är skärmarna som är giftet i vårt samhälle, både i och utanför skolan" ("Screens are the poison in our society, both inside and outside school"); "Resultaten sjunker överallt. Det är en internationell trend och det är skärmpandemin som nu ger fullt utslag" ("Results are falling everywhere. It is an international trend and it is the screen pandemic that is now having full effect"); "Vi riskerar att få en hel generation av funktionella analfabeter." Lotta Edholm (L): "Här sjunker ju läsförståelse och den matematiska kunskapen även för de högpresterande." Government press release attributes decline to "ökad mobilanvändning och mindre läsning" at the societal level and to "elevers uthållighet att lösa komplexa uppgifter" at the pupil level; credits mobile-free schools, 12 million books, analogue early-years instruction, new teacher education. Sources: https://www.vilarare.se/nyheter/pisa-2025/mohamsson-vi-ar-mitt-uppe-i-en-skarmpandemi; https://www.tv4.se/artikel/e0rddVzw2gvhm7PzvOTvY/skyller-pa-skaermarna-pandemi; https://www.skolledaren.se/aktuellt/nyheter/2026/9/skarmpandemi-och-nationellt-nodlage--sa-kommenterar-politikerna-pisaraset; https://www.regeringen.se/pressmeddelanden/2026/09/forsamrade-pisa-resultat-visar-att-regeringens-skolsatsningar-ar-nodvandiga/ (all 8 Sept 2026). Evidence offered: OECD's international trend; no Sweden-specific analysis cited.

B1.3 — France (minister). Édouard Geffray, Minister of National Education: social media is an "arme de destruction intellectuelle massive" ("weapon of mass intellectual destruction"); "Les jeunes qui passent 3 heures par jour sur leur téléphone pendant les cours et qui sont distraits. Ils perdent l'équivalent quasiment depuis 2 années scolaires" ("Young people who spend three hours a day on their phone during lessons and are distracted lose the equivalent of almost two school years"). Ministry dossier lists "l'exposition croissante aux écrans et aux réseaux sociaux," "l'utilisation du téléphone portable à des fins de loisirs," and "le recul du suivi de la scolarité par les parents"; credits extending the phone ban to lycées. Sources: https://www.cafepedagogique.net/2026/09/09/pisa-que-dit-le-ministre-edouard-geffray/; https://www.publicsenat.fr/actualites/politique/resultat-du-classement-pisa-les-ministres-passent-et-le-systeme-demeure-martele-max-brisson; https://www.education.gouv.fr/pisa-2025-une-baisse-des-resultats-commune-l-ocde-un-relevement-poursuivre-505690 (8–9 Sept 2026). Evidence offered: OECD device-use figures; no French causal study cited.

B1.4 — Spain (government). Milagros Tolón, Minister of Education, blamed "hábitos de pantallas" (screen habits), rushed reading and short social-media messages, rejecting LOMLOE as a cause; Abelardo de la Rosa, Secretary of State, linked decline to screen use and called for more "lectura pausada." Sources: Infobae and El Plural, 8 Sept 2026 (URLs at A2.8 and https://www.elplural.com/politica/espana/moncloa-genova-chocan-analisis-resultados-informe-pisa-la-lomloe-uso-pantallas_399740102). Evidence: OECD/Halgreen presentation (A2.8).

B1.5 — Denmark (government). Ministry press release (Magnus Heunicke, Minister of Children and Education, 8 Sept 2026): “distraktion fra digitale apparater og medier fortsat udgør en betydelig udfordring” (“distraction from digital devices and media continues to be a significant challenge”), citing 36% of Danish students reporting distraction in most/every science lesson; frames results as part of a “vestlig læringskrise” (Western learning crisis). Prime Minister Mette Frederiksen (per POV International, 9 Sept 2026) advocated classroom discipline, stronger teacher authority and removing mobile phones (paraphrase per source). Sources: https://uvm.dk/aktuelt/nyheder/2026/september/260908-pisa-undersoegelse-danske-elevers-faglige-niveau-er-lavere-end-nogensinde-maalt-foer/; https://pov.international/historisk-daarlige-pisa-tal-regeringen-lover-flere-penge-til-folkeskolen/.

B1.6 — Austria (minister). Christoph Wiederkehr (NEOS), Education Minister: “Nicht begleitete Digitalisierung in Schulen führt zu schlechteren Ergebnissen” (“Unaccompanied digitalisation in schools leads to worse results”); cites early screen activity correlating with weaker results (paraphrase per source). Source: ORF, 8 Sept 2026, https://orf.at/stories/3441557/.

B1.7 — Germany (Federal Drug Commissioner). Hendrik Streeck (CDU): “Wie viel wissenschaftliche Evidenz brauchen wir eigentlich noch, bevor wir handeln?” (“How much scientific evidence do we still need before we act?”), calling for protective measures against manipulative social-media mechanisms. Karin Prien (CDU), Federal Education Minister, also cited declining parental reading and rising smartphone use in families (paraphrase per Tagesspiegel). Sources: https://www.zdfheute.de/politik/deutschland/pisa-studie-streeck-prien-historisch-schlecht-reaktionen-100.html; Tagesspiegel (A2.5).

B1.8 — Korea (ministry official). Unnamed Ministry of Education official: “숏폼과 SNS 이용의 확대 등이 학생들의 읽기 역량 저하에 영향을 미친 것으로 보고 있다” (“We see the expansion of short-form content and SNS use as having affected the decline in students’ reading competence”); and “하나의 요인이라기보다 최근 디지털 매체 활용 증가, 숏폼 콘텐츠와 SNS 유행 등이 복합적으로 작용한 것으로 추정된다” (“Rather than a single factor, it is presumed that increased digital-media use, the short-form content trend and SNS acted in combination”). Sources: https://edaily.co.kr/News/Read?mediaCodeNo=257&newsId=03578486645578152; https://www.segye.com/newsView/20260908510225; https://en.sedaily.com/society/2026/09/09/korean-students-reading-scores-fall-14-points-trailing-china (8–9 Sept 2026). Evidence: reading fell 515→501; Level-1-and-below rose 14.7%→17.8%.

B1.9 — Australia (minister and commentators). Jason Clare, federal Education Minister: “the more kids use screens, the more their results decline” (Meyka, 9 Sept 2026), flagged device use for the October education ministers’ meeting, while acknowledging (ABC) that students need digital literacy. Meyka cites 57% of Australian students spending >2 hours/day on devices at school vs 30% OECD. Sources: https://meyka.com/blog/australias-pisa-scores-hit-20-year-low-as-screen-time-soars-in-classrooms-0909/; https://www.abc.net.au/news/2026-09-08/australia-records-worst-ever-reading-and-maths-results-pisa/107101504. Counter-framing: Karmactive (Sunita Somvanshi, 9 Sept 2026) argues that Australia’s strong Learning-in-the-Digital-World score (532 vs 500) means the target should be “unmanaged distraction, weak attendance and uneven basics rather than technology itself”: https://www.karmactive.com/australia-pisa-2025-reading-maths-digital-world-score/.

B1.10 — Estonia. Estonian World (Silver Tambur, 8 Sept 2026) cites 36% of Estonian students reporting device distraction in science lessons vs 28% OECD, alongside the reading decline: https://estonianworld.com/knowledge/estonia-remains-among-the-worlds-best-education-systems-in-pisa/.

B1.11 — Global (Education International). David Edwards, EI General Secretary: “Our students are growing up in a world constantly vying for their attention, a world full of distractions and lacking connection.” Source: https://www.ei-ie.org/en/item/32952:new-pisa-report-reveals-decline-in-student-performance-amid-chronic-teacher-shortage-and-shrinking-attention-spans (8 Sept 2026).

B1.12 — Counter-claim (screens do not explain Singapore). Carl Hendrick (education researcher/author), as reported by Joanne Jacobs (9 Sept 2026): the reading decline predates the pandemic and has not recovered, and Singapore “contradicts screen-saturation theory with high scores” (paraphrase per source). Source: joannejacobs.com (A2.4). Stig Toke Gissel (UCL Professionshøjskole, Denmark) likewise notes the “paradoxical finding” that 4–5 hours daily of digital learning correlates with better reading scores in the Danish data (paraphrase per Folkeskolen, https://www.folkeskolen.dk/historisk-fald-i-pisa-aldrig-har-dansk-skole-klaret-sig-daarligere/, 8 Sept 2026).

B2. Digitalisation of schooling itself (tablets, ed-tech, “digital learning”)

B2.1 — Norway (minister). Kari Nessa Nordtun (Ap), Minister of Education: “Jeg har sagt tidligere at satsingen på nettbrett og andre digitale ressurser har vært kunnskapsløst og fullstendig feilslått, og at det har vært en katastrofe for læringsresultatene i skolen” (“I have said before that the investment in tablets and other digital resources was ignorant and completely misguided, and that it has been a catastrophe for learning outcomes in school”); “Samtidig med at skjermene kom inn, så forsvant fysiske lærebøker og annet praktisk utstyr i skolen ut” (“As screens came in, physical textbooks and other practical equipment disappeared from schools”); but also, in VG’s reported speech, that no single reason for the fall in learning outcomes has been identified (“Kunnskapsministeren viser til at det ikke pekes på en enkelt grunn til fallet i læringsresultatene”). Sources: https://www.vg.no/nyheter/i/L4Xg3x/norsk-skole-faar-enda-en-kilevink-i-pisa-undersoekelsen; https://skolemagasinet.no/pressemelding/kommentar-pa-fremleggelsen-av-pisa-2025/ (8 Sept 2026). Evidence offered: minister’s own assessment; “researchers identified digitalisation as a contributing factor” (unspecified). Helle Christin Nyhuus (Lektorlaget): “Digitaliseringen har altfor ofte løpt foran de faglige vurderingene” (“Digitalisation has all too often run ahead of professional judgement”) (VG). VG editorial (8 Sept 2026): “Norsk rikdom gjorde det mulig. Vi kunne putte en skjerm i hendene på barn helt fra første klasse. På dette området ble rikdommen vårt åk” — https://www.vg.no/nyheter/i/7poM4V/katastrofalt-daarlige-pisa-resultater.

B2.2 — Finland (opposition and coalition). Riikka Purra (Finns Party, Finance Minister) lists “Digiuskovaisuus, inkluusio, avotilat, itseohjautuvuudet, ilmiöoppimiset, armoviitosten hyväksyminen, vaatimusten puute, jatkuvien oikeuksien lisääminen, auktoriteettien puute” (“Digital zealotry, inclusion, open-plan spaces, self-direction, phenomenon-based learning, acceptance of mercy passes, lack of demands, the continual addition of rights, lack of authority”) as causes;

also “Suomalainen sisu on kateissa” (“Finnish sisu is missing”). Pia Lohikoski (Left Alliance): “painettu oppikirja tukee oppimistuloksia paremmin kuin digitaalinen” (“the printed textbook supports learning outcomes better than a digital one”). Sources: https://www.suomenuutiset.fi/purra-pisa-tuloksista-ymmarretaanko-jo-mita-on-tehty-vaarin-digiusko-inkluusio-ja-vaatimusten-puute-soivat-osaamisen/ (8 Sept 2026); https://www.pyhtaanlehti.fi/paikalliset/9777257 (9 Sept 2026). Evidence offered by Purra: PISA perseverance item (Finland lowest), employer feedback on basic skills; no studies cited.

B2.3 — Sweden (government). Government credits its pivot to “analog” instruction in early grades and printed textbooks (regeringen.se, B1.2).

B2.4 — Austria. Wiederkehr’s “unaccompanied digitalisation” claim (B1.6) covers school digitalisation as well as leisure use.

B3. AI chatbots and cognitive offloading

B3.1 — Global (EI). David Edwards: “AI is being pushed as the magic solution when in fact excessive use results in children outsourcing their learning and their thinking to a bot.” Sources: EI (B1.11); IBTimes UK, https://www.ibtimes.co.uk/oecd-report-decline-reading-scores-2025-1818573 (8 Sept 2026).

B3.2 — United States (academic). Jonathan Seiden (Vanderbilt): ed-tech and AI “are not magic bullets to solve these problems … not replacements for effective teachers or substitutes for family support for learning at home”; implementing them without attention to engagement “could actually exacerbate the disparities we see between low and high-performing students.” Source: K-12 Dive (A2.3).

B3.3 — Scotland (union). Andrea Bradley, EIS General Secretary: the report “clearly demonstrates that young people who do not use AI in their studies, particularly in reading, out-perform those who do.” Source: https://www.eis.org.uk/latest-news/pisaresults (8 Sept 2026). Evidence: OECD Chapter 4 associations.

B3.4 — Spain (minister). Milagros Tolón cited “screen overuse and AI impact” (El Plural, B1.4).

B3.5 — Estonia, Austria, Italy (press). Estonian World: 58% of Estonian students use AI weekly; users for summarising/drafting scored ~20 points lower in science. ORF (Austria): nearly 50% use chatbots weekly; summarisers score lower. Il Giornale (Italy): 50% weekly use flagged as a concern. Sources: B1.10; B1.6; https://www.ilgiornale.it/news/politica/la-scuola-italiana-resta-in-piedi-valditara-sui-dati-ocse-pisa-siamo-sopra-la-media-diventiamo-un-riferimento-internazionale/.

B3.6 — France (academic). Claude Diebolt (CNRS/University of Strasbourg, drawing on work with Nadir Altinok), The Conversation France, September 2026 (published after 8 Sept; exact date not retrieved): AI makes it tempting to think fundamentals matter less, whereas “PISA souligne que certaines des compétences en compréhension de l’écrit qui diminuent sont précisément celles dont nous avons besoin dans un environnement informationnel saturé.” Source: https://theconversation.com/pisa-2025-et-classements-scolaires-nous-posons-nous-les-bonnes-questions-291492. (Diebolt’s quotations here and in B12.5 and B15.9 were not rechecked before publication: the site blocks automated access.)

B4. Declining reading habits, reading for pleasure, attention spans

B4.1 — Netherlands (Leescoalitie). “Het onderwijs kan het niet alleen” (“Education cannot do it alone”); “ook thuis begint het stimuleren van taal en lezen al op jonge leeftijd”; blames insufficient societal reading culture. Source: https://cpnb.nl/nieuws/reactie-leescoalitie-op-gedaalde-pisa-resultaten-steviger-inzetten-op-leescultuur/ (8 Sept 2026). Evidence: PISA-2025 report’s call for joint effort; Masterplan Basisvaardigheden.

B4.2 — Estonia (teachers’ association). Kaja Sarapuu, Chair, Estonian Association of Mother Tongue Teachers: “At school, we mostly just read and find a fact in a text,” i.e., pupils lack deep textual analysis; students read fewer long texts and consume short fragments. Source: Estonian World (B1.10).

B4.3 — Iceland (press context). Iceland Review (9 Sept 2026) recalls education officials’ 2023 finding that Icelandic vocabulary was declining because of “reduced reading and the growing presence of English in children’s daily lives,” without asserting a causal link to PISA 2025. Source: https://www.icelandreview.com/news/iceland-pisa-2025-worst-results/.

B4.4 — Finland (national PISA director). Arto K. Ahonen (University of Jyväskylä): reduced reading of printed books and dwindling perseverance (Yle reports that only 41 percent of Finnish pupils now say they try harder when a task feels difficult; Ahonen: “This was the smallest share among the different countries”). Source: Yle, https://yle.fi/a/74-20245132 (8 Sept 2026); Finnish Government release https://valtioneuvosto.fi/en/-/1410845/pisa-2025-students-in-finland-still-perform-well-above-oecd-average-in-science.

