Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Measurement-bounded predictive inference in international large-scale assessment: how the plausible-value ceiling governs attainable accuracy and the stability of predictor rankings in PISA 2022

Jonas A. Mandalunes

jonas.mandalunes@perpetual.edu.ph

AMA Computer University

Research Article

Keywords: plausible values, measurement precision, reliability, international large-scale assessment, PISA 2022, cross-validation, predictor importance, creative thinking

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10732065/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background. Research on international large-scale assessments increasingly uses flexible predictive models to report how accurately student achievement can be predicted and which characteristics matter most. Achievement in these assessments is not observed but represented by plausible values drawn from a posterior distribution, so part of the analysed variance is measurement noise that no predictor can explain. This study asks how large the resulting bound on attainable predictive accuracy is, whether it differs across domains and education systems, and whether it also governs the stability of the predictor rankings that applied studies report.

Methods. Using the PISA 2022 international database, covering 80 education systems for mathematics, reading and science and 63 for creative thinking, we estimated a measurement-precision ceiling for each domain in each system as the intraclass correlation of the ten released plausible values, applying a finite-draw correction and a school-cluster bootstrap computed jointly across systems to obtain rank intervals. A weighted linear baseline and a gradient boosting model were fitted on identical school-disjoint cross-validation folds, refitted once per plausible value, and predictor rankings were compared across those refits. A paired design compared domains within system, holding students, schools, predictors and folds fixed so that only the measured construct varied.

Results. The ceiling ranged from 0.654 to 0.933 across domains and systems, and correlated only weakly across domains within the same students. Attainable accuracy and ranking stability both scaled with the ceiling in a graded manner across six paired domain comparisons, with the relationship approaching zero as the ceiling difference approached zero. In the least precisely measured domain, over 40 per cent of a five-predictor list changed between two plausible values of the same data. Random student-level splits inflated performance in 604 of 606 analyses, by an amount governed by the number of sampled schools.

Conclusions. Predictive performance from these databases is not comparable across domains or systems without reference to the measurement bound, and predictor rankings from a single plausible value are one realisation of a wide distribution. Both problems can be addressed with quantities computable from the released data.

Introduction

International large-scale assessments are now a standard source for quantitative research on education, and an increasing share of that research uses flexible predictive models to identify which student, family and school characteristics are associated with achievement. The usual reporting format is a measure of predictive accuracy together with a ranked list of important predictors, from which substantive and sometimes policy conclusions are drawn. Three properties of these assessments make such conclusions harder to interpret than the reporting format suggests, and all three are routinely left unaddressed.

The first is that achievement is not observed. It is represented by a set of plausible values drawn from a posterior distribution, so a portion of the variance in the analysed outcome is measurement noise that no predictor can explain. The second is that students are sampled in school clusters, so a random student-level split of the data allows a model to learn school-level structure that then reappears in the evaluation sample. The third is that reported accuracy and predictor rankings are compared across education systems as though the quantity of learnable signal were the same everywhere.

The principle underlying the first of these is long established and we claim no part of it. That unreliability in a measure attenuates its observed associations dates to Spearman (1904). That the reliability of a prediction target places a ceiling on achievable out-of-sample accuracy has been demonstrated directly, with simulation and with four large empirical datasets, for brain–behaviour prediction (Gell et al., 2024), and achievable performance ceilings have been derived from classical test theory within educational measurement itself for automated essay scoring (Uto, 2026); comparable noise-ceiling methods are established in computational modelling (Lage-Castellanos et al., 2019). Nor is the decomposition of plausible-value variance new. The technical documentation of the National Assessment of Educational Progress has for many years partitioned the estimation-error variance of a group statistic into sampling and measurement components, observed that the resulting ratio "roughly corresponds to reliability in classical test theory", and noted that the measurement share varies across subject areas and is largest where there is least information about a student's scale score (National Center for Education Statistics, 2008). Analytical choices in the scaling model for international assessments have likewise been examined directly (Robitzsch, 2022; Robitzsch & Lüdtke, 2022), and reliability heterogeneity has been studied through multiverse methods in adjacent fields (Parsons, 2022).

What has not been established, to our knowledge, is what follows from these facts for the predictive research now being conducted on these databases. The existing demonstrations concern a single population, a single domain, or the error variance of a group statistic that the programme in question explicitly declines to interpret at the individual level. None estimates individual-level scale precision separately for every participating education system; none asks whether that precision governs the stability of the predictor rankings that such studies report; and none establishes whether any such relationship is graded, which is the property that would distinguish a substantive mechanism from a coincidence.

This paper addresses those questions using the PISA 2022 international database, covering 80 education systems for mathematics, reading and science and 63 for creative thinking. We make four contributions. First, we estimate a measurement-precision ceiling on attainable predictive performance for every domain in every participating system, using the released plausible values as a direct estimator with a finite-draw correction, and we show empirically that this quantity is distinct from the published group-statistic proportion with which it might be confused. Second, we show that the ceiling varies enough between systems and domains, and is sufficiently weakly related across domains within the same students, that raw predictive performance is not comparable without reference to it. Third, using a paired design in which the same students, schools, predictors and folds are used for each domain in turn so that only the measured construct changes, we show that both attainable accuracy and the stability of predictor rankings scale with the ceiling in a graded manner, and that the relationship approaches zero where the ceiling gap approaches zero. Fourth, we quantify what conventional random splitting costs in this setting

and identify the parameter that governs it. We close with an exploratory analysis of why precision varies, which we report with its limitations rather than as an established result.

