

Which disadvantaged students beat the odds in STEM? A design-consistent, fairness-audited machine learning study of 143,573 adolescents in 79 PISA 2022 education systems

Omar Allohibi

oalohibi@taibahu.edu.sa

Taibah University

Article

Keywords: academic resilience, STEM education, PISA 2022, machine learning, explainable AI, algorithmic fairness, educational equity, large-scale assessment, socio-economic disadvantage

Posted Date: September 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10713695/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

Socio-economically disadvantaged students who nonetheless reach solid proficiency in mathematics and science (“STEM-resilient” students) show that disadvantage does not preclude STEM success, and identifying the factors that accompany such resilience is a policy priority. Machine learning studies of international large-scale assessment data have begun to address related questions, but existing work pools students without respecting the assessments’ complex survey design, evaluates models only on in-distribution splits, applies a single outcome definition, and has not yet examined whether models err evenly across student groups. Guided by an ecological, contextually situated account of academic resilience, this study used PISA 2022 data from 143,573 disadvantaged adolescents (weighted bottom socio-economic quartile within each of 79 systems) to predict STEM resilience, defined as proficiency Level 3 or above in both mathematics and science, treating all ten plausible values as multiple imputations, weighting estimation and evaluation, and grouping cross-validation by school. STEM resilience characterized 12.5% of disadvantaged students internationally (population-weighted; SE = 0.3%), with system rates ranging from below 0.1% to 62%. A gradient-boosting model predicted resilience accurately and with excellent calibration (AUC = 0.912; 95% cluster-bootstrap CI 0.909–0.914; PR-AUC = 0.688; slope = 0.996), against 0.862 for logistic regression and 0.801 for a system-base-rate-only baseline. Mathematics self-efficacy was the strongest predictor, and rankings were nearly invariant across plausible values ($\tau = .90$). Models transported imperfectly to unseen systems (median held-out AUC = 0.847), and system-specific factor hierarchies diverged from the pooled hierarchy (median $\tau = .30$) beyond a sampling-noise benchmark ($\tau = .45$), although self-efficacy ranked first under every outcome definition. At a prevalence-matched threshold, the model detected resilient girls (59.4% vs 65.6%) and rural students (53.9% vs 66.1%) less often than their counterparts; an attribute-blind refit removed the gender gap but not the rural gap. Ignoring weights or clustering changed results little; changing the outcome definition changed them most. Protective factors associated with STEM resilience are only partly universal, and single-threshold screening uses would overlook resilient girls and rural students. The study offers a design-consistent, transport-tested, fairness-audited template for machine learning with international assessment data.

Introduction

The association between students’ socio-economic background and their performance in mathematics and science is among the most consistently replicated findings in education research (OECD, 2023b; X. Wang et al., 2023). Because competence in these domains anchors access to science, technology, engineering, and mathematics (STEM) study and careers, socio-economic gaps in mathematics and science proficiency at age 15 foreshadow unequal participation in the STEM pathways that societies increasingly depend on (Caspi & Gorsky, 2024; Sevilla et al., 2023; Vooren et al., 2022; Zhao & Perez-Felkner, 2022). Yet the association is far from deterministic. In every education system that participates in the Programme for International Student Assessment (PISA), some students from the least advantaged quarter of households nonetheless perform at levels that prepare them for continued STEM learning. The Organisation for Economic Co-operation and Development (OECD) calls such students

academically resilient and treats their prevalence as a key equity indicator (OECD, 2011, 2023b). Understanding what accompanies this “beating the odds” is a question of clear policy relevance for STEM education: unlike socio-economic origin itself, several of the factors implicated (self-efficacy, anxiety, engagement, school climate, instructional exposure) are at least partly malleable (Agasisti et al., 2021; García-Crespo et al., 2021; H. Huang et al., 2026).

Two research streams motivate the present study. The first is the rapidly growing application of machine learning (ML) to international large-scale assessments (ILSAs). ML methods accommodate many correlated predictors and nonlinear, interactive relationships, and, coupled with explainable-AI tools such as Shapley additive explanations (SHAP), they can rank large candidate sets of factors by their contribution to predictions (Gabriel et al., 2018; Gamazo & Martínez-Abad, 2020; Lezhnina & Kismihók, 2022; Masci et al., 2018; Yoo, 2018; F. Wang et al., 2026). Within STEM education specifically, this approach has been used to identify determinants of science achievement across 53 systems (J. Hu et al., 2022), of reading and mathematics proficiency in single systems (Bernardo et al., 2021; Y. Huang et al., 2024; H. Liu et al., 2022), and of anxiety-related predictors via ML variable selection (Immekus et al., 2022), and, at the largest scale, of a binary “STEM competency” outcome among 522,802 adolescents in PISA 2022 (J. Liu et al., 2026). A smaller strand has turned ML toward academic resilience: random forests with SHAP applied to mathematics resilience in the pooled PISA 2022 data (Cheung et al., 2024), support vector machines for digital-reading resilience in seven Asian economies (Zheng et al., 2024), analyses within single systems (Jin et al., 2022), and a regional explainable-ML preprint comparing resilience definitions across nine Latin American countries (Delprato & Sandoval-Hernández, 2025).