B4.5 — Austria (IQS). Institute for Quality Assurance in Education: shorter attention spans, changed media consumption, difficulty with tasks requiring perseverance on longer texts (paraphrase per ORF, B1.6).

B4.6 — Germany (Philologenverband). Susanne Lin-Klitzing, DPhV chair: the share of “flüchtig Lesender” (hasty readers) nearly doubled 2018–2025, undermining mathematics and science too. Source: https://www.dphv.de/2026/09/08/pisa-2025-dphv-fordert-blick-auf-die-voraussetzungen-guter-bildung-statt-stereotyper-kritik-am-deutschen-schulsystem-und-den-lehrkraeften-verband-mahnt-konsequente-vorschulische-sprachfoerderung-un/ (8 Sept 2026).

B5. COVID-19 pandemic (as cause, and as rejected cause)

B5.1 — United States (NCEE). Tracey Burns, NCEE: "There's an ongoing global decline, and it seems like the pandemic has a long, long tail." Source: Chalkbeat, https://www.chalkbeat.org/2026/09/08/pisa-results-2025-show-us-gaining-as-other-countries-decline/ (8 Sept 2026).

B5.2 — Korea (ministry official). On mathematics polarisation, a ministry official (per edaily) linked the difficulties of this year's third-year middle-school and first-year high-school students to the disruption of 2020–2021, when they were in grades 4–6 (paraphrase; the Korean wording previously quoted here could not be found in the source). Source: edaily (B1.8). Evidence: low achievers in maths rose 16.2%→18.6%.

B5.3 — Netherlands (national PISA team, NKO). "PISA liet in 2022 wereldwijd een flinke achteruitgang zien … Mogelijk speelden de gevolgen van de coronapandemie voor het onderwijs hierbij een rol" ("Possibly the consequences of the corona pandemic for education played a role"); "Drie jaar later vertonen veel landen, waaronder Nederland, nog weinig tekenen van herstel." Source: https://www.nko.nl/nieuws/pisa-2025-leesvaardigheid-daalt-verder-leerlingprestaties-in-exacte-vakken-gelijk (8 Sept 2026).

B5.4 — Scotland (academic). Lindsay Paterson (Edinburgh): "Science has bounced back a little from the harm caused by shutting schools during Covid," but only to 2018 levels. Source: https://www.enlighten.scot/scotland-and-the-2025-pisa-results/ (8 Sept 2026).

B5.5 — Austria (IQS). Long-term pandemic effects that "den sozialen Kontext nachhaltig verändert haben" (ORF, B1.6).

B5.6 — Poland (minister). Barbara Nowacka frames current results as reflecting "the post-pandemic period, post-gymnasium era, and early reform changes" (paraphrase per Głos Nauczycielski, https://glos.pl/pisa-2025-ibe-wyniki-ujawniaja-globalny-kryzys-edukacji-na-tym-tle-polska-wypada-stabilnie-szefowa-men-jestesmy-zielona-wyspa-ma-tle-europy, 8 Sept 2026).

B5.7 — Rejections of COVID as main cause. Schleicher (A1.2, A2.3); Carl Hendrick (B1.12); Hans Christensen (Substack, 9 Sept 2026, restating OECD: "a decline in reading was already visible well before the pandemic hit," https://hanschristensen.substack.com/p/skills-and-knowledge-decline-pisa); VIVE Denmark (Vibeke Tornhøj Christensen: "Frem til 2015 lå de danske PISA-resultater ret stabilt. Siden har de taget et løbende dyk," https://www.vive.dk/da/nyheder-og-debat/2026/danske-elevers-pisa-resultater-fortsaetter-med-at-falde/).

B6. Absenteeism and attendance

B6.1 — Australia. ABC (Conor Duffy, Miwa Blumer, 8 Sept 2026) reports 44% of Australian students skipped school in the two weeks before the test vs 24% OECD; Karmactive lists "weak attendance" among root causes. Sources: B1.9.

B6.2 — Denmark. Stig Toke Gissel: increased absenteeism (28% vs 20% previously) among factors; VIVE: since 2012, 60% arrive late, 30% skip classes, 28% miss whole days within two weeks. Sources: Folkeskolen (B1.12); VIVE (B5.7).

B6.3 — New Zealand (minister, as credit). Erica Stanford: "You cannot learn if you are not at school"; regular attendance in Term 2 rose from 39.6% (2022) to 64.3% (2026), linked to disadvantaged students gaining "around a full year of learning in science." Source: https://www.scoop.co.nz/stories/PA2609/S00092/disadvantaged-students-drive-significant-closing-of-education-equity-gap.htm (9 Sept 2026).

B6.4 — Norway (VG editorial). "Motivasjonen faller og fraværet øker" ("Motivation falls and absence rises") (B2.1).

B6.5 — Italy. Il Giornale notes 63% of Italian students reported absences in the testing period (B3.5).

B7. Classroom discipline and disruption

B7.1 — Australia. Jonas Bertling (ACER): "Disruption remains a challenge in our Australian classrooms." Blaise Joseph (Centre for Independent Studies): "Students cannot learn effectively when classrooms are regularly disrupted," advocating "clear expectations, explicitly taught routines and consistent approaches to behaviour." Evidence: 41% report noise/disorder vs 31% OECD; Australia ranked 69th of 80 on disciplinary climate. Source: ABC (B1.9).

B7.2 — Denmark (PM). Mette Frederiksen: classroom discipline and teacher authority (B1.5).

B7.3 — Poland (minister). Barbara Nowacka: "Edukacja wymaga dyscypliny, znacznego prowadzenia uczniów przez nauczycieli" ("Education requires discipline, substantial guidance of pupils by teachers") (B5.6).

B7.4 — Scotland (Lib Dems). Duncan Dunlop MSP cites absenteeism and classroom violence. Source: STV, https://news.stv.tv/scotland/scotland-holds-competitive-position-in-education-but-core-subjects-decline (8 Sept 2026).

B7.5 — Norway (VG editorial). "Klasserommene er blitt mer urolige. Flere elever trenger spesialundervisning" ("Classrooms have become more restless. More pupils need special education") (B2.1).

B8. Student motivation, perseverance, effort and attitudes

B8.1 — Finland. Ahonen (B4.4); Purra’s “sisu” (B2.2); Adlercreutz (Education Minister, Swedish People’s Party — party affiliation per general knowledge; the Pyhtäänlehti summary mislabels him): schools must teach pupils to concentrate, to tolerate boredom and to try again (“keskittymään, sietämään tylsyyttä ja yrittämään uudelleen,” per Pyhtäänlehti, B2.2).

B8.2 — Denmark. Christian Christrup Kjeldsen (school researcher): “Tilsammen viser de her tal, at der er tale om markante fald i elevernes tilhørsforhold til skolen” (“Taken together these figures show marked declines in pupils’ attachment to school”); 22% see school as a waste of time (28% boys, 16% girls). Source: Folkeskolen (B1.12).

B8.3 — Japan (MEXT/NIER). NIER summary notes Japanese students’ curiosity items are “OECD平均よりもおおむね低い” (“generally lower than the OECD average”) and that reading-fluency correct rates “有意に低下” (“significantly declined”), but offers no causal mechanism. Source: https://www.nier.go.jp/05_kenkyu_seika/pisa/pdf/2025/01_point.pdf (8 Sept 2026).

B8.4 — Norway (VG). “Barn leser mindre og sliter med konsentrasjonen” (B2.1).

B8.5 — United States (NCEE, as protective factor). Tracey Burns: US students show “relatively positive relationships with teachers, higher growth mindset, and more access to peer-to-peer tutoring,” possibly explaining smaller US declines. Source: Chalkbeat (B5.1).

B9. Migration, language background and demographic change

B9.1 — Germany (commentators asserting migration as main cause). Ferdinand Knauß, Cicero (8 Sept 2026): “Die Kompetenzen der Schüler sinken schon seit 2015, und der Zusammenhang mit der Asyl-Migration ist unübersehbar” (“Student competences have been falling since 2015, and the connection with asylum migration is unmistakable”); accuses PISA authors of ignoring it; no numerical evidence given. https://www.cicero.de/innenpolitik/pisa-studie-migration-bildungspolitik. Josef Kraus, Tichys Einblick (9 Sept 2026): “Das aktuelle desaströse Pisa-Ergebnis ist auch eines der vielen desaströsen Merkel-Ergebnisse” (“The current disastrous PISA result is also one of Merkel’s many disastrous results”); evidence offered: IQB 2021 fourth-grade trends, state-level shares of migrant-background pupils (Bremen 61.3%, Berlin 53.4%), and intelligence studies by Klauk, Rindermann and Henss. https://www.tichyseinblick.de/kolumnen/josef-kraus-lernen-und-bildung/pisa-ergebnis-merkel.

B9.2 — Germany (national PISA team rejecting migration as the dividing line). Samuel Greiff (TU Munich, German PISA lead): “Die Trennlinie läuft nicht zwischen Jugendlichen mit und ohne Zuwanderungshintergrund, sondern zwischen Jugendlichen aus sozial privilegierten und sozial benachteiligten Familien” (“The dividing line does not run between young people with and without a migration background, but between those from socially privileged and socially disadvantaged families”); “In Deutschland sind nur zwei von 100 sozial benachteiligten Schülern leistungsstark. Unter sozial Privilegierten sind es 25 von 100.” Source: https://www.migazin.de/2026/09/08/pisa-absturz-deutschlands-schueler-trennlinie/ (dpa, 8 Sept 2026). Olaf Köller (IPN Kiel) nonetheless points to language gaps among recent migrants from Arabic-speaking and African regions (Tagesspiegel, A2.5).

B9.3 — Germany (Philologenverband: pre-school German). Lin-Klitzing: “Wer mehr Chancengerechtigkeit will, muss dafür sorgen, dass Kinder die deutsche Sprache altersgemäß beherrschen, wenn sie eingeschult werden” (B4.6).

B9.4 — Sweden (Sweden Democrats). Jimmie Åkesson: “Språket är en grundförutsättning för lärande…” (“Language is a basic precondition for learning…”), framing immigration and language as the cause. Source: Vi Lärare (B1.2).

B9.5 — Finland (rejection by minister). Anders Adlercreutz on the immigration explanation: “Se on väärä väite, ja lähinnä ideologinen” (“That is a false claim, and mainly ideological”). Source: Pyhtäänlehti (B2.2). Suomen Uutiset (Finns Party outlet) ran the counter-headline “Antikainen: Maahanmuutto laskee myös Suomen Pisa-tuloksia” (https://www.suomenuutiset.fi/antikainen-maahanmuutto-laskee-myos-suomen-pisa-tuloksia/, not fetched).

B9.6 — Austria (FPÖ). Hermann Brückl: “Wenn Österreich beim Lesen mittlerweile sogar hinter der Türkei liegt” (“When Austria in reading now even lies behind Türkiye”), criticising the abolition of compulsory German-support classes. Source: ORF (B1.6). Unzensuriert.at headline: migrant pupils “mehr als drei Lernjahre hinten” (https://unzensuriert.at/342531-pisa-studie-schockiert-migrantenschueler-in-oesterreich-mehr-als-drei-lernjahre-hinten/, not fetched).

B9.7 — Estonia. Russian-language school transition: 47-point science gap between Russian- and Estonian-background students; reform began 2024 “without adequate teacher preparation” (paraphrase per Estonian World, B1.10).

B9.8 — Flanders (VRT). Students with migration backgrounds score lower; the effect diminishes when controlling for socio-economic status. Source: https://www.vrt.be/vrtnws/nl/2026/09/07/nieuw-pisa-onderzoek-bevestigt-dalende-trend-voor-wiskunde-en-le/ (8 Sept 2026).

B9.9 — Iceland. Iceland Review: 28% of foreign-origin pupils reach baseline reading; ministry says the gap “narrows when social and economic background is taken into account, but does not disappear” (B4.3).

B10. Socio-economic inequality, equity and funding

B10.1 — Germany (government and researchers). Karin Prien: “Was wir heute sehen, ist ohne jeden Zweifel ein besorgniserregender Befund”; “Das Aufstiegsversprechen im Land funktioniert im Moment nicht” (“The promise of social advancement is not working at the moment”). Ludger Wößmann (ifo): family background determines success more strongly in Germany than elsewhere; early tracking and late intervention. Sources: ZDF (B1.7); https://www.sonntagsblatt.de/artikel/epd/pisa-schock-ifo-oekonom-fordert-kurswechsel-der-bildungspolitik (8 Sept 2026). GEW (Anja Bensinger-Stolze): “Ein Vierteljahrhundert PISA-Maßnahmen haben unterm Strich nichts gebracht” (“A quarter-century of PISA measures has, on balance, achieved nothing”); socio-economic background explains 19% of variance vs 11.6% OECD; 125-point gap vs 85 OECD; deferred “financial and personnel problems.” https://www.gew.de/aktuelles/detailseite/pisa-schock-20 (8 Sept 2026).

B10.2 — United States (NCES, NYU). Matthew Soldner (Acting NCES Commissioner): the US has the largest top–bottom gap in reading among participants; “struggling students are not getting the full benefit of what we know how to do.” Michael Kieffer (NYU): “there was not a single country with significantly wider inequality in reading scores than the US … What is clear is that we are under-serving many students.” Sources: Education Week (B1.1); Al Jazeera (A2.1).

B10.3 — Sweden (opposition). Magdalena Andersson (S): “Vi måste utlysa nationellt nödläge för skolan” (“We must declare a national emergency for schools”); alleges the government blames everyone else while failing to fund schools, remove screens and cut class sizes; “Det är de senaste fyra åren som resultaten sjunkit dramatiskt.” Anders Ygeman (S): “ett haveri av Tidöpartiernas skolpolitik.” Sources: Vi Lärare; Skolledaren (B1.2). Sveriges Lärare (union) demands “more time for teaching, smaller groups and proper support for students” (https://www.sverigeslarare.se/om-oss/nyheter/pisa-2025/, page text not retrievable).

B10.4 — Scotland (EIS). Andrea Bradley: “nearly one child in four in Scotland living in poverty”; results reflect “levels of education funding, inequality, social conditions and public policy choices.” Source: B3.3.

B10.5 — Australia (press). Meyka: public-school underfunding, with the federal funding deal not fully effective until 2034 (B1.9).

B10.6 — Denmark (government and union). Heunicke: “Det kræver flere midler, og dem er vi fra regeringen klar til at investere”; Gordon Ørskov Madsen (Danmarks Lærerforening): “Det er investeringer, der skal helt ud i klasseværelserne.” Source: POV (B1.5).

B10.7 — Spain (PM). Pedro Sánchez blamed PP-run regions for not executing education funds (headline claim, https://www.elespanol.com/observatorio-educacion/20260909/sanchez-admite-preocupacion-resultados-informe-pisa-culpa-ccaa-pp-no-ejecutar-recursos/1003744377366_0.html, 9 Sept 2026, not fetched).

B10.8 — Finland (Left Alliance). Lohikoski: government cuts push municipalities to larger classes and fewer materials (B2.2).

B10.9 — France (unions). SE-UNSA blamed progressive state disinvestment: school and class closures, weakened RASED support, insufficiently paid staff, declining continuing education, inadequate resources, and expanding publicly subsidized private education (paraphrase per Café pédagogique, B1.3). Sgen-CFDT: “un soutien enseignant et familial plus faible que dans la moyenne de l’OCDE.” https://www.sgen-cfdt.fr/actu/pisa-et-maintenant-on-fait-quoi/ (8 Sept 2026).