Methods

Data

The analysis uses the public international database of the OECD Programme for International Student Assessment 2022 (cycle code CY08MSP). Four files were retrieved: the student questionnaire file, containing 1278 variables for 613744 students in 80 education systems; the school questionnaire file, containing 431 variables for 21629 schools; the creative thinking file, containing 301 variables for 499843 students in the 63 systems that administered the creative thinking assessment; and the official codebook. Files were downloaded from the OECD file server in August 2026. Byte sizes and SHA-256 hashes of every source file are reported in Additional file 1 so that the exact release used here can be identified unambiguously.

Two features of the release require comment because they are easy to miss and consequential. First, the creative thinking plausible values are not contained in the student questionnaire file; they are distributed in a separate creative thinking file under the variable names PV1CRTH_NC to PV10CRTH_NC, reported on a number-correct metric. Second, that file contains no sampling weights of any kind, so the final student weight must be merged from the student questionnaire file on the international student identifier. The creative thinking file additionally carries its own re-scaled plausible values for mathematics, reading and science, drawn from a population model that includes creative thinking as an additional dimension. Analyses that combine creative thinking with a core domain therefore face a choice between two scaling models, and we report sensitivity to that choice.

Weighted domain means computed from the acquired files reproduce the figures published by the OECD to within rounding, which confirms that the files are the international database and not a derivative. All analyses use the final student weight. The 80 balanced repeated replication weights are present and complete, with a Fay factor of 0.5; the design effect they imply for a country mean is of the expected magnitude, and a worked check is reported in Additional file 1. School membership is identified by the international school identifier and is used throughout as the clustering unit.

The measurement-precision ceiling

Achievement in PISA is not observed. It is represented by ten plausible values drawn from a posterior distribution conditioned on a student's item responses and on background information (Sinharay et al., n.d.; von Davier et al., 2009). Writing the mth draw for student i as

$\theta_i^m = \mu_i + \varepsilon_i^m$, $\varepsilon_i^m \sim N(0, \sigma_i^2)$, m = 1, …, M, with M = 10,

where μ_i is the posterior mean and σ_i² the posterior variance, the ε component is by construction independent of any predictor set. It follows that for any model predicting a single plausible value,

$R^2 \le \rho = \mathrm{Var}(\mu) / [\mathrm{Var}(\mu) + E(\sigma^2)]$.

We call ρ the measurement-precision ceiling. It is the intraclass correlation of the ten draws treated as repeated measures within student, and it is estimable from the released data alone. Writing $\bar{y}_i$ for the mean of a student's ten draws and s_i² for their unbiased variance, and using the final student weight throughout,

$MSW = \Sigma\, w_i\, s_i^2 / \Sigma\, w_i$,

$\mathrm{Var}(\mu) = \mathrm{Var}_w(\bar{y}) - MSW / M$,

$\rho = \mathrm{Var}(\mu) / [\mathrm{Var}(\mu) + MSW]$.

The term − MSW/M is a finite-draw correction. The variance of the student means is itself inflated by residual measurement noise that shrinks only as 1/M rather than to zero, and omitting the correction biases ρ upward. A simulation with a known ρ recovers 0.697, 0.850 and 0.950 at true values of 0.70, 0.85 and 0.95 respectively (Additional file 2).

An estimator using the variance of a single plausible value in place of $\mathrm{Var}_w(\bar{y})$ is consistent but discards nine of the ten draws when estimating the total variance, and is correspondingly noisier. We report ρ from the ten-draw estimator throughout and give the single-draw values for comparison.

Relation to an existing published statistic

The technical documentation of the National Assessment of Educational Progress partitions the estimation-error variance of a group statistic into a sampling component U and an imputation component B, and observes that U/V "roughly corresponds to reliability in classical test theory" (National Center for Education Statistics, 2008). Because that quantity concerns the error variance of an estimator rather than the variance of the outcome across students, it is not the same as ρ. We test this directly rather than assert it: computing both on the same PISA 2022 data for all 80 systems, with U obtained from the 80 replication weights, the two correlate at r = 0.195 (Spearman 0.327). The reason is structural: U carries the school-clustering design effect and the effective sample size, whereas B does not, so U*/V tracks survey design rather than individual-level precision. Reporting a distinct individual-level statistic is therefore warranted.

Uncertainty in ρ and in the ordering of systems

Standard errors and intervals for ρ were obtained by a school-cluster bootstrap that resamples schools with replacement within system, respecting the two-stage sample design, with B = 400 replications. Because the substantive interest includes where a system stands relative to others, the bootstrap was computed jointly across systems: within each replication every system is resampled and ρ recomputed, so that ranks can be recorded per replication. This yields valid rank intervals rather than intervals on ρ alone. We report the 2.5th and 97.5th percentiles of ρ, the corresponding rank interval, and the bootstrap probability that a system occupies the highest rank or the top five.

Predictor set

Predictors are the derived context indices and background variables of the PISA 2022 questionnaires, comprising creativity-related context, socioeconomic status and home resources, school climate and wellbeing, motivational dispositions, time use, demographics and test-taking effort, together with school-level characteristics from the school questionnaire. Every variable name was verified against the file metadata rather than taken from documentation, and every construct was checked for availability within each system. Variables entirely missing in a system were dropped for that system, which leaves between 38 and 46 predictors depending on system. The full inventory, with per-system missingness, is given in Additional file 3.

Achievement plausible values are excluded from the predictor set. They are alternative measurements of the same latent construct family as the outcome and their inclusion would constitute target leakage.