The second stream is a longer, mostly regression-based literature on academic resilience itself. This work has established that resilience prevalence differs sharply across systems; that truancy, self-efficacy, emotional stability, aspirations, and orderly, supportive school climates are recurrent correlates (Agasisti et al., 2018, 2021; García-Crespo et al., 2021; Sandoval-Hernández & Białowolski, 2016); and, critically, that conclusions can shift with the operational definition of resilience adopted, of which the literature contains dozens (Rudd et al., 2021; Ye et al., 2021).

Reading these streams together exposes four unresolved problems that this study addresses.

First, a design problem. ILSAs are not simple random samples: PISA uses two-stage stratified sampling with final student weights, balanced repeated replication (BRR) weights for variance estimation, and ten plausible values (PVs) that represent proficiency as a posterior distribution rather than a point score (OECD, 2009, 2024; Rutkowski et al., 2010). The ML strand has largely set this machinery aside: averaging plausible values or using only the first, ignoring weights, and splitting students at random across training and test sets despite school clustering (F. Wang et al., 2026). Each shortcut has known consequences: mishandled PVs distort measurement-error propagation; unweighted analyses describe the dataset rather than any student population; random splits allow information to leak between schoolmates, inflating apparent accuracy. How much these shortcuts matter for a resilience classification task is, to date, unquantified.

Second, a generalizability problem. Pooled multi-country models are routinely interpreted as identifying “key determinants” writ large, yet they are validated only on in-distribution splits of the same pooled data (Cheung et al., 2024; J. Liu et al., 2026). Whether a model trained on other systems predicts resilience in a system it has never seen (and whether the protective-factor hierarchy it implies holds there) is a different question, analogous to external validation in clinical prediction research (Steyerberg & Harrell, 2016). Education studies have examined model transfer across institutions and populations (Gardner et al., 2023; Ocumpaugh et al., 2014), but no published ILSA study was located that performs out-of-system validation of a student-level prediction model.

Third, an equity-of-errors problem. A classifier can be accurate on average while distributing its errors unevenly: missing resilient girls more often than resilient boys, for example. Algorithmic-fairness auditing is well developed for institutional learning analytics (Baker & Hawn, 2022; Deho et al., 2022; Gardner et al., 2019; Kizilcec & Lee, 2022; Opoku et al., 2025; Yu et al., 2021), but no fairness audit of any ILSA-based prediction model was found. This omission matters for models whose implied use is equity monitoring: students whom a resilience model fails to recognize are also those whom an equity policy relying on it would fail to reach.

Fourth, a definition problem. ML resilience studies typically fix a single outcome definition, although definitional sensitivity is one of the field’s most robust cautions (Rudd et al., 2021; Ye et al., 2021). Whether the predictive hierarchy of protective factors is itself definition-dependent remains untested at scale.

To date, no existing study addresses more than part of this space. The closest, Cheung et al. (2024), models mathematics-only resilience in pooled PISA 2022 data with internal splits, without design-based estimation, transport tests, fairness auditing, or definitional sensitivity; regional and single-system studies narrow the scope further (Delprato & Sandoval-Hernández, 2025; Jin et al., 2022; Zheng et al., 2024); and the largest-scale STEM prediction study to date (J. Liu et al., 2026) concerns the full student population and a sample-relative competency label rather than resilience among the disadvantaged. The present study combines these elements: a STEM-specific resilience outcome (proficiency in both mathematics and science) among disadvantaged students only, modeled with design-consistent ML, validated across systems, audited for subgroup fairness, and stress-tested across outcome definitions.

Theoretical framework

Academic resilience is framed not as a trait but as an inference from two observations: exposure to adversity and better-than-expected adaptation (Masten, 2001; Ungar, 2008). In Masten’s (2001) account, such adaptation is “ordinary magic”: it arises when fundamental adaptive systems (self-beliefs and self-regulation, attachment and family support, and effective schools) operate normally despite adversity. This protective-factor view supplies the candidate predictors and, importantly, a warrant for focusing on disadvantaged students specifically: the correlates of high performance in the full population need not

be the factors that distinguish thriving from struggling among the disadvantaged, because protective processes may be adversity-specific.