B10.10 — Wales (minister, as strength). Anna Brychan: “our most disadvantaged learners compare well internationally.” Source: https://www.wales247.co.uk/new-pisa-results-reinforce-need-to-raise-standards-in-wales (9 Sept 2026).

B11. Teacher shortages, teacher quality and working conditions

B11.1 — Global (EI). Edwards: “While the global teacher shortage remains unaddressed”; four in ten students in schools where principals report shortages hinder instruction (B1.11). Run-up: Marjolaine Perreault (CSQ, Québec), EI, 17 July 2026: students with special needs “whose numbers in regular classrooms have grown exponentially … The workload associated with this for education personnel is very significant.” https://www.ei-ie.org/en/item/32756:pisa-2025-education-unions-get-ready-for-action-on-new-data.

B11.2 — Germany (DPhV). Lin-Klitzing: “Wo Fachlichkeit und das Prinzip der grundständig ausgebildeten Fachlehrkraft geschwächt werden, wirkt sich das unmittelbar auf die erreichten Kompetenzen der Schülerinnen und Schüler aus”; 52% of German students in schools reporting shortages vs 39% OECD (B4.6). Hessian Philologenverband (hphv), 9 Sept 2026: “Basiskompetenzen stärken, Lehrkräfte entlasten und Fachunterricht sichern” (title; body not fetched): https://www.dphv.de/2026/09/09/hphv-zu-pisa-2025-bildungsqualitaet-braucht-verlaessliche-rahmenbedingungen-hessischer-philologenverband-fordert-konsequenzen-aus-erneutem-leistungsrueckgang-basiskompetenzen-staerken-leh/.

B11.3 — Norway (union). Geir Røsvoll (Utdanningsforbundet): “Svaret er å styrke lærerprofesjonen” (“The answer is to strengthen the teaching profession”) (VG, B2.1).

B11.4 — France (Café pédagogique). Teachers’ pay “bien loin des 3000 euros nets annoncés tambour battant à la rentrée 2026” (“far from the €3,000 net announced with great fanfare at the start of the 2026 school year”) and recruitment difficulties throughout the Macron years (B1.3).

B11.5 — Estonia. 65% of students in schools where principals cite staffing shortages (Estonian World, B1.10).

B11.6 — Scotland (EIS). Scarcity of GTCS-registered teachers in early-learning settings (B3.3).

B11.7 — England (union, as credit). Paul Whiteman (NAHT): results reflect "the professionalism, expertise and dedication of the school workforce." Source: Tes, https://www.tes.com/magazine/news/general/pisa-2025-scores-revealed-england-uk-ranking (8 Sept 2026).

B11.8 — Canada/Alberta (as credit). Demetrios Nicolaides (Alberta Minister): results are "a testament to the hard work of students, teachers, families and school communities." https://www.todayville.com/alberta-students-rank-among-worlds-best/ (8 Sept 2026).

B12. Curriculum and pedagogy (knowledge vs skills, phonics, explicit instruction, inclusion, reform churn)

B12.1 — England/Scotland/Wales (knowledge-rich vs "progressive"). Toby Young, Daily Sceptic, 9 Sept 2026: credits Gove-era reforms — free schools and academies, phonics, spelling/grammar tests, maths mastery, EBacc, Progress 8 — and contrasts Scotland and Wales's "skills-based 'curriculum for excellence'"; "As Gove rightly concluded, there's no evidence schools can teach artistic expression, problem solving or character traits like resilience." https://dailysceptic.org/2026/09/09/what-more-evidence-does-the-blob-need-that-michael-gove-was-right/. Lindsay Paterson (Edinburgh): Scotland "has neglected knowledge, and celebrated skills as if skills could exist without knowledge," versus England's "structured knowledge" (B5.4). Greg Ashman (Filling the Pail, 8 Sept 2026): outcomes "continue to diverge in the United Kingdom, where broadly progressive education reforms have been adopted by Scotland and Wales while England has stuck with a traditionalist approach introduced by former education minister, Michael Gove"; Finland should no longer be promoted; Japan's 2022 rise attributed to "active learning" was contradicted by 2025 declines. https://fillingthepail.substack.com/p/a-first-look-at-pisa-2025. Duplicate: Ashman's England-over-Finland point as relayed by Joanne Jacobs (A2.4).

B12.2 — England (DfE). Department for Education: "strong results are not an argument for standing still," and "in reading and maths our children's scores in PISA assessments haven't improved in ten years"; cites Estonia to show "academic and vocational excellence can co-exist." Georgia Gould, Schools Minister: "We are determined to build on these positive results." Sources: Schools Week, https://schoolsweek.co.uk/pisa-2025-rise-across-the-board-but-science-gender-gap-opens/; Tes (B11.7). Academic notes: John Jerrim (UCL): England "pulls ahead" by "stretching" its highest-scoring students; Christian Bokhove (Southampton): "the whole distribution has become slightly more positive" (Tes). Jenni Ingram (Oxford, England report lead): "It's great to see pupils in the UK doing so well"; Mary Richardson (UCL): pupils in England, Wales and NI "reading just as well as they were in 2022." https://www.education.ox.ac.uk/news/pisa-2025-results-reveal-uk-students-knowledge-of-science-reading-and-maths/.

B12.3 — Scotland (government credits forthcoming reform). Màiri McAllan: "Scotland has stabilised its international performance, which is very welcome"; the Curriculum Improvement Cycle "is the largest change to Scotland's curriculum in over ten years." https://www.gov.scot/news/international-education-performance-competitive/ (8 Sept 2026). Opposition: Katherine Sangster (Labour): "Things have flatlined since 2022's historically poor results"; Duncan Dunlop (Lib Dem): "Once the best in the world, Scottish education just isn't what it used to be" (STV, B7.4).

B12.4 — Wales (minister). Anna Brychan: inherited "an education system that lacks clarity, coherence and a sure sense of direction"; Senedd Research notes the 2025 cohort was "the last cohort that did not study the Curriculum for Wales." Sources: Wales247 (B10.10); https://research.senedd.wales/research-articles/how-did-wales-perform-in-pisa-2025/.

B12.5 — France (reform churn vs 1980s pedagogy vs Macron reforms). Max Brisson (LR senator): "Les ministres passent et le système demeure"; blames ideological pedagogical choices of the 1980s; "L'école n'a pas besoin d'une réforme, elles ont toutes échoué." Pierre Ouzoulias (PCF senator): "dix ans de réformes macroniennes ont affaibli l'Éducation nationale." Laurent Lafon (centrist): "stop-and-go" policy. Source: Public Sénat (B1.3). Sgen-CFDT: "une politique éducative marquée par un pilotage centralisé et excessivement prescriptif," "l'accumulation des réformes," and making "la fluence un objectif en soi, alors qu'elle doit rester au service de la compréhension" (B10.9). Geffray: tested pupils "n'ont connu quasiment aucune des réformes menées depuis 2017" (reform-lag argument) and "On s'est beaucoup focalisés sur la réduction du nombre d'élèves par classe, nécessaire, peut-être au détriment de la réflexion sur la pédagogie" (Café pédagogique, B1.3). Diebolt (The Conversation): "trop de réformes tue les réformes" (B3.6).

B12.6 — Spain (LOMLOE). PP: Alberto Núñez Feijóo blames "socialismo educativo" and calls for evaluation and merit ("Evaluar no es castigar, evaluar es saber el grado de conocimiento"); PP regional governments (Galicia, Murcia, Extremadura, Aragón) called LOMLOE a failure in varying terms; Murcia's government: "suspenso rotundo y sin paliativos … la peor Ley de Educación que ha tenido el país." Government (Elma Saiz): "sería bastante simple culpar a la LOMLOE de estos resultados." Sources: El Plural (B1.4); https://www.infobae.com/espana/agencias/2026/09/09/sanchez-culpa-al-pp-del-mal-resultado-de-pisa-y-feijoo-senala-al-socialismo-educativo/ (headline only).

B12.7 — Finland. Purra's list (inclusion, open-plan, self-directed and phenomenon-based learning, "mercy passes") (B2.2). Adlercreutz credits enhanced learning support, more maths/reading hours and phone restrictions, with effects to come (Valtioneuvosto, B4.4).

B12.8 — Germany (curriculum). Olaf Köller (IPN): German curricula remain "verhaftet im 20. Jahrhundert" ("stuck in the 20th century") and fail to connect with young people's lives (Tagesspiegel, A2.5). Heidi Reichinnek (Die Linke): end "Bulimie-Lernen und Druck" (ZDF, B1.7).

B12.9 — Netherlands (curriculum standards). Judith Tielen, State Secretary: new core objectives implemented this school year are "much more concrete and ambitious than the old ones from 2006" (paraphrase per NOS, https://nos.nl/artikel/2630148-bijna-40-procent-15-jarigen-kan-niet-goed-lezen-laaggeletterdheid-dreigt, 8 Sept 2026). Gert Rijlaarsdam (Amsterdam), Neerlandistiek, 9 Sept 2026: advocates integrated reading-writing-discussion across subjects but stresses PISA cannot show whether Dutch reading instruction worsened (see B16.4). https://neerlandistiek.nl/2026/09/pisa-zes-jaar-later-reden-voor-zorg/.

B12.10 — Australia (mathematics curriculum; explicit teaching). Ashman: Australia's maths trajectory "deeply worrying," attributed to "poverty of our mathematics curriculum" (B12.1). Clare (ABC): "This shows why the reforms we are rolling out are so important. The UK results are impressive. It's also hard-earned. It's something almost 20 years in the making" (B1.9). Dissent: "Marty" (Bad Mathematics blog, 8 Sept 2026): "Whatever could have gone wrong? Maybe some explicit teaching will fix it?" — dismissive of both PISA and the explicit-teaching prescription; https://mathematicalcrap.com/2026/09/08/pisa-2025/.

B12.11 — New Zealand (credits). Stanford credits structured literacy and maths, an hour a day of reading/writing/maths, a knowledge-rich curriculum and early intervention (B6.3).

B12.12 — Flanders. Zuhal Demir (N-VA): "Zonder mensen met voldoende kennis bouw je geen sterke samenleving"; Bruno Vanobbergen (Catholic Education Flanders): vocational-stream reading deficits "beschamend"; Koen Pelleriaux (GO!): realigning secondary minimum objectives with primary standards (VRT, B9.8).

B12.13 — Poland (reform instability). Nowacka criticises the abolition of gimnazja as an example of disruptive reform and stresses political stability (B5.6).

B12.14 — Türkiye (credit to exam alignment and model). Yusuf Tekin, Minister: increased funding over 20 years, new classrooms, lower pupil–teacher ratios, transport and boarding, the FATİH project, phone restrictions (paraphrase per Ekonomim, 9 Sept 2026). Yelkin Diker Coşkun (Yeditepe): the LGS entrance exam has been closely aligned with PISA's context-based items for the last two or three years. Özgenur Korlu (ERG): alignment of central exams with PISA methodology; averages mask inequality. https://www.ekonomim.com/egitim/pisa-2025-turkiyenin-yukselisinin-arkasinda-neler-var-haberi-917178.

B12.15 — Japan (MEXT/NIER framing of strengths). NIER attributes Japan's standing to orderly science classrooms and "理科の授業における認知の活性化" (cognitive activation in science lessons, index 0.30 vs 0.00 OECD) (paraphrase per NIER summary, B8.3).

B12.16 — Singapore (MOE credits). MOE press release, 8 Sept 2026: results "affirm Singapore students' strengths in applying knowledge and skills to solve complex problems and being engaged with learning"; credits EdTech Masterplan, upstream support for weaker students; Liew Wei Li (Director-General of Education): "they are curious, they persevere even when learning gets hard." Also notes a widening top–bottom gap and weaker family support (49% weekly family discussion vs 60% OECD). https://www.moe.gov.sg/news/press-releases/20260908-pisa-2025-singapore-students-demonstrate-strong-real-world-problem-solving-skills-in-computational-thinking-domain. (The MOE quotations were not rechecked before publication: the site blocks automated access.)

B12.17 — Korea (ministry response). Ministry plans "수업과 연계한 독서교육을 활성화" and "질문과 토론 중심의 교실 수업 문화" (reading education linked to lessons; question- and discussion-centred classroom culture). https://www.seoulfocus.kr/news/articleView.html?idxno=308928 (8 Sept 2026).

B12.18 — Italy (minister credits system). Giuseppe Valditara: "La scuola italiana si è mostrata un sistema solido"; "Per la prima volta ci collochiamo significativamente al di sopra della media internazionale"; credits work on "criticità" and narrowing North–South gaps. Sources: https://www.orizzontescuola.it/pisa-2025-italia-sopra-la-media-ocse-valditara-diventiamo-un-riferimento-internazionale-lasciandoci-alle-spalle-vecchi-pregiudizi/; Il Giornale (B3.5). No named policy credited.

B12.19 — Ireland (minister). Hildegarde Naughton: "Ireland continues to perform strongly"; credits "sustained investment in science, technology, engineering and mathematics education" generally; acknowledges "a trend of decline that has been observed both here and internationally." https://www.gov.ie/en/department-of-education/press-releases/major-international-study-shows-irish-students-are-first-for-reading-third-for-science-and-fourth-for-mathematics-out-of-27-eu-countries/; https://www.irishexaminer.com/news/arid-41908009.html (8 Sept 2026).

B12.20 — Iceland (minister). Inga Sæland: results a "red card"; "We have reached the bottom … now there is no option other than going up" (paraphrase per Daily Northern, https://www.dailynorthern.com/31015/iceland-records-steepest-pisa-decline-among-participating-systems/, 9 Sept 2026); promises curriculum review and ISK 455 m for learning materials.

B13. Governance, structure and political framing

B13.1 — United States (federal bureaucracy; school choice). Linda McMahon: "our one-size-fits-all federal education bureaucracy has shortchanged our children and stifled our future … we must enact a hard reset that stops protecting a failed status quo and instead builds a system that empowers state leaders and embraces innovative learning options through school choice." Source: Al Jazeera (A2.1). Evidence offered: US reading at near-25-year low; largest top–bottom reading gap.

B13.2 — Germany (federalism/complacency). Greiff (dpa): complacency after the 2000 reforms, inadequate early childhood education, teacher shortages, digitalisation gaps (paraphrase per Migazin, B9.2). Anna Stolz (KMK President): "Wir fangen nicht bei null an" (ZDF).

B13.3 — Poland. Nowacka: "Na tle Europy jesteśmy zieloną wyspą" ("Against the European backdrop we are a green island"); lack of "stabilności politycznej" as the systemic weakness (Do Rzeczy headline; Głos, B5.6).

B13.4 — Wales. Brychan's "lacks clarity, coherence and a sure sense of direction" (B12.4).

B13.5 — Finland (data gap). Martin Paasi (Kokoomus): "Kun ei tutkita, ei tiedetä" ("If it isn't studied, it isn't known") — Finland lacks national tracking data on comprehensive-school performance (Pyhtäänlehti, B2.2).

B14. Family support, parents and early childhood

B14.1 — Germany. Prien: insufficient early-childhood/primary priority; declining parental reading (Tagesspiegel, A2.5). Katharina Dröge (Greens): the foundations must already be laid in kindergarten (reported speech, ZDF, B1.7). Sigrid Maurer (Austrian Greens) similarly (ORF, B1.6).

B14.2 — Singapore (MOE). Weaker perceived family support (B12.16).