Validation architecture

Students in PISA are sampled in school clusters. A random student-level split therefore permits a model to learn school-level structure that reappears in the evaluation fold. All primary estimates use school-disjoint five-fold cross-validation, in which entire schools are assigned to folds so that no school contributes students to both training and evaluation. Every preprocessing step, including imputation and scaling, is fitted inside the training folds only, within a pipeline, so that no information from an evaluation fold can influence it.

To quantify what the conventional alternative costs, the identical pipeline was also evaluated under a random student-level split with the same number of folds, and the difference in out-of-fold R^2 is reported as an inflation.

Models

Two model classes were fitted, always in parallel and always on identical folds. The statistical baseline is a weighted ridge regression with median imputation and a missingness indicator. The machine-learning comparator is a weighted histogram-based gradient boosting regressor. Hyperparameters were held fixed across all systems and domains, since the quantity of interest is the relationship between attainable performance and measurement precision rather than maximal accuracy, and holding the pipeline fixed is what makes the paired comparisons below interpretable. Performance is the weighted out-of-fold coefficient of determination, computed against the weighted mean of the outcome. The random seed was 20220 throughout.

Explanation stability as an estimand

For each system, domain and model class, the entire pipeline was refitted once per plausible value, with folds and hyperparameters held fixed. Successive replications therefore differ only in latent measurement noise. Predictor importance was obtained by weighted permutation importance on a held-out fold, and the stability of the resulting rankings across the ten replications was summarised by the mean pairwise Kendall rank correlation and the mean pairwise Jaccard overlap of the top five predictors. We also report the range of out-of-fold R^2 across the ten plausible values.

A null for attributing instability

Permutation importance distributes credit arbitrarily among correlated predictors, and several of the indices used here are substantially correlated. Rank churn could therefore arise from the importance estimator rather than from measurement noise. To separate the two, three conditions were compared with folds and hyperparameters fixed: the outcome varied across the ten plausible values with the estimator seed fixed; the outcome fixed at a single plausible value with the estimator seed varied; and both fixed. The second and third conditions contain estimator noise alone. Instability is attributed to measurement noise only where overlap is lower in the first condition than in the second.

The paired within-system design

Correlating ρ against performance across systems confounds measurement precision with every other respect in which education systems differ. The primary analysis therefore pairs within system. For each system, the same students, the same schools, the same predictors, the same folds and the same hyperparameters were used to model each available domain in turn. Only the measured construct, and therefore ρ, changes. Differences are taken within system and tested with the Wilcoxon signed-rank test; a sign-flip permutation test was used as a distribution-free check. With four domains there are six orderable domain pairs, which span a range of paired ceiling gaps and permit examination of whether downstream differences scale with the gap.

Analysis of coding intensity

Creative thinking is the only PISA 2022 domain composed entirely of open-ended tasks scored by human coders against rubrics; mathematics is essentially machine-scored, and reading and science lie between. To examine whether measurement precision varies with the volume of responses a system had to code, ρ and the measurement error were regressed on the natural logarithm of the number of students assessed in creative thinking, separately by domain. Because measurement error is expressed on different score metrics across domains, raw slopes are not comparable between domains; all cross-domain comparisons therefore use scale-free quantities, namely the noise-to-signal ratio and the elasticity of log measurement error. The paired within-system difference in log measurement error between two domains provides a test in which any sample-size artefact common to both domains cancels. This analysis is exploratory and was not specified in advance; its limitations are set out in the Limitations section.

Software and reproducibility

Analyses were carried out in Python 3.14 with numpy 2.4.4, pandas 3.0.2, scikit-learn 1.9.0, scipy 1.17.1, statsmodels 0.14.6 and pyreadstat 1.3.6. The random seed was 20220 for every stochastic procedure. Source data files were treated as read-only and never modified. All analysis code, the derived data tables, and a machine-readable file of the exact values plotted in each figure are archived and citable; see Availability of data and materials. Every headline quantity reported in this paper was independently recomputed by a second, different implementation as a check on the first, and the comparison is included in the archive.

Results

The measurement-precision ceiling varies by domain and by education system

Table 1 reports the ceiling for each domain across all participating systems. The ceiling differs substantially both between domains and between systems within a domain. Creative thinking spans 0.654 to 0.894 across the 63 systems that administered it, a range of 0.240. Mathematics spans 0.811 to 0.933 across 80 systems, reading 0.766 to 0.915, and science 0.790 to 0.915. Domain medians order as creative thinking 0.802, reading 0.857, science 0.871 and mathematics 0.900. Figure 1 shows the full distributions.

Table 1
Measurement-precision ceiling by assessment domain

Domain	Systems	Students	Minimum	25th percentile	Median	75th percentile	Maximum	Range	Mean CI width	Mean PV-averaging inflation
Creative thinking	63	499843	0.654	0.778	0.802	0.822	0.894	0.24	0.023	0.174
Reading	80	613744	0.766	0.843	0.857	0.873	0.915	0.149	0.018	0.126
Science	80	613744	0.79	0.847	0.871	0.889	0.916	0.125	0.018	0.119
Mathematics	80	613744	0.811	0.871	0.9	0.917	0.933	0.122	0.017	0.097

Ceiling estimated as the intraclass correlation of the ten plausible values with a finite-draw correction, weighted by the final student weight.

Table 1
b Design-based standard errors for the measurement-precision ceiling

Domain	Systems	Mean ceiling	Mean BRR standard error	Minimum SE	Maximum SE	Mean 95 per cent CI width
Creative thinking	63	0.8014	0.0047	0.0019	0.0071	0.0185
Reading	80	0.8571	0.0037	0.0011	0.0085	0.0144
Science	80	0.8654	0.0035	0.0012	0.0082	0.0139
Mathematics	80	0.8909	0.0033	0.0008	0.011	0.0129

Standard errors by balanced repeated replication using the 80 replicate weights supplied with the database and a Fay factor of 0.5, recomputing the full estimator once per replicate weight.