Where those adaptive systems sit is organized by Bronfenbrenner's bioecological model (Bronfenbrenner & Morris, 2006): person characteristics (gender, immigrant background, expectations, self-efficacy, anxiety, dispositional attributes), proximal processes in microsystems (family support, teacher support, cognitive activation, instructional exposure, disciplinary climate, bullying), and features of the surrounding context (school resources, staffing, location, sector). The predictor set operationalizes each layer with PISA 2022 measures, and the interpretive analyses ask which layers carry predictive weight.

Two further theoretical commitments generate the study's distinctive questions rather than merely organizing its variables. The first comes from social-cognitive theory: self-efficacy (belief in one's capability to organize and execute action) is hypothesized to be a proximal engine of persistence and performance under adversity (Bandura, 1997), and motivational theory more broadly positions competence beliefs as central to achievement choices (Eccles & Wigfield, 2020). If protective factors have a universal core, domain-specific self-efficacy is its most likely occupant, and it should dominate predictions under any defensible outcome definition. The second comes from Ungar's (2008, 2011) social-ecological account of resilience: protective processes are negotiated within cultural and institutional contexts, so their relative importance should be *contextually variable*. This claim has rarely been tested quantitatively at scale. This study operationalizes it as a transportability question: if resilience processes are largely universal, a model trained on other systems should predict well in an unseen system and imply a similar factor hierarchy there; if they are contextually contingent, transported performance should degrade and hierarchies should diverge. Cross-system validation therefore tests a theoretical prediction rather than serving only as a robustness check.

The present study

Within this framework, and given the four gaps above, this study asks:

RQ1 (Predictability under design consistency). How accurately can STEM resilience among disadvantaged adolescents be predicted from student-, family-, and school-level factors when the PISA survey design (final student weights, all ten plausible values, school clustering) is fully respected, and how do design-naïve shortcuts change the answer?

RQ2 (Protective-factor hierarchy). Which factors contribute most to model predictions of STEM resilience, in which direction, and with what nonlinearities?

RQ3 (Transportability). To what extent do the model and the protective-factor hierarchy it implies transport to education systems excluded from training?

RQ4 (Fairness). Are prediction errors distributed evenly across gender, immigrant background, home-language, and urban–rural subgroups?

The study contributes to STEM education research in three ways. Substantively, it provides a global, design-consistent map of factors that accompany dual-domain STEM resilience among disadvantaged adolescents, and a quantitative test, apparently the first at this scale, of the universality-versus-contextuality of those factors across 79 systems. Methodologically, it demonstrates, and quantifies the consequences of, a fully design-consistent ML pipeline for ILSA data, including a transport-validated country-grouped cross-validation scheme and what is, to date, the first subgroup fairness audit of an ILSA-based classifier. For policy and practice, it clarifies both what such models can offer (well-calibrated risk stratification; identification of malleable correlates such as mathematics self-efficacy) and what they cannot yet be trusted to do (transport unmodified across systems; screen subgroups evenhandedly).

Methods

Data source and participants

This study analysed the public-use PISA 2022 database (OECD, 2023a), which contains nationally representative samples of 15-year-old students drawn through a two-stage stratified design (schools, then students within schools). The student questionnaire file comprises 613,744 students from 80 education systems and 21,629 schools. Figure 1 documents the construction of the analytic sample. Costa Rica was excluded because the socio-economic index (ESCS) is unavailable for all of its 6,113 records in the public file, making disadvantage undefinable. In the remaining 79 systems, 19,355 students (3.2%) missing ESCS were excluded for the same reason. Following the OECD definition, students were classified as socio-economically disadvantaged if their ESCS fell at or below the weighted (final student weight, W_{FSTUWT}) 25th percentile of ESCS within their own system (OECD, 2023b). The analytic sample comprised $N = 143,573$ disadvantaged students in 79 systems and 19,345 schools (51.3% girls; 6.7% first-generation and 8.4% second-generation immigrant students; 38.2% attending schools in villages or small towns; 12.0% in private schools). All 79 systems contributed between 711 and 6,817 students (median 1,583).