B14.3 — Netherlands. Leescoalitie: reading stimulation “begint thuis” (B4.1).

B14.4 — France (ministry). “le recul du suivi de la scolarité par les parents” (B1.3).

B14.5 — Finland (SDP). Tuula Haatainen: focus on pre-school and primary; stronger parent–school collaboration (paraphrase; SDP release via STT, https://www.sttinfo.fi/tiedote/72302821/).

B15. Claims that the results are not meaningful, that PISA is flawed, or that sampling problems undermine comparisons

B15.1 — United States (sample bias). Chalkbeat: US results carry an asterisk; the OECD cannot “rule out bias” (B5.1). OECD’s own statement in A1.9.

B15.2 — Norway (exclusions may flatter results). Fredrik Jensen (Norwegian PISA project leader): “hvis deltagelsen hadde vært bedre, så kunne man jo anta at resultatene for de norske elevene ville vært enda svakere” (“if participation had been better, one could assume the Norwegian results would have been even weaker”); with 2015 exemption rates the decline would have been 4 points larger. Simon Malkenes: “de norske resultatene [kunne] potensielt vært enda verre.” Source: VG, https://www.vg.no/nyheter/i/n1l7ea/pisa-2025-hoeyest-andel-elever-med-fritak (8 Sept 2026). Evidence: 10.4% exclusion, highest in Norway’s nine cycles.

B15.3 — Denmark (inclusion of dyslexic students). Kjeldsen: including more dyslexic pupils had a small effect on scores but did not cause the 29-point reading decline (paraphrase of the Danish original); Tornhøj Christensen: OECD verified declines are real, not methodological (paraphrase per Folkeskolen, B1.12). Exemptions fell from 11.7% (2022) to 8.3% (2025).

B15.4 — Netherlands (PISA cannot say why). Gert Rijlaarsdam: “PISA vertelt dat er iets aan de hand is. Niet wat er aan de hand is” (“PISA tells us something is going on. Not what is going on”); “Uit: Nederlandse jongeren lezen slechter volgt dus niet automatisch: het Nederlandse leesonderwijs is slechter geworden”; digital distraction is noted by PISA authors as possible but unproven; he revises his earlier scepticism and now finds “reden tot serieuze zorg.” Source: Neerlandistiek (B12.9). NOS notes sample limitations (B12.9).

B15.5 — England (sampling). Schools Week: England’s 5,619 pupils in 183 schools “may not be entirely representative”; non-respondents were “slightly more likely to speak English as a first language, to have been eligible for free school meals in the last six years and to receive a form of special educational needs support”; “results should be interpreted cautiously, particularly when considering trends.” Source: B12.2.

B15.6 — Canada (non-response bias). CMEC: “the data from Canada, especially from Newfoundland and Labrador, Nova Scotia, Quebec, Ontario, Manitoba, Alberta, and British Columbia, require careful interpretation due to potential non-response bias.” https://www.newswire.ca/news-releases/oecd-report-ranks-canadian-students-among-top-performers-globally-in-science-844515596.html (8 Sept 2026).

B15.7 — Spain (Catalonia excluded). Público notes Catalonia’s results were excluded from international comparison because of a 23% exclusion rate (B1.4).

B15.8 — Singapore (PISA measures tuition, not schooling) — run-up. EveryChild.SG (Jom, 27 March 2026): “Does PISA really measure the strength of an education system? Or is it better understood as a gauge of a population’s willingness (and financial ability) to inflict extracurricular, for-profit drilling upon their offspring?”; evidence: S$764/month average tuition spend, class sizes of 33, PSLE at 12. https://www.jom.media/beyond-the-rankings-how-does-singapore-fare-against-pisa-peers/.

B15.9 — France (tests are not education). Diebolt: “les tests ne sont pas l’éducation et PISA n’est pas l’école”; rankings matter less than trajectories; “Être vingt-cinquième, vingt-septième ou trentième nous apprend finalement moins de choses que de savoir d’où l’on vient” (B3.6).

B15.10 — Australia (dismissal). “Marty,” Bad Mathematics: “I couldn’t give a stuff” (B12.10).

B15.11 — UK (against league tables). Tes notes the OECD “warns against compiling league tables of very different education systems” (B11.7).

B15.12 — Scotland (improvement is illusory). Paterson: the science recovery only restores 2018 levels and leaves a “14 months of learning” deficit against 2009 (B5.4).

B15.13 — Türkiye (coverage). Ekonomim notes the Turkish sample covers about 72% of the 15-year-old population (B12.14).

B15.14 — Iceland (press criticism of coverage). Vitinn (8 Sept 2026) faults mbl.is for stating without a source that Iceland “sitji neðst allra Norðurlanda og hafi dalað mest.” https://vitinn.news/grein/68c2f428-b46a-4057-9c21-3de41d4b21eb.

B16. Country-specific reactions crediting or blaming specific policies (cross-reference)

For convenience, the entries above that attach a result to a named policy: England — Gove reforms/phonics/knowledge-rich (B12.1), DfE (B12.2); Scotland — Curriculum for Excellence blamed (B12.1, B5.4), Curriculum Improvement Cycle credited (B12.3); Wales — Curriculum for Wales not yet in cohort (B12.4); Sweden — mobile ban, books, analogue early years credited; Tidö policies blamed by opposition (B1.2, B10.3); Norway — tablet policy blamed (B2.1); Finland — inclusion, phenomenon-based learning, digital blamed; learning-support reform credited (B2.2, B12.7); France — phone ban and post-2017 reforms credited/lagged; 1980s pedagogy and Macron reforms blamed (B1.3, B12.5); Spain — LOMLOE blamed by PP, defended by government (B12.6); Germany — Merkel 2015 migration policy blamed by commentators (B9.1), early tracking blamed by ifo (B10.1); Austria — abolition of German-support classes blamed by FPÖ (B9.6); Poland — gimnazjum abolition blamed (B12.13); New Zealand — structured literacy, attendance drive credited (B6.3, B12.11); Australia — explicit-teaching reforms and funding deal (B12.10, B10.5); Türkiye — LGS/PISA alignment and Maarif model credited (B12.14); Korea — reading and discussion-based instruction promised (B12.17); Japan — orderly classrooms and cognitive activation credited (B12.15); Singapore — EdTech Masterplan credited (B12.16); United States — federal bureaucracy blamed, school choice proposed (B13.1); Denmark — 100-school programme, books, teacher recruitment promised (B1.5); Iceland — curriculum review and materials (B12.20); Italy — none named (B12.18); Ireland — STEM investment generally (B12.19).

Part C. Frequency of explanation types (count of distinct third-party sources/speakers, excluding the OECD)

The count below tallies distinct speakers or organisations (not articles) in Part B that offered each explanation, and lists the countries in which it appeared. Counting is approximate where a speaker offered several explanations at once; no judgement about validity is implied.

Explanation type	Distinct third-party speakers/organisations	Countries where offered
Smartphones / social media / screen time / digital distraction	16	US (press), Sweden (2 ministers + govt), France (minister + ministry), Spain (2 officials), Denmark (minister + PM), Austria (minister), Germany (Streeck, Prien), Korea (ministry), Australia (minister), Estonia (press), EI (global)
Socio-economic inequality / underfunding / austerity	14	Germany (Prien, Wößmann, GEW, SPD, Greens, Linke), US (Soldner, Kieffer), Sweden (S, union), Scotland (EIS), Denmark (minister, union), Spain (PM), Finland (Left Alliance), France (SE-UNSA, Sgen-CFDT), Australia (press)
Curriculum/pedagogy: knowledge-rich or explicit instruction credited; skills-based/“progressive” reform blamed	9	England/Scotland/Wales (Young, Paterson, Ashman, DfE), Finland (Purra), Australia (Ashman, Clare, Joseph), New Zealand (Stanford), Flanders (Demir)
Reform churn / policy instability / governance	8	France (Brisson, Ouzoulias, Lafon, Sgen-CFDT, Diebolt), Poland (Nowacka), Wales (Brychan), Finland (Paasi)
Teacher shortages / teacher quality / workload	8	EI (global, CSQ), Germany (DPhV, hphv), Norway (Røsvoll, Nyhuus), France (Café pédagogique), Estonia (press), Scotland (EIS)
Migration / language background as main cause	6	Germany (Knauß, Kraus), Sweden (Åkesson), Austria (FPÖ, Unzensuriert), Finland (Finns Party outlet)
Migration explicitly rejected as main cause	3	Germany (Greiff), Finland (Adlercreutz), OECD (A1.6)
Digitalisation of schooling / tablets / ed-tech (as distinct from leisure screens)	6	Norway (Nordtun, Nyhuus, VG), Finland (Purra, Lohikoski), Sweden (govt), Austria (Wiederkehr)
COVID-19 long tail (as contributing cause)	6	US (Burns), Korea (ministry), Netherlands (NKO), Scotland (Paterson), Austria (IQS), Poland (Nowacka)
COVID rejected as sufficient cause	4	OECD (Schleicher, Cormann), Hendrick, Christensen, VIVE
AI chatbots / cognitive offloading	6	EI, US (Seiden), Scotland (EIS), Spain (Tolón), France (Diebolt), press in Estonia/Austria/Italy
Declining reading habits / attention / perseverance / motivation	9	Netherlands (Leescoalitie), Estonia (Sarapuu), Finland (Ahonen, Purra), Austria (IQS), Germany (DPhV), Denmark (Kjeldsen, Gissel), Norway (VG), Japan (NIER, descriptive)
Absenteeism / attendance	5	Australia (press), Denmark (Gissel, VIVE), New Zealand (Stanford, as credit), Norway (VG), Italy (press)
Classroom discipline / disruption	5	Australia (Bertling, Joseph), Denmark (Frederiksen), Poland (Nowacka), Scotland (Dunlop)
Family support / early childhood / language readiness	7	Germany (Prien, Dröge, DPhV), Austria (Maurer), Singapore (MOE), Netherlands (Leescoalitie), France (ministry), Finland (Haatainen)
Sampling, exclusions or response rates undermine comparisons	7	US (press), Norway (Jensen, Malkenes), Canada (CMEC), England (Schools Week), Netherlands (NOS), Spain (Catalonia), Denmark (Kjeldsen, partial)
PISA measures the wrong thing / cannot explain / tests are not education	5	Netherlands (Rijlaarsdam), Singapore (EveryChild.SG), France (Diebolt), Australia (Marty), UK (Tes on league tables)
Federal bureaucracy / school choice (US-specific)	1	US (McMahon)

Explanation type	Distinct third-party speakers/organisations	Countries where offered
Exam alignment or “teaching to PISA” as explanation of improvement	2	Türkiye (Coşkun, Korlu)

The most frequently offered explanation across countries is smartphones/screens/social media, which appears in ministerial statements in Sweden, France, Spain, Denmark, Austria, Australia and Korea and in the OECD’s own materials. Socio-economic inequality and funding is the second most frequent, offered mostly by unions, opposition parties and researchers, and is dominant in Germany. Curriculum-and-pedagogy explanations cluster in the anglophone countries (England, Scotland, Wales, Australia, New Zealand) and Finland. Migration-as-cause appears chiefly in German-language and Nordic right-of-centre commentary, and was explicitly contested in Germany and Finland. Sampling caveats were raised in every country the OECD asterisked (United States, Canada, Netherlands, New Zealand, Norway) and in England, Spain and Denmark.

Source list (all URLs cited above, for reference)

OECD: Volume I text (vol1.txt); US country note PDF (https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/09/pisa-2025-results-volume-i-country-notes_88d1164e/united-states_87c833c3/0c8cbc7c-en.pdf); https://oecdedutoday.com/the-state-of-global-education-according-to-pisa/; https://finance.yahoo.com/news/teen-reading-slumps-worst-century-111013754.html. United States: Chalkbeat; Education Week; K-12 Dive; Al Jazeera; joannejacobs.com; EI (three items). United Kingdom: Schools Week; Tes; Daily Sceptic; Oxford Department of Education; gov.scot; STV; EIS; Enlighten (Paterson); Wales247; Senedd Research; IBTimes UK (two items); Filling the Pail (Ashman). Germany: ZDF; Tagesspiegel; GEW; DPhV; Migazin; Cicero; Tichys Einblick; Spektrum; Sonntagsblatt (epd/ifo); newzs.de (FAZ). France: Café pédagogique; education.gouv.fr; Public Sénat; VousNousIls; Sgen-CFDT; The Conversation (Diebolt). Netherlands/Flanders: NOS; NKO; Neerlandistiek; CPNB/Leescoalitie; VRT. Nordics: Skolledaren; Vi Lärare; TV4; regeringen.se; Sveriges Lärare; VG (three items); Skolemagasinet; uvm.dk; Folkeskolen; VIVE; POV International; Pyhtäänlehti; Suomen Uutiset; Yle; valtioneuvosto.fi; Iceland Review; Daily Northern; Vitinn. Australia/New Zealand/Canada: ABC; Teacher Magazine (ACER); Karmactive; Meyka; Bad Mathematics; Scoop (Stanford); CMEC/Newswire; Todayville (Alberta). Asia: NIER (Japan); resemom; np-schools; edaily, Segye, Seoul Focus, Seoul Economic Daily (Korea); MOE Singapore; Jom (Singapore). Other Europe: Estonian World; gov.ie; Irish Times; Irish Examiner; SRF (Switzerland); ORF (Austria); Orizzonte Scuola and Il Giornale (Italy); Infobae, Público, El Plural, El Español (Spain); Głos Nauczycielski (Poland); Ekonomim (Türkiye).

Appendix G. Adversarial review of the draft, and the changes made in response

Reviewed: `report/main_report.md` (10,392 words) and `G/us_annex.md` (9,465 words), against `A/pisa_trends.csv`, the supporting sections in `/A`, `/B`, `/C`, `/D`, `/E`, the OECD text (`inputs/vol1.txt`), and the cited external sources. The brief was bias and omission, not style. Each finding gives the section, the passage, the standpoint from which it reads as tilted, the evidence that standpoint would cite (named), and a fix. Findings are ranked by how much they would change what a careful reader takes away. Numerical checks and the space-versus-evidence count follow; the overall judgment is at the end.

Method note on the web checks: web search was no longer available when this review began, so external sources were verified by direct fetch where a URL was known (Jakubowski et al. 2025; Beneito and Vicente-Chirivella 2022; UNICEF Report Card 18) and otherwise cited from the reviewer's knowledge with the citation given so the authors can check it. Those are marked "(not fetched)".

Short version

The report is careful, and most of its numbers are right. Where it leans, it leans in two directions at once. It leans towards the OECD's own reading of the decline (real, driven by engagement with long texts, digital in origin) because its predict-then-check method can only test what PISA measures, and PISA measures effort and in-school device use, not out-of-school phone hours, curriculum content or spending changes; the report acknowledges this for effort and not for the others. And it leans against the pandemic and spending harder than its own supporting sections do: the summary calls the pandemic a contributor of "a few points" while the causal section's verdict is "a third to a half of the mathematics decline since 2018", the spending verdict is "contradicted" though the prediction that matters (change in spending) was never tested, and the main report drops the annex's finding that US schools that stayed open all year still lost 0.17 SD. The largest single omission is that Sweden, used throughout as the flagship case for every "it isn't the pandemic, curriculum or funding" argument, has a 2018 baseline that Sweden's own National Audit Office found to be inflated by over-exclusion of newly arrived immigrant students. Three factual errors in the US material (one in the main report, two in the annex) favour, respectively, reading the US decline as gentler than it is and the NCLB accountability explanation.