The practical reading of these values is direct. In a system at the lower end of the creative thinking distribution, no model of any kind can explain more than about 65 per cent of the variance in a single plausible value, however many predictors are added and whatever algorithm is used. In a system at the upper end of the mathematics distribution the corresponding figure is about 93 per cent. Reported model performance is therefore not comparable across domains or across systems without reference to this bound.

The finite-draw correction is small in magnitude but systematic: its mean effect across all domains and systems is + 0.0001, with a maximum absolute shift of 0.0126 relative to the single-draw estimator.

Two further features of Table 1 are worth noting. First, averaging the ten plausible values, a shortcut that the developers of the method explicitly discourage (von Davier et al., 2009), raises the ceiling by a domain- and system-dependent amount: a mean of 0.174 for creative thinking, 0.126 for reading, 0.119 for science and 0.097 for mathematics, reaching 0.296 in the most affected system. An analyst who averages plausible values therefore obtains a higher apparent bound than one who does not, by an amount that is confounded with both domain and country. Second, the bootstrap intervals on ρ are narrow, with a mean width of 0.017 to 0.023 depending on domain, so the between-system differences reported here are not attributable to estimation error in ρ.

Measurement precision is close to orthogonal across domains

Within the 63 systems assessed in both, the creative thinking ceiling and the mathematics ceiling correlate at r = 0.148 (Fig. 2). Measurement quality in one domain is therefore a poor guide to measurement quality in another within the same system and the same students. The correlation with the domain mean score is also weak and negative (r = − 0.295), so higher-scoring systems do not enjoy more precise measurement of creative thinking; if anything the reverse.

Only two of the 63 systems measure creative thinking more precisely than they measure mathematics; in the remaining 61 the order is reversed. One consequence is that a system's standing on measurement precision can differ sharply by domain. One system ranks 1st of 63 on the creative thinking ceiling, with a 95 per cent bootstrap rank interval of 1 to 2 and a bootstrap probability of 0.733 of occupying the highest rank, while ranking 55th of 80 on mathematics. Figure 6 gives the ceiling with rank intervals for the highest and lowest systems.

Attainable performance tracks the ceiling

Across systems within a domain, out-of-fold R^2 is positively associated with ρ in every domain and both model classes: for gradient boosting, r = 0.640 for creative thinking, 0.695 for reading, 0.775 for science and 0.711 for mathematics; for the linear baseline the corresponding values are 0.267, 0.532, 0.599 and 0.516. Stability of importance rankings is also positively associated with ρ across systems, but more weakly and inconsistently between model classes. Because between-system comparisons confound measurement precision with every other difference between education systems, we do not rely on them; they are reported in Additional file 4 for completeness. The primary analysis is the paired design that follows.

The paired within-system analysis: a graded relationship

Table 2 reports, for each of the six orderable domain pairs, the paired within-system difference in the ceiling and in four downstream quantities. Every comparison holds the students, schools, predictors, folds and hyperparameters fixed, so that only the measured construct differs.

Table 2
Paired within-system differences across six domain pairs

Model	Domain pair	Systems	Delta rho	Delta R2	p (R2)	Positive (R2)	Delta Kendall tau	p (Kendall tau)	Positive (Kendall tau)	Delta Top-five Jaccard	p (Top-five Jaccard)	Positive (Top-five Jaccard)	D R o a P
Gradient boosting	Mathematics minus Creative thinking	63	0.0882	0.0904	6e-11	57 of 63	0.1157	1.4e-11	61 of 63	0.1221	5.2e-08	55 of 63	-(
Gradient boosting	Science minus Creative thinking	63	0.0619	0.0664	2.8e-11	57 of 63	0.0882	1.3e-10	57 of 63	0.0638	0.00042	43 of 63	-(
Gradient boosting	Reading minus Creative thinking	63	0.0524	0.0811	8e-12	62 of 63	0.1002	1.4e-11	61 of 63	0.079	9.8e-06	47 of 63	-(
Gradient boosting	Mathematics minus Reading	80	0.0339	0.0097	0.045	51 of 80	0.0148	0.005	53 of 80	0.0387	0.025	50 of 80	-(
Gradient boosting	Mathematics minus Science	80	0.0256	0.0227	4.3e-08	62 of 80	0.0255	2.7e-07	59 of 80	0.0515	0.00028	52 of 80	-(
Gradient boosting	Science minus Reading	80	0.0083	-0.013	0.006	32 of 80	-0.0107	0.16	36 of 80	-0.0128	0.24	34 of 80	0
Linear baseline	Mathematics minus Creative thinking	63	0.0882	0.0889	2.5e-10	56 of 63	0.1278	1.5e-11	59 of 63	0.1763	2.3e-11	56 of 63	-(
Linear baseline	Science minus Creative thinking	63	0.0619	0.0637	9e-11	58 of 63	0.0793	1.5e-10	58 of 63	0.1183	4.3e-08	50 of 63	-(
Linear baseline	Reading minus Creative thinking	63	0.0524	0.0782	5.4e-12	62 of 63	0.092	2.1e-11	58 of 63	0.1114	7.1e-08	51 of 63	-(
Linear baseline	Mathematics minus Reading	80	0.0339	0.0114	0.023	52 of 80	0.0344	1.5e-07	61 of 80	0.0519	0.001	50 of 80	-(
Linear baseline	Mathematics minus Science	80	0.0256	0.0244	1.2e-08	65 of 80	0.0461	3.2e-12	71 of 80	0.0563	0.00018	51 of 80	-(
Linear baseline	Science minus Reading	80	0.0083	-0.013	0.0089	34 of 80	-0.0118	0.028	35 of 80	-0.0044	0.91	42 of 80	0