Outcome: STEM resilience

Because PISA proficiency is reported as ten plausible values (PVs) per domain (random draws from each student's posterior proficiency distribution), the outcome was constructed separately within each PV, never from averaged PVs (OECD, 2009; Rutkowski et al., 2010; von Davier et al., 2009). For plausible value m ($= 1 \dots 10$), a disadvantaged student was classified as STEM-resilient if they reached proficiency Level 3 or above in *both* mathematics ($PV_m\text{MATH} \geq 482.38$) and science ($PV_m\text{SCIE} \geq 484.14$), using the official PISA 2022 level boundaries (OECD, 2023b). This criterion-referenced definition has three advantages over sample-relative alternatives: it is substantively interpretable (Level 3 exceeds the Level-2 baseline that OECD regards as minimum proficiency, in both core STEM domains, indicating a foundation for further STEM learning); it is identical across systems, so the label does not encode

country membership; and it does not depend on the estimation sample. Because definitional choices are known to matter (Rudd et al., 2021; Ye et al., 2021), the core analysis was repeated under three alternatives: S1, Level 2 or above in both domains (“baseline resilience”); S2, the OECD (2023b) national relative definition, scoring in the weighted top quarter of mathematics within one’s own system; and S3, Level 3 or above in mathematics only, which isolates the contribution of the dual-domain requirement.

Predictors

Fifty-three predictors were selected a priori to operationalize the three layers of the theoretical framework (full definitions, PISA variable names, and codings in Supplementary Table S2). *Person* measures included gender, age, relative grade, immigrant background, grade repetition, expected educational attainment, mathematics self-efficacy in formal/applied tasks (MATHEFF) and in mathematical reasoning/21st-century tasks (MATHEF21), mathematics anxiety, effort and persistence in mathematics, subject preference and perceived ease and motivation, self-reported test effort, and the PISA 2022 dispositional scales (curiosity, perseverance, cooperation, empathy, assertiveness, stress resistance, growth mindset), plus self-directed-learning self-efficacy. *Microsystem* measures included family support, family support for self-directed learning, teacher support in mathematics, cognitive activation (fostering reasoning; encouraging mathematical thinking), disciplinary climate, exposure to formal/applied mathematics tasks and to mathematical-reasoning tasks (opportunity to learn), sense of belonging, bullying victimization, feeling safe, student–teacher relationship quality, and out-of-school time use (paid work, household work, study time, sport), truancy, and lateness. *Context* measures from the school questionnaire included school size, mathematics class size, student–teacher and student–mathematics-teacher ratios, proportion of fully certified teachers, material and staff shortages, school selectivity, autonomy, sector (public/private), and community location. ESCS itself was retained as a predictor of within-quartile socio-economic depth, together with home ICT resources; ESCS’s own components (HISEI, PAREDINT, HOMEPOS) were *not* entered separately, to avoid the composite-versus-components redundancy that can distort importance rankings. Continuous questionnaire scales are OECD-derived weighted-likelihood estimates standardized across OECD countries (OECD, 2023b).

Design-consistent estimation

Three design features were respected throughout. Weights. All model fitting and pooled evaluation used senate-style weights, W_FSTUWT rescaled to sum to a constant (1,000) within each system’s disadvantaged subsample, so every system contributes equally to pooled analyses, while within-system descriptive statistics and per-system evaluation used W_FSTUWT itself; design-based standard errors for descriptive quantities used the 80 BRR replicate weights with Fay’s adjustment ($k = 0.5$) (OECD, 2024). Plausible values. The entire pipeline (model fitting, prediction, evaluation, interpretation) was executed ten times, once per PV, and estimates were pooled following multiple-imputation logic: point estimates as PV means, and uncertainty combining within-PV (cross-fold) and between-PV variance with the standard $(1 + 1/M)$ correction (Rubin, 1987; OECD, 2009). PVs were never averaged into a single

score. Clustering. Cross-validation was grouped by school (GroupKFold on school ID, five folds), so schoolmates never straddled a training/test boundary; all preprocessing (imputation medians) and all hyperparameter selection were confined to training folds.

Models, tuning, and evaluation

Five classifiers spanning the linearity spectrum were compared: logistic regression; ridge-penalized logistic regression; random forest; histogram-based gradient boosting (scikit-learn); and LightGBM gradient boosting (Ke et al., 2017), designated a priori as the primary model. Two reference points bound these comparisons from below: a system-base-rate-only “model” (each student assigned their system’s training-fold resilience rate), which quantifies how much apparent accuracy reflects geography alone, and main-effects logistic regression, which quantifies the gain from nonlinearity. Tree-based gradient boosting handled missing predictor values natively; logistic and random-forest models used training-fold median imputation. Hyperparameters were selected once, before any test fold was evaluated, using only the training partition of the first outer fold: inner two-fold school-grouped cross-validation on a 50% school subsample, over 17 logged trials covering eight distinct LightGBM configurations (two grid stages) and small grids for ridge (C) and random forest (minimum leaf size). Following the one-standard-error parsimony principle (Hastie et al., 2009), the most parsimonious LightGBM configuration within one standard error of the best was adopted (31 leaves, learning rate 0.05, 400 trees, minimum 200 observations per leaf; inner AUC 0.9086 ± 0.0019 vs 0.9097 ± 0.0016 ; full trial log in Supplementary Table S3). Because this selection reused data that later serve as test folds 1–4, the absence of selection optimism was verified on the fully untouched fold 0, whose test AUC (0.913, mean over ten PVs) did not fall below the all-fold average (0.912). The primary model was evaluated across all ten PVs (50 model fits); comparison models across PVs 1–3 (60 fits), which is sufficient because between-PV variation in their performance proved negligible (≤ 0.002 AUC). ROC-AUC is reported together with the precision–recall AUC (the informative metric given the 15.2% senate-weighted prevalence of the positive class), the Brier score, and calibration slope and intercept from logistic recalibration of out-of-fold predictions; all pooled metrics are senate-weighted. For the headline discrimination estimate and the fairness gaps, 95% confidence intervals are additionally reported from a school-cluster bootstrap (200 resamples of schools with replacement over the pooled out-of-fold predictions), which respects the clustering that conventional binomial intervals ignore; the SE reported alongside Rubin-pooled quantities combines between-PV and between-fold variability and should be read as such rather than as a full design-based standard error.