Ranked findings

1. Sweden's 2018 baseline is unexamined, and Sweden carries more argumentative weight than any other country

Where. Main report, "Candidate explanations" (pandemic, curriculum, effort, TIMSS divergence); "Three facts constrain any combination"; Table 6; "How to read the results". The passage: "Native-born students in high-income systems with stable schooling, short or no closures, high spending and no major reform (Finland, Norway, Denmark, Sweden, the Netherlands) fell 35 to 45 points … the same cohort in several of these systems did not fall on TIMSS grade-8 mathematics." And: "Sweden's grade-8 mathematics rose 14 points on TIMSS while its PISA mathematics fell 38."

The problem. Every Swedish figure in the report is a change from 2018. In 2018 Sweden's overall exclusion rate was 11.1%, the highest in that cycle (PISA 2018 Table I.A2.1, in `A/raw/pisa2018/annexA2_888934028862.xlsx`, column 12), and its Coverage Index 3 dropped to 0.857 from 0.936 in 2015 before recovering to 0.891 in 2022 and 0.915 in 2025 (`pisa_trends.csv`, `coverage_index_3`). The Swedish National Audit Office (Riksrevisionen, RiR 2021:12, *Pisa-undersökningen 2018 – arbetet med att säkerställa ett tillförlitligt resultat*, not fetched) found that Sweden had excluded newly arrived immigrant students who should have been tested and that the OECD's follow-up review was inadequate. The 2018 Swedish mean is therefore the one cycle in the series where the tested population was narrowest, and Sweden's 2018-to-2025 declines of 39 (reading) and 38 (mathematics) are measured from that point. Sweden's 2015-to-2025 changes (reading 500 to 466, −34; mathematics 494 to 464, −30; Table 3 of the trajectory section) are smaller and start from a properly sampled cycle. This does not remove the Swedish decline; it means a slice of it, plausibly several points, is a baseline artefact of exactly the kind the validity section catalogues for Canada, the Netherlands and New Zealand, and the report never mentions it. It matters for three standpoints at once: the measurement reader (a baseline sampling artefact), the demographic reader (the excluded students were immigrants, so the 2018 figure is compositionally unlike 2025), and the pandemic reader (Sweden's "0 full closure weeks against a 38-point fall" is the report's single most-used datum against closures).

Fix. Add Sweden 2018 to Table 4's companion paragraph on the 2022 baseline ("The 2022 baseline had more problems"), quote the 11.1% exclusion rate and the coverage dip, cite Riksrevisionen 2021:12, and either restate the Swedish comparisons from 2015 or attach the caveat every time Sweden is used as the decisive case (four places in the causal section, once in the constraining facts, once in the conclusion).

2. The pandemic is treated more dismissively in the synthesis than in the sections it synthesises, and the strongest published counter-evidence is missing

Where. Summary: "Four are contradicted as general accounts while surviving as contributors of a few points or as explanations for particular countries: the pandemic, demographic composition, spending, and a reversal of the long-run rise in cognitive test scores." Causal section, pandemic verdict: "perhaps a third to a half of the mathematics decline since 2018 and less of reading." Table 8: "A third to a half of mathematics since 2018."

The problem, in four parts.

(a) A third to a half of a 24-point mathematics decline is 8 to 12 points, not “a few”. The summary sentence groups the pandemic with spending (verdict: contradicted, “for nothing identifiable” in the causal review) and the Flynn reversal. A pandemic-focused reader will regard the summary as misrepresenting the section beneath it; the causal review’s own closing paragraph uses the same “few points” phrase, so the error originates there, but the synthesis is where readers stop.

(b) The report’s closure test (“The closure-length prediction fails (rho +0.24 across systems, the wrong sign; +0.14 points per week with GDP controlled, SE 0.13”) is presented as decisive without citing the published study that reaches the opposite conclusion on the same UNESCO closure data and the same PISA 2022 outcome: Jakubowski, Gajderowicz and Patrinos, “COVID-19, school closures, and student learning outcomes. New global evidence from PISA”, *npj Science of Learning*, 2025 (https://www.nature.com/articles/s41539-025-00297-3, fetched), which finds about 0.44 points of mathematics loss per week of full closure, losses of 9 to 12% of an SD in the shortest-closure countries and 17 to 23% in the longest, and effects concentrated among boys, immigrant and disadvantaged students. The difference from the report’s null is specification (they net out pre-pandemic country trends, which the report’s cross-sectional regression does not). The report must at least cite it and say why its own specification is preferred.

(c) The closure correlation excludes Hong Kong, Macao and Chinese Taipei because UNESCO does not cover them (trajectory section, section 5). Hong Kong is the third-largest 2018-to-2025 three-domain decliner in Table 6’s set (−31.6) and had among the longest school closures of any high-income system (2020 to 2022, under zero-COVID). Omitting the one rich system with both long closures and a large decline biases the “no gradient within rich systems” finding in the direction the report reports. This should be said.

(d) The annex reaches a different conclusion for the United States and the main report does not carry it. Annex section 3.1: “even the most in-person states show losses, so closure length explains variation across states, not the national floor”, citing Goldhaber, Kane et al. (schools in person all year still lost about 0.17 SD). That is the strongest reply to the report’s closure test, because it says the pandemic’s effect is not indexed by closure weeks at all; it is nowhere in the main report, which treats “no closure gradient” as equivalent to “no pandemic effect”. The report’s own hedge in the causal review (“The absence of a closure gradient shows only that the marginal effect of additional weeks is invisible at country level with a coarse measure”) was also dropped in the synthesis.

Fix. In the summary, replace “contributors of a few points” with the section’s own range for the pandemic. In the pandemic paragraph, cite Jakubowski et al. 2025 and state the specification difference; note the Hong Kong omission; add one sentence carrying the Goldhaber finding and its implication that closure length is a poor proxy for pandemic exposure. Keep the verdict; it is defensible. Change its framing from “contradicted as the main cause” to “contradicted as the sole cause; a 10 to 15-point mathematics deficit remains the best-identified single component”.

3. Three factual errors in the US material, two of which favour a particular reading

(a) Main report, “The picture since 2000”, on the United States: “none of its window changes is significant, because its standard errors are about 7 points, but its 2018-to-2025 changes in reading (−15.6) and mathematics (−15.0) are.”

Both parts are wrong against the CSV. The US 2022-to-2025 reading change is −14.42 with SE 7.35 (Table I.B1.2a.36), z = 1.96, and is flagged significant in the OECD workbook’s own conditional formatting (`change_to_2025_sig = 1`) and reported as “sig” in the annex’s Table 2 and in the NCES statement the annex quotes (“lower than its scores in the two most recent assessment cycles”). The 2018-to-2025 changes are −15.8 (reading) and −15.4 (mathematics), as Table 6 of the same report correctly shows; the −15.6 and −15.0 are transcription errors inherited from `trajectories.md` (which also gives the 2022-2025 reading change as −13.8). The effect of the sentence as written is to make the US look flatter than the OECD’s tables say. The D2 placement is unaffected (the rule needs two significant domains).

(b) Annex 3.6 (NCLB): “The NCLB gains are among the better-identified effects in the US record … and they are the only period in which the bottom of the US distribution rose.”

False on the annex’s own dataset. In `us_series.csv`, the NAEP grade-8 mathematics 10th percentile rose from 215.0 (1990) to 226.7 (2000) and 229.9 (2003) before NCLB took effect, a larger rise than the 229.9 to 237.4 of 2003-2013; the LTT age-13 mathematics 10th percentile rose from 213.3 (1978) to 233.8 (1999). The claim as written is an argument for the accountability explanation that the data do not support. Table 5 of the annex starts at 1999/2000, so the reader cannot see this.

(c) Annex 3.6 and 1.1, minor: “Grade-12 … mathematics −3.3 over 2019-2024”: 150.3 − 146.9 = 3.4 in the series file. Trivial, but the annex quotes both −3.3 and −3.4 in different places for the same quantity.

Fix. Correct (a) to “only one of its three-year window changes, reading 2022-2025 (−14.4, SE 7.3), reaches significance, and only just”; correct the two 2018-2025 values. Delete the “only period” clause in (b) and replace with “the bottom of the distribution rose through the 1990s and 2000s and turned down after 2013”. Reconcile (c).

4. The 2025 contractor change is larger than the report says

Where. “How much weight the results can bear”: “the whole assessment ran on a new delivery platform from a new contractor, mentioned only in two endnotes on timing thresholds.”

The problem. Volume I's acknowledgements (lines 271-278) and Annex A2 (lines 41158-41172) state that the Core 1 contractor for PISA 2025, responsible for "instrument development, scaling and analysis, and all data products" and "migration of trend units", was ACER, with Open Assessment Technologies building the platform. ACER ran PISA from 2000 to 2012 under the Rasch model; ETS ran 2015 to 2022 under the 2PL/GPCM model with partial invariance, the change that Robitzsch et al. (2020) and Jerrim et al. (2018) analysed. The report's validity section describes the 2015 contractor and model change in detail and then describes 2025 as a platform change. A measurement-artefact reader will say the 2025 change is of the same kind as 2015's, in reverse, and that Volume I nowhere states whether the ETS scaling model and item parameters were retained. The reader would also note that the science framework was developed by Oxford University Press under contract (Core 2), which the PISA-as-agenda reader will regard as worth a sentence.

Fix. State the contractor change for scaling explicitly, name ACER and ETS, and add "whether the 2015-2022 scaling model, item parameters and partial-invariance decisions were carried over is not stated in Volume I" to the list of things the Technical Report must document.

5. The effort explanation is given the strongest correlations but not the statistics that qualify them

Where. Causal section, effort paragraph: "A regression of the reading change on the effort change gives 17 points per effort point (SE 7.3), which at the OECD-35 effort change of −0.52 attributes about 9 of the 24.6 points to effort."

The problem. The regression's R^2 is 0.07 (`E/computed/analysis_log.txt`, section 1: "n=69, R2=0.07"). A slope that explains 7% of the cross-country variance in reading change is being used to attribute 36% of the OECD-average decline. The report calls the slope "an upper bound" but never gives the R^2; a reader from any standpoint who dislikes the effort story (and the phone, pandemic and curriculum readers all do, for different reasons) will cite it. Two further omissions from the validity review that the synthesis dropped: the 2026 multilevel study of PISA 2022 finding no country-level association between self-reported effort and achievement ($\beta = -0.01$, $p = 0.93$; *Large-scale Assessments in Education*, cited in the validity review), and Michaelides and Ivanova's finding that rapid guessing correlates positively with country means. Both are in `/D`; neither is in the main report, which keeps only the evidence in the effort story's favour (Gneezy et al., Borghans and Schils, Zamarro et al.) and the Borgonovi "effort is a trait" point.

Fix. Report the R^2. Add one sentence on the null cross-sectional country-level findings. Keep the "a few points to a third" bound.

6. "Advantaged students fell most" is load-bearing and rests on an index the report itself says was recomputed

Where. Summary, "What the 2025 results show", causal section (used against the pandemic, composition, spending and mental-health accounts and in favour of smartphones and effort), and the "three facts".

The problem. The report notes in passing that "ESCS-quarter means for 2015 and 2018 differ between the two workbooks by up to about 4 points for individual countries, which the 2025 volume does not comment on" (build log: the US 2018 reading gap is 98.8 in one workbook and 96.4 in the other). It never connects this to the finding it leans on most. ESCS quarters are national and relative; if the index was re-derived for 2025 (PISA revised ESCS in 2022 and the 2025 workbook restates 2015 and 2018 values), the "top quarter" of 2025 is not the same population slice as the top quarter of 2018, and a 2 to 4-point restatement is a material fraction of the 5 to 7-point difference between the two quarters' 2022-2025 declines (top −20.4, bottom −3.5 in reading; the gap between the changes is what the argument uses). The measurement reader and the inequality reader will both raise this, from opposite motives.

Fix. Say, where the finding is first stated, that it depends on an index the OECD re-derived without documentation, and give the size of the restatement; note that the 2022 values are identical in both workbooks so the 2022-2025 comparison is on a consistent index and the 2018-2022 comparison is not.

7. Spending: the verdict is sharper than the test

Where. Causal section: "Spending and austerity … Verdict: contradicted." Table 8: "Fits: None." Summary: "the evidence supports none of these at the scale the claims imply."

The problem. The causal review states: "I did not construct a change-in-spending variable (Education at a Glance indicator C1 gives it; the OECD pages could not be fetched programmatically), so the lagged prediction is untested beyond these cases." The synthesis drops that sentence and keeps the verdict. What was tested is spending *level* (2019) against change, which the report itself says is "the same fact as 'higher-scoring systems fell more'". An untested prediction cannot be "contradicted"; the honest verdict is "level uncorrelated in the predicted direction; change not tested; a few case comparisons (UK, Italy, Spain versus Finland, Norway, Denmark) run against the hypothesis". The funding reader would add three things the report omits: UNICEF Innocenti Report Card 18 (2023, https://www.unicef.org/innocenti/reports/child-poverty-midst-wealth, fetched), which finds child income poverty rose by at least 10% between 2012-14 and 2019-21 in France, Iceland, Norway, Switzerland and the United Kingdom (three of them among the largest PISA decliners) while falling by more than 30% in Poland, Slovenia, Latvia and Lithuania (two of them the largest decliners), so the evidence cuts both ways and should be shown rather than left out; that the feature set in the trajectory analysis contains no measure of private or voucher-school share or of grade inflation, though "the market school" is a named explanation in the inventory (B10.3) and Sweden, the Netherlands, Denmark and Chile all combine high non-public shares with large declines; and, in the annex, that Jackson, Wigger and Xiong's recession-cut effects fall on cohorts schooled during 2009-2015, which are precisely the NAEP grade-8 cohorts of 2017-2024, so "cuts coincide with the NAEP peak, not the decline" mis-times a lagged effect.

Fix. Restore the “not tested” sentence; change the verdict to “not supported by level data; change untested”; add the UNICEF trends in one sentence with both directions; note the absent private-share feature as a limitation; reword the annex spending paragraph to acknowledge the lag.

8. Smartphones: the report says the relevant quantities are unmeasured, but external measures exist, and two ban studies with the “right” sign are missing

Where. Causal section: “the quantities the hypothesis is about, out-of-school hours and age at first phone, are not in the trend tables”; “The causal literature on school bans gives effects of 0.02 to 0.14 SD, concentrated among low achievers”. Summary: “The public debate has been dominated by smartphones, funding and curriculum, and the evidence supports none of these at the scale the claims imply.”

The problem. (a) The phone reader will cite Beneito and Vicente-Chirivella, “Banning mobile phones in schools: evidence from regional-level policies in Spain”, *Applied Economic Analysis*, 2022 (https://ideas.repec.org/a/eme/aeapps/aea-05-2021-0112.html, fetched), a PISA-based difference-in-differences on Galicia and Castilla-La Mancha finding gains equivalent to 0.6 to 0.8 years in mathematics and 0.7 to about 1 year in science and reduced bullying; it is the one ban study on PISA scores and the one with an effect of the size the debate assumes, and it is absent from the causal review and the report. (b) Abrahamsson’s Norwegian study (in the causal review) found effects on girls’ GPA, mental-health consultations and bullying; the report’s one-line summary (“0.02 to 0.14 SD, concentrated among low achievers”) loses the mental-health and bullying results, which bear on the report’s own finding that bullying change correlates at −0.43 with reading change. (c) Out-of-school time-use series exist and are not cited: the Common Sense Census (2015, 2019, 2021: US teen entertainment screen time 6h40 to 8h39 a day; not fetched), Pew 2024 (in the annex only), WHO/HBSC 2022 problematic social-media use 7% to 11% (in the causal review, dropped from the report). The claim that the hypothesis is untestable is true of PISA and not of the evidence base. (d) The summary’s “supports none of these at the scale the claims imply” is inconsistent with the smartphone verdict two sections later (“compatible with anything from near zero to most of the decline”). A phone reader will say the report contradicts itself; a phone sceptic will say the conclusion (“An explanation that fits all three has to operate outside school …”) leans towards phones on a verdict the summary had just called unsupported.