Each row compares two domains within the same systems, holding students, schools, predictors, folds and hyperparameters fixed. p-values from the Wilcox

For the pair with the largest ceiling gap, mathematics against creative thinking ($\Delta\rho = 0.088$, $n = 63$), the linear baseline gives a paired difference of + 0.089 in out-of-fold R^2 (Wilcoxon $p = 2.5 \times 10^{-10}$), + 0.128 in Kendall's τ between importance rankings ($p = 1.5 \times 10^{-11}$), + 0.176 in top-five Jaccard overlap ($p = 2.3 \times 10^{-11}$), and − 0.022 in the range of R^2 across plausible values ($p = 2.7 \times 10^{-11}$). Gradient boosting gives + 0.090, + 0.116, + 0.122 and − 0.019 respectively, with comparable p-values.

The pattern is graded. Ordering the six pairs by the size of the ceiling gap, every downstream difference declines with it, and statistical significance degrades in step (Fig. 3). At $\Delta\rho = 0.088$ all four quantities are significant at $p < 10^{-7}$. At $\Delta\rho = 0.052$ and 0.062 they remain significant at $p < 10^{-5}$. At $\Delta\rho = 0.026$ and 0.034 the differences are smaller and the p-values weaken to between 10^{-7} and 0.02. At the smallest gap, science against reading, where $\Delta\rho = 0.008$, the differences are near zero or slightly negative and are not significant for Jaccard overlap ($p = 0.91$ for the linear baseline, $p = 0.24$ for gradient boosting).

Across the six pairs, the correlation between the ceiling gap and the downstream gap is 0.988 for top-five Jaccard overlap, 0.959 for Kendall's τ, 0.927 for out-of-fold R^2, and − 0.965 for the range of R^2 across plausible values, using the linear baseline; for gradient boosting the corresponding values are 0.936, 0.935, 0.923 and − 0.957. Fitting a straight line to the Jaccard relationship gives an intercept of − 0.014 for the linear baseline and − 0.009 for gradient boosting. The intercept was not constrained. That the fitted line passes close to the origin, and that the pair with almost no ceiling gap shows almost no downstream gap, are the two features of this analysis least easily explained by confounding.

Importance rankings are unstable, and the instability is attributable to measurement noise

In absolute terms the instability is substantial. Median top-five Jaccard overlap between importance rankings that differ only in which plausible value was used as the outcome is 0.543 for creative thinking under the linear baseline and 0.612 under gradient boosting, against 0.729 for mathematics under both. An overlap of 0.543 means that a little over 40 per cent of a five-predictor list changes between two plausible values of the same students, the same predictors and the same folds.

Table 3 reports the attribution null. In all six system by model cells examined, top-five overlap is lower when the outcome varies across plausible values than when the outcome is held fixed and only the estimator seed varies, by between 0.105 and 0.453. The instability is therefore not an artefact of the importance estimator (Fig. 5).

Table 3
Attribution of predictor-ranking instability to measurement noise

System	Model	Predictors	Kendall tau, varying PV	Kendall tau, fixed outcome varying seed	Kendall tau, both fixed	Jaccard, varying PV	Jaccard, fixed outcome varying seed	Jaccard, both fixed	Excess attributable to measurement noise
PHL	Linear baseline	46	0.768	0.816	0.822	0.684	0.844	0.882	0.161
PHL	Gradient boosting	46	0.661	0.795	0.8	0.476	0.651	0.785	0.174
KOR	Linear baseline	43	0.657	0.787	0.803	0.642	0.8	0.819	0.158
KOR	Gradient boosting	43	0.511	0.701	0.683	0.405	0.551	0.542	0.147
HKG	Linear baseline	43	0.528	0.837	0.845	0.332	0.785	0.785	0.453
HKG	Gradient boosting	43	0.543	0.75	0.758	0.484	0.589	0.526	0.105

Mean pairwise agreement between predictor rankings across ten replications under three conditions, folds and hyperparameters fixed throughout.

Table 4
Inflation of predictive performance under random student-level splitting

Model	Analyses	Positive inflation	Spearman with sampled schools	Maximum inflation	Median inflation, school quartile 1	Median inflation, school quartile 2	Median inflation, school quartile 3	Median inflation, school quartile 4
Linear baseline	303	302 of 303	-0.442	0.746	0.0246	0.0195	0.0151	0.011
Gradient boosting	303	302 of 303	-0.365	0.148	0.0398	0.0303	0.0295	0.0216

Inflation is the difference in out-of-fold R-squared between a random student-level split and a school-disjoint split, identical pipeline and data.

The null condition is nevertheless not innocent. With the outcome held completely fixed, top-five overlap is still only 0.526 to 0.882. Permutation importance over correlated context indices is itself unstable, and the two sources of instability compound. A study that reports a single ranking from a single plausible value with a single seed is reporting one draw from a distribution it has not characterised.