Transportability analysis

To test whether models generalize to unseen systems (RQ3), country-grouped eight-fold cross-validation was used: systems were randomly partitioned into eight groups, and each group was predicted by a model trained on the other seven, so every student’s transported prediction comes from a model that saw *no data from their system*. This is the internal–external validation logic of clinical prediction research

(Steyerberg & Harrell, 2016) applied at the system level. This computationally efficient scheme was verified to approximate strict leave-one-system-out validation by also fitting strict LOCO models for 12 randomly sampled systems (PV1; 11 estimable, one excluded for a degenerate outcome distribution); agreement was high ($r = .84$; mean absolute difference = .010), so country-grouped estimates are reported for all 79 systems (PVs 1–3). To compare protective-factor *hierarchies*, within-system models were additionally fitted (systems with ≥ 300 disadvantaged students and ≥ 20 resilient cases; 67 systems qualified), using a lighter configuration suited to small samples (≤ 300 trees, minimum 20 observations per leaf), and Kendall's τ was computed between each system's SHAP importance ranking and the pooled ranking. Because small-sample SHAP rankings are noisy, an explicit noise benchmark was constructed: 30 random subsamples of the *pooled* data at the median system size ($n = 1,583$), refitted with the same lighter configuration, yield the τ distribution expected under sampling noise alone (i.e., with no true between-system heterogeneity); observed cross-system τ values are interpreted against this floor.

Interpretation and fairness audit

Global and local interpretation used TreeSHAP (Lundberg et al., 2020; Lundberg & Lee, 2017) applied to the primary model refitted on the full pooled sample, separately per PV, on weighted subsamples of 15,000 students; the analysis reports PV-averaged mean absolute SHAP values with between-PV variability, directional associations, and dependence plots, and quantifies the stability of the importance ranking across PVs with Kendall's τ . Because SHAP quantifies contributions to *predictions*, not causal effects, all rankings are interpreted associationally. The fairness audit (RQ4) followed the error-rate-parity tradition (Baker & Hawn, 2022; Gardner et al., 2019; Hardt et al., 2016): out-of-fold predictions were thresholded at a prevalence-matched operating point (students in the top 15.2% of predicted probability classified positive, matching the senate-weighted prevalence per PV), and true-positive rates, false-positive rates, subgroup AUC, and subgroup calibration were compared for gender, immigrant background (first-generation vs non-immigrant), home language (different from vs same as test language), and school location (rural/small town vs town/city), pooled by Rubin's logic across the ten PVs. Because protected attributes were among the model's features, a design choice with known fairness trade-offs (Yu et al., 2021), the audit was repeated with an attribute-blind model (gender, immigrant-background, and community-location features removed; PV1), and how the gaps behave across alternative operating points (top 10%, 15%, 20%, 25%) was examined.

Design-ablation experiment

To quantify the consequences of common design shortcuts (RQ1), the primary model was re-run under a factorial ablation holding data, features, and algorithm fixed: {school-grouped vs random student-level splits} $\times$ {senate-weighted vs unweighted}, plus a fifth cell replacing the criterion-referenced outcome with a benchmark-style label, binarizing the average of PV-averaged mathematics and science scores at the pooled unweighted mean (cf. J. Liu et al., 2026), under random unweighted splitting.