Fix. Cite Beneito and Vicente-Chirivella and say why a single regional study should not carry the estimate; restore the Abrahamsson detail; cite the time-use series; make the summary sentence and the verdict say the same thing (for instance: “the evidence is compatible with a large role for screens and does not establish one”).

9. Curriculum: the within-UK natural experiment was dropped, and the largest decliners get no country-level scrutiny

Where. Causal section, curriculum paragraph; Table 6 (no Scotland, Wales, Latvia, Slovenia, Iceland, Denmark rows); “What has been claimed publicly”.

The problem. (a) The causal review calls the England-Scotland-Wales contrast “the strongest natural experiment in the set”; the synthesis mentions Scotland and Wales only as speakers in the claims inventory. Whether one believes the knowledge-rich account or not, that is the test a curriculum reader most wants to see, and it is missing. (b) The curriculum paragraph never discusses Latvia (largest 2022-2025 decliner, −45 reading), Slovenia (second-largest since 2018, −51), Iceland (largest since 2018, −52), Denmark (−41) or Norway (−47) at the country level, though each had a reform in the right window that a reform-blaming reader will name: Latvia’s *Skola2030* competence-based curriculum (phased in 2020-2023, so the 2025 cohort is the first schooled under it in lower secondary) and its transition to Latvian-only instruction (from 2023); Denmark’s 2014 *folkeskolereform* (longer days, “learning goals”) and 2012 inclusion law, on which VIVE’s evaluations found no achievement gains (not fetched); Norway’s LK20 (2020) and its one-to-one device policy; Finland’s 2011 three-tier inclusion reform, which Purra’s list names and the report never tests. The report’s case-by-case method therefore examines the countries where reform timing does not fit (Finland 2016, Australia 2012, Sweden 2011) and not the countries where it might. (c) The reverse tilt exists too. The England paragraph puts “consistent with the reforms having protected England” first and the sampling explanation second, and Türkiye’s exam alignment is called “the clearest case of a national policy that predicts its PISA trend” without noting that a knowledge-rich advocate and a teaching-to-the-test critic will read the same sentence in opposite ways. The Mississippi treatment in the annex is balanced.

Fix. Restore the UK-nations contrast with its confounds (a paragraph exists in `/E`). Add one paragraph on Latvia, Slovenia, Iceland, Denmark and Norway with the reforms and their timing, stating for each whether the cohort was exposed. Reorder the England sentence so the two readings are given equal weight.

10. Composition: three omitted facts, two of which help the demographic reader and one of which hurts

Where. “What the 2025 results show” and the composition paragraph.

The problem. (a) The immigrant share nearly doubled between 2018 and 2025 in Iceland (5.6% to 10.1%), Malta (8.8% to 17.1%) and rose by more than five points in Slovenia (8.9% to 14.1%), Norway (12.4% to 18.7%) and the United Kingdom (19.8% to 27.7%) (`system_features.csv`); the report names only Germany, Israel and Portugal. (b) Denmark’s exemption rate fell from 11.7% (2022) to 8.3% (2025) as more dyslexic pupils were included (claims inventory B15.3); the report’s statement that “rising exclusions work the other way” is true for Norway and the Netherlands and the reverse holds for Denmark, the fifth-largest 2022-2025 decliner, and is not mentioned. (c) Against the demographic reader: the OECD-29 series in the CSV shows immigrant students’ reading falling by 24.6 points over 2018-2025 (459.2 to 434.5), the same as non-immigrants’ 22.3, so the composition effect is purely a share effect; the report could say this in one clause and does not. (d) Israel, used twice as a composition and TIMSS example (−38 reading 2022-2025), tested in spring 2025 after eighteen months of war-related disruption, reservist mobilisation and displacement; the report offers no country context for it at all.

Fix. Give the share changes for the five systems; add the Denmark exemption fact next to the Norway/Netherlands sentence; add the immigrant-student decline; one sentence of context for Israel.

11. The governance critique of PISA disappears between the validity review and the synthesis

Where. "How much weight the results can bear"; "How this report was produced".

The problem. The validity review has a section, "The broader critiques and the OECD's responses", covering Meyer and Benavot, the 2014 open letter to Schleicher, Sjøberg, Zhao and Hopfenbeck, and it records that the Preface "moves directly from results to policy prescriptions … the practice the governance critics object to". None of this is in the main report. The reader who believes PISA is used to push a policy agenda will also note: that the same volume that warns about screens launched the "Learning in the Digital World" innovative domain (line 59) and that the OECD's own Education 2030 project promoted the competence-based curricula that the reform-blaming reader indicts, neither of which the report mentions; that the report adopts OECD vocabulary throughout ("baseline of proficiency", "hasty readers", "advantaged strugglers") without noting that "hasty" is a construct on re-normalised thresholds; that its method (predict-then-check against PISA's questionnaire) structurally favours explanations PISA measures, a point the report makes for effort ("because it is the only one PISA measures directly") and not for the others; and that "How this report was produced" asserts, in the past tense and before this review existed, that the draft "was subjected to an adversarial review … to remove any sharpening of hedged findings", which reads as self-certification.

Fix. Restore a short paragraph from `/D` on the governance critique, with the OECD's responses; add the Learning in the Digital World and Education 2030 points as one sentence each; rewrite the production note after this review, listing what was changed.

12. The pre-pandemic reading decline is stated as fact where its individual steps are within link error

Where. Summary: "the decline did not begin with the pandemic: on the aggregate of 23 countries with data in every cycle since 2000, reading was flat until 2012 and has fallen in each cycle since". Section 1: "declined in each subsequent cycle: −4.1 to 2015, −4.4 to 2018".

The problem. The 2012-to-2015 step crosses the paper-to-computer transition that the validity section calls "the weakest [link] in the series" (reading link error 6.1 to 2012), and neither −4.1 nor −4.4 is individually significant on the OECD's link errors; the combined 2012-to-2018 OECD-23 change (−8.5, SE about 3.8 using the 3.74 link error the trajectory section quotes from PISA 2018 Table I.A7.1) is significant only marginally ($z \approx 2.2$), and OECD-35 reading 2015-to-2018 (−2.7) is not. The measurement reader will say the "pre-pandemic decline" on the OECD average is one marginal step across a mode change; the pandemic reader will say the same. The country-level evidence (Finland, Australia, the Netherlands, Korea) is much stronger and should carry the claim.

Fix. Qualify the summary sentence ("fell in each cycle since, though the two pre-2018 steps are within link error individually and the first crosses the 2015 mode change; the pre-pandemic decline is clearest at country level in …").

13. A Pearson correlation is reported in a paragraph of Spearman correlations

Where. "The picture since 2000": "across all 75 systems with 2018-to-2025 data the 2018 level correlates at −0.46 with the subsequent change, and within the 37 OECD members at −0.23, not significant." Preceding sentence gives Spearman −0.25, −0.28, −0.32.

The problem. Recomputed from the CSV, the Spearman correlation between the 2018 three-domain mean and the 2018-2025 three-domain change is −0.33 ($n = 75$, $p = 0.004$; −0.30 without Zambia); −0.46 is the Pearson. Within the OECD, Spearman −0.16 ($p = 0.36$), Pearson −0.23 ($p = 0.18$). The Pearson value is inflated by Zambia and Cambodia (PISA-for-Development baselines, coverage 0.36 and 0.28). The substantive point (starting level discriminates, more than any feature) survives at −0.33 but the "regression to the mean" reading, which the report offers as one of four possibilities, looks weaker.

Fix. Label the statistic; give both; note the influence of the two PISA-D systems.

14. Testing-date changes as an unmentioned comparability issue

Where. "How much weight the results can bear" lists three 2025-specific issues (link, science test, sampling).

The problem. Box I.2.2 (vol1.txt lines 13952-13965) says its new learning-gain estimates exploit "countries/economies where testing dates changed over the 2015-2025 period" and lists ten: Austria, Brunei, Cambodia, Ireland, the Netherlands, Serbia, England, Northern Ireland, Scotland and the United States. A testing-date shift changes the grade composition of the age-eligible sample; the OECD uses this to estimate a 12 to 28-point yearly gain. The report never mentions that these ten systems' trends are across a change in testing date, and the annex never mentions it for the United States.

Fix. One sentence in the validity section and one in annex 1.3.

15. The annex omits Table I.2.7's US-specific "year of learning" estimate where it bears directly on a quoted claim

Where. Annex 1.1 and claims inventory B1.1: McMahon's "14-point reading drop is equivalent to a seven-month loss in learning".

The problem. Table I.2.7 (vol1.txt lines 13990-14030) gives the US yearly gain as 13.8 points in science and 12.0 in reading, among the lowest of the ten systems (Serbia 13.7/11.4; England 24.5/24.0). On the OECD's US-specific estimate, a 14-point reading fall is more than a year, not seven months. The annex has the table in its sources and does not use it; the main report cites the table for the UK and Serbia only.

Fix. Add the US row to annex 1.1 with the caution Box I.2.2 attaches.

16. Smaller items

- Main report, "What has been claimed publicly": "reactions from the OECD and from twenty-six countries". The inventory's coverage list has 25 (counting the United Kingdom once).
- Main report, effort paragraph: "the cross-test prediction … not for Israel or the United States" omits Hungary, Chile and Hong Kong, which the causal review lists as systems where PISA and TIMSS fell together; leaving them out makes the Nordic pattern look more general than it is.
- Main report, composition paragraph, drops the causal review's two cautions (the native/all difference includes larger declines among immigrant students themselves; the background means use the sub-sample with valid data).
- Main report, section 1, reports the science 10th-percentile fall without the validity section's caveat that adaptive designs (science newly adaptive in 2025) change precision at the tails.
- Annex 3.7 and 2.3 do not mention that NAEP's inclusion rates for students with disabilities and English learners rose after the 2010 NAGB inclusion policy, a compositional factor a demographic reader would add to the race decomposition (it works in the same direction as the annex already notes for English learners).
- Annex 3.9 tests no US-specific spending series (per-pupil real spending fell 2009-2013 and did not regain 2008 levels nationally until about 2017; Census F-33/NCES Digest, not fetched); "which I have not tested" is honest but the section is then 80 words against 971 for the Common Core.

Standpoint summary

Standpoint	Where the report reads as tilted	Omitted evidence the reader would cite
Measurement artefact	Understates the 2025 contractor change (finding 4); states the pre-2018 decline as fact (12); never applies its own baseline-bias reasoning to Sweden 2018 (1); ESCS restatement not connected to the "advantaged fell most" claim (6); science tails reported without adaptive caveat (16)	Riksrevisionen 2021:12; the 2026 country-level null on effort (in `/D`); R^2 = 0.07; Box I.2.2 testing-date changes (14)
Real, caused by phones	Summary says the evidence supports none of the popular explanations at scale while the verdict allows "most of the decline" (8d); "quantities unmeasured" when external series exist (8c)	Beneito and Vicente-Chirivella 2022; Abrahamsson's mental-health results; Common Sense Census; HBSC 2022
Real, caused by pandemic closures	Summary's "a few points" versus section's "third to a half" (2a); closure test presented as decisive (2b); Hong Kong excluded from closure data (2c); annex's "in-person schools lost 0.17 SD" not carried up (2d); grade-4 signal described as "something happened … without saying what" when it is the pandemic's predicted signature	Jakubowski, Gajderowicz and Patrinos 2025; Engzell, Frey and Verhagen (PNAS 2021: 0.08 SD loss in the Netherlands after eight weeks, not fetched); Israel's 2023-25 disruption (10d)
Curriculum (reform caused decline)	Case selection examines the reforms that don't fit and not the ones that might (9b); inclusion policies never tested	Latvia Skola2030; Denmark 2014 reform and 2012 inclusion law; Norway LK20; Finland 2011 three-tier support
Curriculum (report too quick to dismiss, or too kind to explicit instruction)	Within-UK contrast dropped (9a); England "protected" listed first; Türkiye "clearest case" (9c)	The `/E` UK-nations paragraph; Machin, McNally and Viarengo (in `/E`, not in report)
Demographic change	Only Germany, Israel and Portugal named; Sweden 2018 exclusion story absent (1); Denmark's falling exemptions absent (10b); share doublings in Iceland, Malta, Slovenia absent (10a)	Riksrevisionen 2021:12; Heller Sahlgren IFN PP71 and Skolverket 2016 (in `/E`, dropped)
Funding and inequality	"Contradicted" on an untested change prediction (7); no private-share or grade-inflation feature; annex mis-times recession cuts	UNICEF Report Card 18; Jackson, Wigger and Xiong on exposed cohorts; the OECD's own 2018-22 gap widening, which the report reports but does not weigh
PISA as policy instrument	Governance critique dropped (11); OECD vocabulary adopted; method favours PISA-measured explanations; production note self-certifies	Meyer and Benavot 2013; 2014 open letter; Sjøberg 2015; Zhao 2020; LDW domain; Education 2030

Numerical checks

Twenty-four claims, chosen across sections; nine from the annex. "Source" is what the claim was checked against.