Random splits inflate performance, and the inflation depends on the number of sampled schools

Random student-level splits produced higher out-of-fold R^2 than school-disjoint splits in 604 of 606 system by model cells across the four domains. The magnitude is governed by how many schools a system sampled rather than by the learner: Spearman correlation between the number of sampled schools and the inflation is − 0.442 for the linear baseline and − 0.365 for gradient boosting (Fig. 4). The largest single inflation observed is 0.746. In the most extreme system, which sampled 46 schools, the school-disjoint estimate of R^2 is 0.002 while the random-split estimate of the same pipeline on the same data is 0.361.

That clustered data require grouped cross-validation is not itself a new observation. What these figures add is the magnitude in international large-scale assessment and the identification of the sampled-school count as the governing parameter.

Normalising by the ceiling reorders systems

Expressing performance as a share of what is attainable, R^2/ρ, changes the ordering of systems substantially. Between 53 and 70 of the 63 to 80 systems change rank, depending on domain and model class, with a maximum movement of 20 rank positions. The movement is systematic rather than noise: the correlation between ρ and rank gain is negative in all eight domain by model combinations, ranging from − 0.515 to − 0.724, meaning that systems with lower

ceilings systematically gain position once performance is expressed relative to the bound. Reordering is largest in creative thinking, which has the widest ceiling spread, and smallest in mathematics, which has the narrowest.

Exploratory analysis: precision, coding intensity and the number of students assessed

Measurement precision declines with the number of students assessed, and the size of the decline is graded by how much human open-response coding the domain requires. Regressing the noise-to-signal ratio on the logarithm of the number of students assessed gives a slope of 0.0850 for creative thinking, which is composed entirely of open-ended human-coded tasks ($p = 2.9 \times 10^{-6}$, $R^2 = 0.303$); 0.0401 for science, which contains some open-response items ($p = 0.005$); 0.0229 for reading ($p = 0.072$); and 0.0164 for mathematics, which is essentially machine-scored ($p = 0.327$). Expressed as elasticities of log measurement error the ordering is the same: 0.154, 0.081, 0.056 and 0.038 respectively.

The paired within-system test, which removes any sample-size artefact common to both domains, gives a difference in log measurement error between creative thinking and mathematics of 0.117 (SE 0.021, $p = 5.6 \times 10^{-7}$, $R^2 = 0.339$). An outlier-robust Theil–Sen estimator gives 0.133 (95 per cent confidence interval 0.088 to 0.178). The relationship survives adjustment for the domain mean score (0.064, $p < 0.0001$) and for the number of subnational reporting regions (0.040, $p = 0.019$), both of which contribute independently. Figure 7 shows the relationship and its localisation: the effect appears in measurement error and not in true-score dispersion, which is unrelated to the number of students assessed ($p = 0.66$).

This analysis is exploratory and its evidence is not uniform across the range. The coefficient is stable under perturbation, with all 63 leave-one-out estimates positive and between 0.111 and 0.128, and a bootstrap interval of 0.081 to 0.147. It is also robust to the choice of estimator and to influence diagnostics. However, deleting the largest systems progressively removes statistical significance while leaving the coefficient largely intact: with the ten largest systems excluded the coefficient is 0.087 and $p = 0.071$, and with fifteen excluded it is 0.083 and $p = 0.115$. The relationship is therefore established across the full range of participating sample sizes and is driven by the contrast between the largest systems and typical ones; it is not demonstrated among typical-sized systems alone. The interpretation in terms of coding capacity is discussed, with its limitations, in the following sections.

Discussion

Cross-system comparison of predictive performance requires a reference point

The ceiling varies from 0.654 to 0.933 across the domain by system combinations examined here. A model that attains an R^2 of 0.45 in a system whose ceiling is 0.65 has captured a little over two-thirds of the variance available to it; the same R^2 in a system whose ceiling is 0.93 has captured less than half. Reporting the two side by side, as equal, misrepresents both. The practical recommendation is modest and easy to adopt: alongside any reported predictive performance from an international assessment, report the ceiling for that domain and system, and report performance as a share of it. The quantity requires no additional data and no assumption beyond those already made in generating the plausible values.

That normalisation is not cosmetic. Expressed as a share of the attainable bound, between 53 and 70 of the systems change rank, by up to 20 positions, and the direction of movement is systematic: lower-ceiling systems gain. Any exercise that ranks education systems by how predictable their achievement is, or that compares model performance across systems as evidence about those systems, is therefore sensitive to whether this correction is applied.

Predictor rankings from a single plausible value are one draw from an uncharacterised distribution

The instability we observe is large in absolute terms. For creative thinking, a little over 40 per cent of a five-predictor list changes between two plausible values of the same students. Because the attribution null shows this exceeds the churn produced by the importance estimator alone, it cannot be dismissed as an artefact of permutation importance, although we also find that permutation importance over correlated context indices is itself unstable and that the two sources compound.

The implication for practice is specific. Studies that report a ranked list of important predictors from an international assessment should fit across all plausible values and report a stability statistic alongside the ranking, rather than reporting a single ranking from a single draw. Where a study's central claim is the identity of the leading predictors, a stability statistic is not a robustness check but part of the estimate.

The graded relationship is the evidence, not the individual contrasts

Any single contrast between two domains could be explained by something other than measurement precision, since domains differ in content, item format and testing time as well as in reliability. The graded pattern is harder to explain that way. Ordering six paired comparisons by the size of the ceiling gap, every downstream difference declines with it, significance degrades in step, and the pair with almost no ceiling gap shows almost no downstream difference. The fitted relationship passes close to the origin without that having been imposed. A confound would have to vary across the six pairs in proportion to the ceiling gap to produce this pattern.