Results

STEM resilience is rare, unevenly distributed, and predictable

Across the 79 systems, 12.52% of disadvantaged students (population-weighted; design-based SE = 0.30%, combining BRR sampling variance and between-PV variance) reached Level 3 in both mathematics and science; weighting systems equally, the average of system rates was 15.25%. Weighted descriptive statistics for the key variables, overall and by resilience status, appear in Table 1. The international averages conceal an enormous range (Fig. 2; Supplementary Table S1): system rates ran from below 0.1% in the lowest-performing systems to 62% in Singapore, suggesting that resilience against a fixed international criterion is as much a property of systems as of students (cf. OECD, 2023c).

Table 1
Weighted descriptive statistics of key variables, overall and by STEM-resilience status (PV1)

Variable	Missing %	Weighted mean	SD	Mean: resilient	Mean: not resilient
AGE	0	15.82	0.293	15.819	15.82
GRADE	4.238	-0.23	0.68	-0.002	-0.263
ESCS	0	-2.033	0.916	-1.294	-2.14
ICTRES	41.235	-1.684	1.268	-0.612	-1.799
EXERPRAC	5.592	4.041	3.704	3.245	4.161
WORKPAY	7.135	1.711	3.022	0.606	1.878
EXPECEDU	29.554	6.113	2.338	6.244	6.089
MATHEFF	24.525	-0.764	1.143	-0.156	-0.891
MATHEF21	22.801	-0.125	0.976	0.032	-0.152
ANXMAT	23.727	0.387	1.114	0.031	0.469
EFFORT1	16.742	8.125	1.836	7.905	8.16
STRESAGR	52.296	-0.091	0.974	-0.012	-0.111
FAMSUPSL	46.473	-0.057	1.05	-0.506	0.027
BELONG	7.936	-0.237	0.848	-0.132	-0.253
BULLIED	6.671	-0.273	1.073	-0.455	-0.246
DISCLIM	12.112	0.061	1.016	0.378	0.011
TEACHSUP	16.823	0.232	1.126	0.186	0.24
COGACRCO	16.384	-0.001	1.171	0.046	-0.008
EXPOFA	17.385	-0.018	1.069	-0.13	0
EXPO21ST	18.836	0.123	1.095	-0.074	0.157
STRATIO	15.78	18.075	11.606	14.182	18.611
SCHSIZE	12.911	939.192	1075.78	986.806	932.504
EDUSHORT	10.733	0.29	1.311	-0.112	0.346
FEMALE	0.017	0.532	0.499	0.458	0.543
REPEAT	3.74	0.167	0.373	0.048	0.182
<i>Note. Means and SDs use final student weights (W_FSTUWT); group means condition on the PV1 outcome. Variable definitions: Supplementary Table S2.</i>					

Variable	Missing %	Weighted mean	SD	Mean: resilient	Mean: not resilient
IMMIG3	6.02	0.038	0.192	0.054	0.036
<i>Note. Means and SDs use final student weights (W_FSTUWT); group means condition on the PV1 outcome. Variable definitions: Supplementary Table S2.</i>					

Table 2 summarizes model comparison under the design-consistent pipeline (school-grouped five-fold cross-validation, senate weights). The primary LightGBM model discriminated well, AUC = 0.912 (combined between-fold/between-PV SE = 0.001; 95% school-cluster-bootstrap CI 0.909–0.914), PR-AUC = 0.688 against the 15.2% senate-weighted prevalence baseline, and was excellently calibrated (slope = 0.996, intercept = 0.022; Fig. 3d), with a Brier score of 0.077. Histogram gradient boosting performed equivalently (AUC = 0.910); random forest trailed (AUC = 0.887) and was poorly calibrated (slope = 1.50); logistic and ridge logistic regression reached AUC = 0.862. Two baselines locate these numbers. A system-base-rate-only model (no student information at all) already achieved AUC = 0.801 (Table 3), so roughly speaking, much of the pooled discrimination reflects *where* a student goes to school rather than student-level differences; the pooled model’s discrimination *within* systems (internal out-of-fold predictions evaluated system by system) averaged 0.85. And the ~ 0.05 AUC advantage of gradient boosting over well-tuned linear models indicates material nonlinearity and interaction structure. Performance was almost identical across the ten plausible values (LightGBM between-PV SD = 0.0006), and Fig. 3b–c shows the ten PV-specific ROC and precision–recall curves to be visually indistinguishable.

