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

The results, how much weight they can bear, and which explanations survive the evidence. With a United States annex.

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 $13,952(2012-13) \mathit{inconstant} 2022-23 \mathit{dollars}, a 5.214,894$) 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).