Two opposing biases operate at once

Plausible values are drawn from a posterior conditioned on principal components of the background questionnaire (Sinharay et al., n.d.). The predictors used in secondary analysis are therefore, in part, the same information on which the outcome was conditioned. Two biases consequently operate in opposite directions in every questionnaire-based prediction of achievement in these databases. Measurement noise deflates attainable performance, which is the bound quantified here. Conditioning inflates the apparent association between questionnaire variables and the outcome, because the posterior mean is

partly a function of those variables. Neither is acknowledged in the applied literature we reviewed. It follows that ρ bounds the prediction of a plausible value, not of an error-free true score, and that the share of the bound attained by questionnaire-based predictors is not directly interpretable as the share attainable by arbitrary predictors. Quantifying the second bias requires access to the conditioning model and is beyond the scope of this paper, but analysts should not assume that the two effects cancel.

Averaging plausible values changes the bound in a country-dependent way

Averaging the ten plausible values before analysis, a shortcut the developers of the method discourage (von Davier et al., 2009), raises the ceiling by a mean of 0.174 for creative thinking and 0.097 for mathematics, reaching 0.296 in the most affected system. Because the inflation depends on both domain and system, two studies that differ only in this analytic choice will differ in apparent bound by an amount confounded with the country they analyse. This is a concrete reason to follow the existing recommendation, beyond correctness of standard errors.

Validation design is a measurement decision, not a convenience

Random student-level splits produced higher performance than school-disjoint splits in 604 of 606 cells, and the inflation is governed by how many schools a system sampled rather than by the learner. In the extreme case the same pipeline on the same data yields 0.002 under school-disjoint evaluation and 0.361 under random splitting. That grouped cross-validation is appropriate for clustered data is well known; what these figures establish is the magnitude in this setting and the parameter that predicts it, which allows a reader to judge how much a published random-split estimate is likely to be inflated from the sampled-school count alone.

Limitations

Several limitations qualify these results, and we state them in the order in which we judge them to matter.

The exploratory analysis of coding intensity is the weakest component and we present it as such. Its predictor is the number of students assessed, which is very highly correlated with the volume of open responses a system had to code but is not a measurement of coding capacity; the number of items per student contributes nothing once sample size is included. The interpretation in terms of coding capacity is therefore an inference from a graded differential across domains, not an observed mechanism. The PISA 2022 public release contains no coder-level, multiple-coding or inter-coder reliability variables in any of its files, which we verified by searching all 28802 variable labels in the official codebook; the interpretation cannot be adjudicated from these data. The coding-reliability statistics that the OECD reports separately would allow it, and we were unable to obtain them during this work. The evidence is also not uniform across the range: while the coefficient is stable under leave-one-out, bootstrap and outlier-robust estimation, statistical significance does not survive deletion of the ten largest systems. The relationship is established across the full range of participating sample sizes and driven by the contrast between the largest systems and typical ones, and should not be read as operating among typical-sized systems alone. Finally, this hypothesis was generated during analysis rather than specified in advance, and the reported p-values are not adjusted for that search.

The ceiling bounds the prediction of a plausible value. Because the posterior mean is itself conditioned on background information, the quantity is not a bound on the prediction of an error-free true score, and the formal bound for the conditioned case is not derived here.

Hyperparameters were held fixed rather than tuned within nested cross-validation, and predictor importance was obtained by permutation rather than by a fuller attribution method. Both decisions were made so that the pipeline is identical across every domain and system, which is what makes the paired comparisons interpretable; both will change absolute performance figures but not the paired contrasts, since those hold the pipeline fixed.

The reported estimates use the final student weight throughout. The 80 replication weights are present, complete and verified functional, and the combination rules for sampling and imputation variance are specified, but the headline tables are not yet adjusted for the full complex design variance. Sampling strata are used descriptively rather than entering the variance estimation.

Finally, all analyses are cross-sectional and associational. Nothing here supports a causal interpretation, and predictor importance is reported strictly as a property of a predictive model rather than as an effect.

Conclusions

How much a predictive model can establish about student achievement in an international assessment is bounded by how precisely the achievement was measured, and that bound is a property of each domain in each education system rather than a constant. Across PISA 2022 it ranges from 0.654 to 0.933. It is weakly related across domains within the same students, so it cannot be inferred from one domain to another. It governs not only attainable accuracy but the stability of the predictor rankings that applied studies report, and it does so in a graded way that approaches zero as the bound difference approaches zero.

Two practical recommendations follow, both inexpensive. Report the ceiling alongside any measure of predictive performance from these databases, and express performance as a share of it, so that figures from different domains and systems can be compared. And when reporting which predictors matter, fit across all plausible values and report a stability statistic, because a ranking from a single draw is one realisation of a distribution that is, in the least precisely measured domain, wide enough to change most of a five-predictor list.

The wider point is that the analytic capacity of a large-scale assessment database is not uniform across the systems it covers. Where measurement is less precise, less can be learned, and conclusions are less stable. Since precision is partly a design and operational property rather than a fact of nature, it is also,

in principle, something that assessment programmes can change.

Abbreviations

• BRR
balanced repeated replication
• CI
confidence interval
• CV
cross—validation
• ICC
intraclass correlation coefficient
• ILSA
international large—scale assessment
• IRT
item response theory
• NAEP
National Assessment of Educational Progress
• OECD
Organisation for Economic Co—operation and Development
• PISA
Programme for International Student Assessment
• PV
plausible value
• RMSE
root mean squared error
• SD
standard deviation
• SE
standard error

Declarations

In accordance with this journal's double-anonymous review policy, the data availability, ethics, competing interests, authors' contributions and acknowledgements statements are provided in the submission system and will be published alongside the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contribution

J.A.M. conceived and designed the study, obtained and verified the data, conducted all analyses, produced the figures and tables, and wrote the manuscript. J.A.M. is the sole author and takes full responsibility for the content.