Table 2

Model comparison under design-consistent evaluation (school-grouped five-fold CV, senate-weighted)

Model	PVs	AUC	PR_AUC	Brier	Cal_slope	Cal_intercept
LightGBM (primary)	1–10	0.912 (0.002)	0.687 (0.010)	0.0766	0.996	0.022
Histogram gradient boosting	1–3	0.910 (0.002)	0.684 (0.010)	0.0772	0.986	0.039
Random forest	1–3	0.887 (0.002)	0.635 (0.008)	0.0874	1.504	0.554
Ridge logistic regression	1–3	0.862 (0.002)	0.567 (0.011)	0.0920	1.002	0.002
Logistic regression	1–3	0.862 (0.002)	0.567 (0.011)	0.0919	0.994	-0.007
System base rate only (no student data)	1–3	0.801	N/A	N/A	N/A	N/A
<i>Note. Cells show mean (SD) across folds × PVs. The primary model additionally has a 95% school-cluster-bootstrap CI for AUC of [0.909, 0.914]. PR-AUC baseline = senate-weighted prevalence (0.152).</i>						

The design-ablation (Supplementary Table S5; Supplementary Fig. S1) contextualizes these results. Ignoring weights changed AUC modestly (0.907 unweighted vs 0.912 weighted) but changed the estimand: performance for an equal-system population versus the pooled respondent pool. Replacing school-grouped with random student-level splits inflated AUC by only ≈ 0.001 in this application, indicating that cluster leakage, though real, was minor here. The most consequential choice was the outcome definition itself: the benchmark-style mean-split composite label yielded AUC = 0.889 on the identical sample and features, an estimate of a different quantity. Sensitivity analyses over resilience definitions reinforced this: predictability was high for the Level-2 variant (S1: AUC = 0.882 at 34.3% senate-weighted prevalence) and for mathematics-only Level 3 (S3: AUC = 0.901 at 18.9% prevalence), and lowest for the fully national-relative OECD definition (S2: AUC = 0.836 at 11.5% prevalence). The dual-domain requirement binds: relative to mathematics-only (S3), requiring science Level 3 as well reclassifies 3.8% of students and reduces the senate-weighted prevalence by 3.7 percentage points, confirming that the STEM construct is not redundant with mathematics alone. The ordering across definitions is equally informative: definitions that retain between-system variation in the label (the primary, S1, S3) are partly predictable *because* systems differ, whereas the within-system definition (S2) removes that source of signal. Pooled discrimination numbers in this literature should therefore not be read as within-system screening accuracy, a point the baseline decomposition above and the transport analysis below make directly.

Table 3
Estimand decomposition and cross-system transport summary

Analysis	Result
Pooled internal CV, school-grouped + senate weights (primary)	AUC 0.912 (95% cluster-bootstrap CI 0.909–0.914)
System-base-rate-only baseline (PVs 1–3)	AUC 0.801
Within-system discrimination of pooled model (internal, PVs 1–3)	mean AUC 0.854
Pooled transported (country-grouped 8-fold, PVs 1–3)	AUC 0.894
Per-system transported AUC, median [IQR]	0.847 [0.826, 0.873]
Strict LOCO vs grouped (11 systems, PV1)	$r = .84$; mean $ \Delta = .010$
Hierarchy consistency: system vs pooled SHAP, median τ [IQR] (67 systems)	0.30 [0.23, 0.38]
Sampling-noise τ benchmark (30 pooled subsamples, $n = 1,583$)	median 0.451 [IQR 0.417, 0.492]; min 0.357
<i>Note. All AUCs senate-weighted unless noted; per-system transported metrics use W_{FSTUWT} within systems.</i>	

A stable global hierarchy of protective factors, dominated by mathematics self-efficacy

Figure 4 presents the global interpretability results. Mathematics self-efficacy in formal and applied tasks was by a wide margin the most important predictor (PV-averaged mean |SHAP| = 0.462), followed by within-quartile socio-economic position (0.360), home ICT resources (0.336), and stress resistance (0.282). Higher self-efficacy, deeper (relatively less disadvantaged) ESCS, more home ICT resources, higher expected education, greater test effort, larger schools, higher relative grade, and more cognitive activation in mathematics lessons were associated with higher predicted resilience; being female, paid work, sport frequency (reflecting overall out-of-school time allocation), grade repetition, and mathematics anxiety with lower predicted resilience. Stress resistance showed a distinctly non-monotonic signature (Fig. 5): predictions were suppressed at low values but plateaued beyond the middle of the scale. Family support for self-directed learning carried a *negative* association (plausibly a compensatory pattern, with families mobilizing support precisely for students who struggle), which underscores the associational, non-causal status of these rankings. Dependence plots (Fig. 5) further show that the self-efficacy gradient is steepest in its upper range, that ICT resources matter most at the very bottom of the resource distribution, and that even modest paid-work involvement is associated with lower predicted resilience.