#	Location	Claim	Source	Result
1	Summary	OECD-35 reading 2022-25 −14.3 (CI −16.9 to −11.8); mathematics −9.4 (−12.0 to −6.7); science −2.8 (SE 3.2)	CSV `OECD average-35`	Pass (−14.34, SE 1.28; −9.36, SE 1.37; −2.82, SE 3.18)
2	Table 1	OECD-23 series, all nine cycles and three change rows	CSV `OECD average-23`	Pass (all cells match to one decimal)

#	Location	Claim	Source	Result
3	Table 2	Fitted trends −27.0 (3.3), −23.9 (3.1), −6.7 (3.1)	CSV `decennial_trend_2015_2025`	Pass
4	Section 1	Latvia 2022-25 reading −44.9 (SE 3.4); eight largest 2018-25 reading declines in that order	CSV	Pass (−44.89, 3.37; Iceland −51.7, Slovenia −50.6, Latvia −49.0, Norway −46.6, Finland −45.8, Cyprus −44.9, Hong Kong −44.2, Netherlands −43.4)
5	Section 1	Eight systems up and nine down in all three domains 2022-25	CSV significance flags	Pass (identical lists)
6	Table 3	Counts by base year	CSV	Pass; 2022 base uses the 74-system Table I.2.6 set, annex tables give 75/76 as the note says
7	Section 1	Cyprus mathematics −6.5, significant	CSV	Pass (−6.53, SE 2.18)
8	Section 1	Science p10 2018-25 −12.6, p90 +2.4 (3.6); 259-point inter-decile range	CSV	Pass (−12.63; +2.39, 3.61; 611.8 − 352.6 = 259.2)
9	Section 1	Below Level 2: reading 25.7 to 30.4, mathematics 30.0 to 33.9, science 23.8 to 25.0	CSV	Pass
10	Section 1	ESCS: reading gap 87.9 to 93.4; 2022-25 top −20.4, bottom −3.5; science bottom +5.2 (3.2); mathematics bottom −2.6 (1.5)	CSV	Pass
11	Section 1	Immigrant share 12.8% to 18.0%; science gap 37.2 to 42.7	CSV `OECD average-29`	Pass
12	Table 6	Reading and mathematics series and 2018-25 changes for 14 systems	CSV	Pass (all 56 means and 28 changes match after rounding)
13	Section 2	Back-calculated science link error 2022-25 = 3.12	CSV: sqrt(3.183² − 0.453² − 0.466²)	Pass (3.12)
14	Section 3	US 2018-25 changes "−15.6" reading, "−15.0" mathematics	CSV	**Fail**: −15.8 (SE 7.1) and −15.4 (SE 6.8); Table 6 of the same report has the correct rounded values
15	Section 3	"none of its [US] window changes is significant"	CSV	**Fail**: reading 2022-25 −14.4 (SE 7.35), z = 1.96, flagged significant in the OECD workbook and in annex Table 2
16	Section 3	2018 level vs 2018-25 change −0.46 (all), −0.23 (OECD)	CSV, recomputed	**Partial**: values are Pearson; Spearman −0.33 ($p = 0.004$) and −0.16 ($p = 0.36$); reported alongside Spearman figures without label
17	Section 3	Closure coefficient +0.20/week (SE 0.16) with log GDP; +0.14 (0.13) in causal section	`analysis_log.txt` section 8	Pass (the two figures are for different outcomes: 2018-25 and 2018-22 three-domain change)
18	Section 5	Effort regression 17 points per effort point (SE 7.3), 9 of 24.6	`analysis_log.txt` section 1	Pass; $R^2 = 0.07$ not reported
19	Section 4	Speaker counts 16, 15, 9, 8, 8, 6, 6, 6, 9, 6; "twenty-six countries"	`claims_inventory.md` Part C	Counts pass; country count **fail** (25 listed)
20	Section 7	Yearly learning gain about 10 (Serbia) to 25-28 (UK, Scotland)	vol1.txt Table I.2.7	Pass (Serbia 13.7/11.4; Scotland 28.1 science, 33.6 reading; US 13.8/12.0 not mentioned)
A1	Annex Table 1	US reading 2025 489.5 (5.8), CI 478.1-501.0, +26.5 vs OECD-35	CSV	Pass
A2	Annex Table 1	US mathematics 2025 462.8 (5.4), −2.6 vs OECD-35, n.s.; 38-member averages 461/463/482	CSV	Pass
A3	Annex Table 2	Reading from 2012: −8.1 (9.2); p10 −46.5 (9.4); p90 +29.9 (11.3)	CSV	Pass
A4	Annex 1.2	Mathematics 2003-25 −20.0 (8.4); below L2 25.7% to 35.2%; science L5+ +4.4 to 12.9%	CSV	Pass
A5	Annex 1.2	Decennial trends −6.9 (7.0, p 0.33), −10.5 (6.5, p 0.11), +3.6 (6.8)	CSV	Pass
A6	Annex 1.2	Reading 90-10 range 312, largest of 91; OECD 264; science 298 second to Luxembourg	CSV	Pass (311.8; 264.5; Luxembourg 299.0, US 297.9)
A7	Annex 1.3, Table 3	2025 response rates 45/54/76; exclusions 6.9%; coverage 0.85	CSV, Reader's Guide	Pass
A8	Annex 2.3-2.4	NAEP g8 mathematics 2019-22 −7.7, 2019-24 −8.2; LTT 13 mathematics 2012-23 −14.3, 2020-23 −9.0; g12 mathematics 2019-24 −3.3	`us_series.csv`	Pass except g12: 150.3 − 146.9 = 3.4 (trivial)
A9	Annex 3.6	NCLB years "the only period in which the bottom of the US distribution rose"	`us_series.csv` p10 series	**Fail**: g8 mathematics p10 215.0 (1990) to 229.9 (2003), +14.9, exceeds 2003-13's +7.5; LTT 13 p10 213 (1978) to 234 (1999)
A10	Annex 3.7	US immigrant students 7-18 points above non-immigrants in 2025	CSV	Pass (7.4, 17.7, 10.9)

Summary: 24 checks, 20 pass, 3 fail (14, 15, A9), 1 partial (16), plus one trivial rounding inconsistency (A8) and one miscount (19). The failures are not random in direction: two make the US decline look gentler than the tables say and one supports the accountability explanation.

Does the synthesis misrepresent the supporting sections?

Mostly not; the figures and verdicts are carried faithfully. The departures are:

1. Pandemic: "contributors of a few points" (summary) against "a third to a half of the mathematics decline since 2018" (causal review and report's own Table 8). The causal review's closing paragraph contains the same phrase, so the inconsistency is inherited, but the summary chose the weaker formulation.
2. Spending: the causal review's admission that the change-in-spending prediction was untested is dropped; the verdict "contradicted" is kept.
3. Effort: the validity review's country-level null findings (2026 multilevel study; Michaelides and Ivanova) are dropped; the evidence for the effort story is kept.
4. Curriculum: the causal review's "strongest natural experiment in the set" (England versus Scotland and Wales) is dropped.
5. Governance: the validity review's "broader critiques" section is dropped entirely.
6. United States: `trajectories.md` says "none of its window changes is significant (2022-2025 reading −13.8 with SE 7.3 is just below the threshold)"; the CSV says −14.4, z = 1.96, significant; the synthesis copied the error. The annex, working from the CSV, has it right, so the report package contradicts itself.
7. Composition: the causal review's two cautions on the native/all comparison are dropped.
8. Correlation labelling (finding 13) originates in `trajectories.md` and is copied.

Nothing in the synthesis states a source section's finding more strongly in the OECD's favour than the section did, except by omission (items 3 and 5). The direction of the omissions is consistent: hedges against the explanations the report favours (effort, digital) were dropped more often than hedges against the ones it disfavours (pandemic, spending), with the pandemic the exception in the other direction.

Space against evidence

Words in the "Candidate explanations" section, plus material on the same explanation elsewhere in the main report:

Explanation	Words in causal section	Elsewhere in report	Total	Evidence base (report's own characterisation)	Comment
Test effort	314	about 450 (validity), about 200 (conclusion)	about 960	Strongest within-PISA correlations; R^2 0.07; external literature mixed	Most space of any explanation; the qualifying statistics are the part omitted
Smartphones	342	about 100 (phone policy)	about 440	"Weakest direct causal evidence at the scale required" (causal review)	Most words in the causal section; defensible on salience; the two ban studies with the "right" sign are missing
Pandemic	316	about 250 (closure analysis)	about 570	Strongest external causal literature (meta-analysis of 42 studies); PISA-based dose-response study omitted	Adequate space, but framing (summary, verdict wording) is harsher than the section's numbers
Curriculum	249	0	249	Country cases only	Under-evidenced relative to public salience (9 speakers) and relative to `/E`, which has more
Composition	186	about 150 (section 1)	about 340	Decomposition and native series	Proportionate
Instrument/sampling	188	about 1,400 (validity section)	about 1,600	Documented	Proportionate for a report of this kind
Reading habits	177	0	177	"The check that matters most cannot be done"	Proportionate
Digitalised schooling	143	0	143	Null causal estimates	Proportionate
Flynn reversal	123	0	123	Contradicted on timing	Proportionate
Absenteeism	106	0	106	Well-identified, country-specific	Slightly thin given the US and English administrative series
Mental health	103	0	103	PISA proxies weak	Proportionate; "contradicted by PISA's well-being indicators" is strong for belonging as a proxy for depression
Spending	100	0	100	Change not tested	Least space for the second most-offered public explanation (15 speakers), and the firmest verdict
Teacher supply	71	0	71	Small effects	Proportionate
Accountability	77	0	77	Two versions cancel	Proportionate

In the annex: Common Core 971 words, pandemic 721, reading instruction 618, accountability 362, smartphones 343, absenteeism 313, demography 259, effort 257, spending about 80. The Common Core allocation is justified by the number of published evaluations; the spending allocation is not justified by the number of published evaluations (Jackson, Wigger and Xiong; Lafortune, Rothstein and Schanzenbach; Jackson and Mackevicius), all of which the annex names and none of which it applies to US timing.

Overall judgment

The report leans, and it should say so. It leans towards reading the decline as real and located in students' engagement with long, low-stakes, reading-heavy tasks, with digital displacement as the likeliest cause and declining test effort as the form in which PISA records it. That is the OECD's Annex A1 position restated with better hedging, and the report arrives at it partly because its predict-then-check method can only confirm or reject what PISA's own questionnaire measures. The report says this once, about effort, and should say it about the method as a whole.

Against the pandemic and spending the report is harder than its own evidence: the summary understates the pandemic's estimated contribution by its own section's numbers, the closure test is presented as decisive without the published study that reverses it on the same data and without the annex's finding that closure length is a poor proxy for pandemic exposure, and spending is "contradicted" on a prediction that was never tested. Against curriculum it is not so much tilted as incomplete: it examines the reforms that do not fit the timing and not the ones in the countries that fell most, and it drops the one within-country natural experiment its source section rated highest. On composition it is fair on the average and thin on the countries where shares doubled. On measurement it is careful about 2022 and 2025 sampling and silent about the two things a sceptic would go to first: Sweden's 2018 baseline and the return of scaling to a different contractor.

The three factual errors found all run one way in the US material: two make the US PISA trend look flatter than the OECD's tables say and one supports the accountability explanation with a claim the annex's own data contradict. They should be corrected before the report circulates.

None of this changes the report's central conclusion, which is that the assembled evidence rules out the confident single-cause explanations and cannot choose among the rest. The findings above would make that conclusion harder to attack from any of the seven standpoints, and would move the report from "leans with the OECD, with hedges" to "does not lean".

Changes made in response

Source of findings: `F/adversarial_review.md`. Each numbered finding below corresponds to the reviewer's numbering. "Verified" means checked against `A/pisa_trends.csv`, `A/build_log.txt`, `A/raw/pisa2018/annexA2_888934028862.xlsx`, `A/raw/pisa2018_vol1.txt`, `inputs/vol1.txt`, `B/system_features.csv`, `E/computed/analysis_log.txt`, or a direct fetch of the external source, before the change was made.

Finding 1: Sweden's 2018 baseline

Verified: PISA 2018 Table I.A2.1 gives Sweden's overall exclusion rate as 11.09%, the highest of the cycle (Israel 10.2% next); Coverage Index 3 for Sweden is 0.936 (2015), 0.857 (2018), 0.891 (2022), 0.915 (2025) in the CSV; the PISA 2018 Volume I text attributes the rise to "a large and temporary increase in immigrant and refugee inflows". The Riksrevisionen report was fetched from the audit office's site (RiR 2021:12, published 29 April 2021; the page title is "… ett tillförlitligt elevdeltagande", not "…resultat" as the reviewer cited). Its summary states that the audit found deficiencies in Skolverket's follow-up of exclusions and that the migrant-inflow explanation offered by Skolverket, the OECD and the government "is not a valid explanation" ("inte en giltig förklaring"). Sweden's 2015-to-2025 changes are −33.7 (reading) and −29.7 (mathematics) in the CSV.

Done: added a paragraph in "How much weight the results can bear" giving the exclusion rate, coverage series, the OECD's 2018 explanation and the audit office's finding, and stating that the 2018-to-2025 Swedish declines are measured from that baseline; attached the caveat where Sweden is used in the pandemic and curriculum checks, in the "three facts" paragraph, in the instrument verdict and in the "what remains open" section. The Riksrevisionen report itself (the PDF) could not be retrieved; the finding is cited from the official summary page.

Finding 2: the pandemic

(a) Done: the summary no longer says "contributors of a few points"; it now gives the causal section's own range ("a third to a half of the mathematics decline since 2018 and less of reading") and the verdict wording is "contradicted as the sole or main cause" with "the best-identified single component of the post-2018 decline" added, in the summary, the causal paragraph, Table 8 and the closing section.

(b) Verified by fetch (nature.com): Jakubowski, Gajderowicz and Patrinos, *npj Science of Learning*, 2025; PISA 2000-2022, 72 countries, UNESCO closure measures; about 0.44 mathematics points per week of full closure; 9-12% SD in shortest-closure countries, 17-23% in longest; larger losses for boys, immigrant and disadvantaged students; country-specific pre-pandemic linear trends and fixed effects. Done: cited in the pandemic paragraph with the specification difference stated and the report's own regression described as "not the last word"; also added to Table 8.

(c) Verified: `trajectories.md` states the UNESCO database does not cover Hong Kong, Macao, Chinese Taipei or Kosovo; Hong Kong's three-domain 2018-2025 change is −31.6 in the CSV. Done: noted in the pandemic paragraph and in the features paragraph. Declined: the reviewer's statement that Hong Kong "had among the longest school closures of any high-income system" is not in any source section and could not be verified without web search, which was not available at the time, so the report says only that Hong Kong is absent from the closure data.

(d) Done: carried the annex's Goldhaber et al. finding (schools in person all year lost about 0.17 SD) and its implication that closure length is a poor proxy for pandemic exposure into the main report, and restored the causal review's hedge that the absence of a gradient shows only that the marginal effect of additional weeks is invisible with a coarse measure. Also added Engzell, Frey and Verhagen (PNAS 2021), verified by fetch: 0.08 SD loss after eight weeks in the Netherlands, up to 60% larger for children of less-educated parents.

Finding 3: US numerical errors

(a) Verified against the CSV: US reading 2022→2025 −14.42 (SE 7.35), `change_to_2025_sig` = 1; 2018→2025 reading −15.83 (SE 7.08), mathematics −15.40 (SE 6.79). Done: the trajectories paragraph now says only one three-year window change (reading 2022-2025, −14.4, SE 7.3) reaches significance and only just, gives −15.8 and −15.4, and notes that the trajectory section's −13.8/−15.6/−15.0 are transcription errors.

(b) and (c) concern the US annex (`G/us_annex.md`), which is outside the file I was asked to revise. Not changed here; flagged for the annex author: the "only period in which the bottom of the US distribution rose" claim (annex 3.6) is contradicted by the annex's own `us_series.csv` per the reviewer, and grade-12 mathematics 2019-2024 is quoted as both −3.3 and −3.4.

Finding 4: the contractor change

Verified in `inputs/vol1.txt` lines 271-278 and 41158-41172: ACER is the Core 1 contractor with "instrument development, scaling and analysis, and all data products"; Open Assessment Technologies subcontractor for "the PISA 2025 platform and migration of trend units"; Oxford University Press Core 2 for the science framework; Westat Core 3 for sampling. Grep of vol1.txt for "partial invariance", "item response", "scaling model", "item parameters" and "ETS" finds no statement on whether the 2015-2022 model was retained. Done: the "science test" paragraph now names ACER, OAT and OUP, states ACER's earlier tenure and ETS's 2015-2022 tenure (from the validity review), and says Volume I does not state whether the scaling model and item parameters were carried over; the summary's fourth qualification now lists "a new scaling contractor"; the "what remains open" section asks the Technical Report to document this.

Finding 5: effort statistics

Verified: `analysis_log.txt` section 1 gives "slope=17.0 points per effort point (SE 7.3), intercept=-12.1, n=69, R2=0.07". The 2026 multilevel study ($\beta = -0.01$, $p = 0.93$ at country level) and Michaelides and Ivanova (rapid guessing positively correlated with country means) are in the validity review. Done: R^2 reported in the effort paragraph and Table 8, with the sentence that a slope explaining 7% of cross-country variance is being used to attribute more than a third of the average decline; both country-level null findings added; the closing section now says the effort correlations "explain little of the cross-country variance". Bound unchanged.