Data Availability

This study analysed publicly available secondary data. The data are the OECD Programme for International Student Assessment (PISA) 2022 international database, available without restriction from https://www.oecd.org/en/data/datasets/pisa-2022-database.html. The specific files used were the student questionnaire file (CY08MSP_STU_QQQ), the school questionnaire file (CY08MSP_SCH_QQQ) and the creative thinking file (CY08MSP_CRT_COG), retrieved from https://webfs.oecd.org/pisa2022/ in August 2026. Byte sizes and SHA-256 hashes identifying the exact release are reported in Additional file 1. All analysis code, all derived data tables, and machine-readable files containing the exact values underlying every figure are openly archived at https://doi.org/10.5281/zenodo.21984245 (Zenodo). No new data were collected from human participants. Software: analysis conducted in Python 3.14 with numpy 2.4.4, pandas 3.0.2, scikit-learn 1.9.0, scipy 1.17.1, statsmodels 0.14.6 and pyreadstat 1.3.6; random seed 20220 throughout. Archived version: https://doi.org/10.5281/zenodo.21984245. Platform independent. Licence: MIT for code, CC BY 4.0 for derived data tables. No restrictions on use.

References

1. APA style, as the journal specifies Every DOI below was resolved against Crossref or DataCite. Web resources are given as numbered reference entries with access dates, as required, rather than as inline links.

2. Gell, M., Eickhoff, S. B., Omidvarnia, A., Küppers, V., Patil, K. R., Satterthwaite, T. D., Müller, V. I., & Langner, R. (2024). How measurement noise limits the accuracy of brain-behaviour predictions. *Nature Communications*, 15. https://doi.org/10.1038/s41467-024-54022-6
3. Gutiérrez-de-Rozas, B., López-Martín, E., & Carpintero Molina, E. (2022). Defining the profile of students with low academic achievement: A cross-country analysis through PISA 2018 data. *Frontiers in Education*. https://doi.org/10.3389/feduc.2022.910039. 7.
4. König, C., Fink, A., & Frey, A. (2026). Effects of the highly adaptive testing design for PISA on plausible value-based population estimates. *Large-scale Assessments in Education*, *14*. https://doi.org/10.1186/s40536-026-00309-7
5. Lage-Castellanos, A., Valente, G., Formisano, E., & De Martino, F. (2019). Methods for computing the maximum performance of computational models of fMRI responses. *PLOS Computational Biology*, *15*(3). https://doi.org/10.1371/journal.pcbi.1006397
6. Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika*, *81*(2), 274–289. https://doi.org/10.1007/s11336-016-9497-x
7. National Center for Education Statistics (2008). Using plausible values to estimate the proportions of variance explained by student sampling and measurement error. NAEP technical documentation. https://nces.ed.gov/nationsreportcard/tdw/analysis/2000_2001/summary_proportion.aspx. Accessed 17 August 2026.
8. OECD. (2023). PISA 2022 results (I): The state of learning and equity in education. *OECD Publishing*. https://doi.org/10.1787/53f23881-en
9. OECD (2024). PISA 2022 results (Volume III): Creative minds, creative schools. OECD Publishing. https://doi.org/10.1787/765ee8c2-en
10. Anonymous (2026). Measurement-bounded predictive inference in international large-scale assessment: analysis code and derived data [Data set]. Zenodo. https://doi.org/10.5281/zenodo.21984245.
11. OECD (2024). PISA 2022 database. https://www.oecd.org/en/data/datasets/pisa-2022-database.html. Accessed 16 August 2026.
12. Parsons, S. (2022). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *Meta-Psychology*, *6*. https://doi.org/10.15626/mp.2020.2577
13. Robitzsch, A. (2022). Exploring the multiverse of analytical decisions in scaling educational large-scale assessment data: A specification curve analysis for PISA databases. *European Journal of Investigation in Health Psychology and Education*, *12*(7), 731–753. https://doi.org/10.3390/ejihpe12070054
14. Robitzsch, A., & Lüdtke, O. (2022). Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. Measurement Instruments for the Social Sciences, 4. https://doi.org/10.1186/s42409-022-00039-w
15. Sinharay, S., Guo, Z., von Davier, M., & Veldkamp, B. P. (2026). August (n.d.). Assessing fit of latent regression models. In M. von Davier & D. Hastedt (Eds.), IERI monograph series: Issues and methodologies in large-scale assessments (Vol. 3, pp. 35–56). Educational Testing Service and IEA. https://ierinstitute.org/fileadmin/Documents/IERI_Monograph/Volume_3/IERI_Monograph_Volume_03_Chapter_2.pdf. Accessed 16.
16. Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72–101.
17. Uto, M. (2026). Has automated essay scoring reached sufficient accuracy? Deriving achievable QWK ceilings from classical test theory. In Lecture notes in computer science. https://doi.org/10.1007/978-3-032-29744-0_37
18. von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In von M. Davier, & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 2, pp. 9–36). Educational Testing Service and IEA.
19. von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: A pedagogical overview. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705. https://doi.org/10.3102/1076998619881789
20. Bhaktha, N., & Lechner, C. M. (2021). To score or not to score? A simulation study on the performance of test scores, plausible values, and SEM, in regression with socio-emotional skill or personality scales as predictors. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.679481

Figures

Figures 1 to 7 are available in the Supplementary Files section.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfiles.zip
- AdditionalfilesLegends.docx