Finding 6: ESCS restatement

Verified in `A/build_log.txt`: 308 cells differ between the 2022 and 2025 workbooks, 156 for 2015, 143 for 2018, 3 for 2022 (all Palestinian Authority, whose 2022 results were recalculated); 2018 differences are mostly 1-3 points (US reading 98.78 vs 96.44; Israel reading 121.44 vs 114.32, the largest at 7.1); 2015 differences reach 25 points (Qatar, United Arab Emirates). Done: the caveat is now stated where the "advantaged fell most" finding first appears, with these sizes, and with the point that the 2022-2025 comparison is on a consistent index and the 2018-2022 widening is not. The factual summary's "up to about 4 points" (which referred to quarter means, not gaps) is superseded by the build-log figures for gaps.

Finding 7: spending

Verified: `causal_review.md` section 11 states the change-in-spending variable was not constructed. UNICEF Report Card 18 fetched: child income poverty 2012-14 to 2019-21 fell 8% across 40 countries; Poland −38%, Slovenia, Latvia and Lithuania each more than −30%; UK +20%, France, Iceland, Norway, Switzerland at least +10%. Done: restored the "not tested" statement; verdict changed to "not supported by the level data, which run the wrong way; the change prediction is untested"; UNICEF trends added with both directions; absence of a private-school-share or grade-inflation feature noted; Table 8 row revised; summary and closing section reworded. Declined: the reviewer's list of countries "combining high non-public shares with large declines" (Sweden, Netherlands, Denmark, Chile) is not in any source section and was not added. The annex's Jackson-Wigger-Xiong timing point concerns the annex and was not changed here.

Finding 8: smartphones

(a) Verified by fetch (RePEc abstract page): Beneito and Vicente-Chirivella, *Applied Economic Analysis*, 2022; two Spanish regions banning phones in 2015 (Galicia named); PISA mathematics and science; synthetic control and difference-in-differences; 0.6-0.8 years in mathematics and 0.72 to near one year in science; reduced bullying. Done: cited with the caution that a single two-region study should not carry the estimate alone; added to Table 8.

(b) Done: Abrahamsson's mental-health and bullying results (from the causal review) restored.

(c) Done: HBSC 7% to 11% (from the causal review) and Pew 2024 (from the US annex) added as external series that move as the hypothesis expects but are not linked to PISA scores. Declined: the Common Sense Census figures could not be verified (the report page fetched does not contain the numbers and the PDF was not retrieved), so they are not cited.

(d) Done: the summary and the closing section now use the same formulation as the verdict, "compatible with a large role and does not establish one"; the sentence "the evidence supports none of these at the scale the claims imply" was removed.

Finding 9: curriculum

(a) Done: the within-UK contrast restored from the causal review, with its confounds (Scotland's 2022 respondent selection; immigration, funding, poverty), and described as the strongest natural experiment in the set.

(b) Partly done: the curriculum paragraph now states as a limitation that the case-by-case check examined the reforms named in public debate and not the reform histories of Latvia, Slovenia, Iceland and Denmark, and that inclusion policies were not tested; Norway's 2020 curriculum (in the causal review) is mentioned. Declined: the specific reforms the reviewer names (Latvia's Skola2030 and language transition, Denmark's 2014 reform and 2012 inclusion law with VIVE's evaluations, Finland's 2011 three-tier support) are not in any source section and could not be verified by fetch (skola2030.lv and the Danish and Latvian Eurydice pages returned 404 or lacked the dates), so they are named as an untested gap rather than asserted.

(c) Done: the England sentence now gives the sampling reading first and the reform reading second as "two readings [that] fit equally"; the Türkiye sentence now states both readings (policy working; students practising PISA-format items).

Finding 10: composition

(a) Verified in `B/system_features.csv`: Iceland 5.59 → 10.10, Malta 8.79 → 17.14, Slovenia 8.88 → 14.07, Norway 12.40 → 18.70, United Kingdom 19.76 → 27.70, Portugal 6.96 → 18.27, Germany 22.17 → 29.22 (2018 → 2025). Done: added.

(b) Verified in claims inventory B15.3 (Denmark exemptions 11.7% to 8.3%). Done: added next to the Norway/Netherlands sentence.

(c) Verified in the CSV: OECD-29 immigrant reading 459.2 (2018) → 434.5 (2025), −24.6; non-immigrant 498.6 → 476.2, −22.3. Done: added.

(d) Declined: the Israel war-disruption context is not in any source section and could not be verified without web search; not added.

Finding 11: governance critique

Done: a paragraph restored from the validity review (Meyer and Benavot; the 2014 open letter; Sjøberg; Zhao; the OECD's technical responses and non-concession on construct and governance), with the "Learning in the Digital World" domain (verified at vol1.txt line 298) and two consequences for this report: its OECD vocabulary and the method's structural preference for PISA-measured explanations. Declined: the OECD Education 2030 point is not in any source section and was not added. The production note was rewritten in the past tense to describe this review and what it changed.

Finding 12: pre-2018 reading decline

Verified: OECD-23 reading 2012 501.4 (SE 0.6), 2018 492.8 (SE 0.5); with the 3.74 link error the trajectory section quotes from PISA 2018 Table I.A7.1, the 2012-2018 change of −8.6 has SE about 3.8, z about 2.3; the reading link error from 2012 to 2025 is 6.11. Done: the summary's first qualification now carries this. Declined: the reviewer's claim that OECD-35 reading 2015-2018 (−2.7) is not significant could not be checked because no 2015-2018 link error is in the sources; omitted.

Finding 13: Pearson/Spearman labelling

Verified by recomputation from the CSV ($n = 75$): Spearman −0.34 ($p = 0.003$), Pearson −0.47; OECD ($n = 37$): Spearman −0.16 ($p = 0.36$), Pearson −0.23 ($p = 0.18$); without Zambia and Cambodia, Spearman −0.29. Done: both statistics now labelled, with the influence of the two PISA-for-Development baselines noted and the “regression to the mean” reading marked as resting on the weaker statistic. (The reviewer’s −0.33 differs from my −0.34 by rounding of the sample used; my recomputation is reported.)

Finding 14: testing-date changes

Verified in vol1.txt lines 13952-13965 (Box I.2.2) and the Table I.2.7 rows. Done: one sentence added in the sampling discussion listing the ten systems.

Finding 15: US “year of learning”

Concerns the annex; not changed here. The main report’s citation of Table I.2.7 (10 in Serbia to 25-28 in the UK/Scotland) is unchanged; the US values (13.8 science, 12.0 reading) were verified in vol1.txt and are available for the annex author.

Finding 16: smaller items

- Country count corrected to twenty-five (the inventory lists 25 with the UK counted once). Done.
- Hungary, Chile and Hong Kong added to the list of systems where PISA and TIMSS fell together. Done.
- The causal review’s two cautions on the native/all-student comparison restored. Done.
- Adaptive-design caveat on science tails added where the 10th-percentile science fall is reported. Done.
- The NAEP inclusion-policy and US spending-series items concern the annex; not changed here.

Other changes made in the same pass

- Length: the revised prose is about 9,500 words (excluding tables, captions and headings), against the coordinator’s “about 9,000”; the additions required by the review were offset by condensing elsewhere without removing any explanation, verdict or load-bearing number.
- Table 8 rows for pandemic, smartphones, curriculum, effort and spending updated to match the revised verdicts.
- The “what remains open” section now includes the scaling-model carry-over question and Sweden’s 2018 baseline.
- The report’s refusal to select an explanation, and its statement that the surviving explanations are not independent and that the effort explanation’s size is a definitional choice, are unchanged.

Items declined or not verifiable, summarised

Not added because not in any source section and not verifiable in this session: Hong Kong’s closure length; Israel’s 2023-25 disruption; Latvia’s, Denmark’s and Finland’s specific reforms and VIVE’s evaluations; the Common Sense Census figures; the private-school-share country list; OECD Education 2030; the OECD-35 2015-2018 reading significance. Items 3(b), 3(c), 7 (annex timing), 15 and the annex parts of 16 belong to the US annex and are left for its author.

Appendix H. Quotation check before publication

Before the report was published, a further set of AI agents checked the quotations attributed to named people, organizations and documents in the main report, the United States annex and the inventory of public claims (Appendix F) against their sources. Each agent fetched the cited page and compared the quoted words with the source text, in the original language where the source was not in English.

About 300 quoted passages were checked. About 200 matched their sources, or were fair paraphrases marked as such. About 40 did not. Most of these were sentences cut short without an ellipsis, or translations and reported speech presented as direct quotation; nine could not be found in their sources in anything like the quoted words. All of these were corrected, turned into paraphrase, or removed, as listed below. About 50 passages could not be rechecked, most of them from OECD publications whose web pages cannot be read by automated tools; these stand as the agents first took them from the text of *PISA 2025 Results (Volume I)* and the other documents cited, and three OECD passages that a checker compared with a different page of Volume I were left unchanged for that reason.

Words not found in the source

Where	What was quoted	What was done
F, B10.1	A sentence attributed to Ludger Wößmann comparing today's 15-year-olds with 13-year-olds of 13 years ago	Removed; not in the cited article
F, B10.3	A "market school experiment" statement attributed to Vänsterpartiet	Removed; no PISA 2025 statement found in the cited outlets. The count of speakers offering inequality and funding explanations falls from 15 to 14 (Part C and the main report)
F, B12.1	Greg Ashman: England's approach "broadly held"	Replaced with Ashman's sentence, in which "broadly" describes the Scottish and Welsh reforms
F, B12.2	Department for Education: "decades of reform to raise academic standards"	Replaced with a sentence from the same statement
F, B12.3	Màiri McAllan: performance stable "while many other countries saw decline"	Replaced with her actual words
F, B12.6	Spanish PP regional governments: LOMLOE a "categorical failure"	No shared phrase exists; replaced with a description and one regional government's words
F, B12.10	Jason Clare: "classroom teaching reforms," "20-year reform trajectory"	Replaced with his actual words
F, B5.2	A Korean ministry official on pandemic learning loss	Quotation not in the source; replaced with a paraphrase of the article
US annex, 2.3	Tom Loveless: "six years of flat NAEP scores extending across all states"	Replaced with a description of his finding and a phrase that is in the report

Wording, attribution or sense corrected

Where	Problem	Correction
F, B15.5	Schools Week's description of England's non-respondents was reversed (more likely, not less likely, to be eligible for free school meals and to have special educational needs)	Quoted in full
Main report, reliability section; F, B15.3	A Danish researcher's view presented as an English quotation	Paraphrased
Main report, claims section; F, A2.8	Tue Halgreen's caution on causality presented as a direct quotation	Paraphrased, with the Spanish reported speech
US annex, 1.3	Reason for the 2003 autumn testing window	Replaced with the NCES wording
US annex, 2.4	Chalkbeat's headline framing presented as a quotation	Paraphrased
US annex, 3.1	Education Recovery Scorecard estimate on absence	Quoted as published
F, A2.2	Schleicher's four drivers: absence is presented in the source as a possible reflection of changed attitudes, not a separate driver	Quoted as published
F, B1.11, B3.1, B3.2, B3.3	Quotations from David Edwards, Jonathan Seiden and Andrea Bradley cut short or run together	Completed
F, B2.1, B2.2, B7.5, B8.1	Norwegian and Finnish quotations truncated, in the wrong grammatical form, or mistranslated ("frustration tolerance" for tolerating boredom)	Corrected in the original language and translation
F, B4.4	A reporter's sentence attributed to Arto Ahonen	Attributed to Yle; Ahonen's own words quoted
F, B10.3	Magdalena Andersson's words	Corrected

Where	Problem	Correction
F, B10.9, B11.4, B12.5	French union and senator statements incomplete, or a translation in quotation marks	Corrected
F, B12.14	Timeframe of the Turkish exam’s alignment with PISA	Corrected to “the last two or three years”
F, B14.1, B14.5, B15.14	Reported speech in quotation marks; a wrong cross-reference; Icelandic wording	Corrected

A correction of reasoning

Where	Problem	Correction
Main report, curriculum paragraph; E, section 5 (England)	The sampling reading of England’s results was stated backwards. If England’s 2022 score was inflated by 7 to 8 points, the unchanged 2022-to-2025 figures mean a real rise, not a return to a properly sampled level; and the coverage-index swing, if real, would have pushed the 2022 mean down, not up	Restated: a larger fall from 2018 to 2022 followed by a recovery from 2022 to 2025, with the 2018-to-2025 changes unaffected by the 2022 bias and 2022 marked as the anomalous sample

Clarifications of wording

Where	Problem	Change
Main report, summary (third qualification) and the section on how much weight the results can bear; D, the same sentence	“Measurement rather than learning” and “a reader who does not” compressed the point about test effort past clarity	Spelled out: the OECD counts the capacity to sustain attention as part of reading, so a fall in it is a fall in reading; a reader who separates reading skill from effort on a test that doesn’t count will see part of the decline as students trying less rather than reading less well
Main report, “How this report was produced”; About box on the web page	Said the agents were given their brief and materials “and nothing else” (About box: “no context whatsoever”). Agents running in the same software on 10 September reported that it supplied the account holder’s name, email address and a short description of his newsletter project, and the original agents ran in the same setting	Now says so
Main report, opening of “What the 2025 results show”	Said the link error is “explained in the next section”, though the list of terms had already defined it	Now refers back to the definition
Main report, summary (causes paragraph)	“Composition” used without explanation	Replaced with “changes in the student population, chiefly a larger share of immigrant students”

Section summaries

The short summaries headed “In brief” at the top of the longer sections of the main report, the United States annex and Appendices A to F were added on 10 September 2026, after the sections were written. Four agents running Claude Fable 5, the model that wrote the sections, drafted them from the section text alone, with a brief that did not say who had commissioned the work; the software they ran in did supply basic account details (a name, an email address and a short description of a newsletter project), which each agent reported it had not used. Two agents running Claude Opus then compared each draft with its section and found 50 problems: qualifications dropped, numbers attached to the wrong period or group, claims stated more strongly than the section states them, and unexplained shorthand. A third check made 10 further accuracy corrections and about 60 edits for readability. The drafts and all three check logs are kept with the report’s working files.

The final check also found four places where sections disagree with one another, which the summaries do not resolve. Appendix C describes the systems that rose since 2018 as starting below 460 points with three exceptions in one place and with one exception in two others. The 25-point reading decline behind the OECD’s estimate of the effect of immigration is labelled the OECD-29 average in Appendices A, C and E and the OECD-35 average in Appendix D. The main report says restated socio-economic gaps differ by more than 20 points for two systems in 2015, while Appendix A says by up to about 4 points. Appendix C reports correlations between starting level and change of −0.46 and −0.23 without saying which kind; the main report gives Spearman values of −0.34 and −0.16 and identifies the other figures as Pearson correlations.

Not rechecked

Quotations from The Conversation France (Claude Diebolt; B3.6, B12.5, B15.9) and from Singapore’s Ministry of Education (B12.16) could not be retrieved and are marked in the inventory. Quotations from OECD publications (Volume I, its country notes and Reader’s Guide, and earlier PISA volumes), and from several NCES reports, could not be read by the checking tools; where a search engine indexed the exact phrase on the OECD’s own pages, as for “complementary rather than competing” and “never the main driving force”, that is the only independent confirmation.