Importantly for practice, this hierarchy was essentially invariant to the plausible-value machinery: the importance ranking's mean between-PV Kendall's τ was .90 (minimum .87; Supplementary Fig. S2), and headline AUCs varied by < 0.002 across PVs. Design consistency, in other words, changed uncertainty statements more than point conclusions. The same cannot be said of the outcome definition. Under the national-relative definition (S2), ESCS and ICT resources dropped out of the top five (which became self-efficacy, gender, stress resistance, expected education, paid work), while under S1 the top five remained (self-efficacy, stress resistance, ICT resources, ESCS, gender): 9 of 10 top-decile features overlapped between S1 and the primary definition versus 7 of 10 for S2. Mathematics self-efficacy ranked first under every definition. Socio-economic depth and home resources, by contrast, behave partly as *between-system* markers: they help predict who clears a fixed international bar, but contribute less to distinguishing nationally exceptional performers within systems. A further check bounds how much the model *depends* on self-efficacy: refitting without both self-efficacy scales reduced AUC only from 0.912 to 0.903, with ESCS, ICT resources, and stress resistance moving up to head the hierarchy; the information self-efficacy carries is thus partially redundant with the remaining motivational and contextual measures, even though it is the single strongest carrier of that information.

Models and hierarchies transport imperfectly across systems

Under country-grouped transport validation (every student predicted by a model trained without any data from their system), pooled discrimination declined from 0.912 to 0.894 (Table 3), and the per-system picture (Fig. 6a) was markedly heterogeneous: median held-out AUC = 0.847 (IQR 0.826–0.873), ranging from 0.598 to 0.993 across the 79 systems. The strict leave-one-system-out spot check agreed closely with the grouped estimates ($r = .84$, mean $|\Delta| = .010$; Fig. 6b). Thus a globally trained model remains

useful in a typical unseen system, but with substantially lower, and highly variable, discrimination than pooled figures suggest, and in the least favorable systems it approaches uselessness.

The hierarchy-level test was starker (Fig. 6c). System-specific SHAP rankings (67 systems) correlated only modestly with the pooled global ranking: median Kendall's $\tau = .30$ (IQR .23–.38), and on average only 1.5 of each system's top-5 factors appeared in the global top-5. Small-sample instability cannot account for this: the sampling-noise benchmark (random pooled-data subsamples of the median system size, fitted identically) produced far higher agreement (median $\tau = .45$, IQR .42–.49, minimum .36), so the observed cross-system divergence lies below the entire noise-floor distribution and reflects genuine between-system heterogeneity (noise nonetheless inflates apparent divergence, so the *true* heterogeneity is more moderate than the raw τ suggests). The global hierarchy is thus a real but weak template for any particular system: quantitative support for the contextual-contingency thesis (Ungar, 2008), with mathematics self-efficacy as the principal universal exception.

The model's errors are not evenly distributed

At the prevalence-matched operating point, subgroup discrimination was uniformly high (AUC .899–.922) and subgroup calibration nearly exact (predicted vs observed rates within 0.5 percentage points for every group; Table 4). Error *rates*, however, diverged (Fig. 7). The model recognized truly resilient boys more often than truly resilient girls (TPR = 65.6% vs 59.4%; a 6.2-point recall gap, 95% cluster-bootstrap CI 4.5–7.9) and urban students more often than rural students (66.1% vs 53.9%; a 12.1-point gap, CI 9.5–13.4); first-generation immigrant and home-language-minority resilient students were recognized *more* often than their counterparts (66.9% vs 61.6%; 70.3% vs 58.9%), with correspondingly higher false-positive rates. The gaps were not artifacts of one operating point: across thresholds from the top 10% to the top 25% of predicted probability, the gender gap ranged from 9.1 to 3.1 points and the rural gap from 12.0 to 9.3. Nor do they all share one mechanism. Refitting the model *without* the protected attributes eliminated (indeed slightly reversed) the gender gap (attribute-blind TPR 63.7% for girls vs 61.6% for boys) at nearly unchanged accuracy (AUC 0.909), indicating that the gender asymmetry was largely attribute-driven: the model had learned to use gender itself against girls' predictions. The rural gap, by contrast, persisted essentially unchanged under attribute blinding (54.4% vs 65.6%), marking it as base-rate- and covariate-driven, the canonical impossibility pattern of fair classification under unequal base rates (Hardt et al., 2016; Kizilcec & Lee, 2022; cf. Yu et al., 2021 on the ambivalent effects of including protected attributes). Substantively: a single-threshold screening use of this model, as estimated, would systematically under-identify resilient girls and resilient rural students, and only the first of those failures is fixable by blinding the model.

Table 4 Subgroup performance at the prevalence-matched operating point (Rubin-pooled over ten PVs)

Group	n	auc	tpr	fpr	obs	pred
Male	69879	0.9122	0.6562	0.0699	0.1706	0.1668
Female	73670	0.9106	0.5943	0.0606	0.1352	0.135
Home language = test language	103572	0.9062	0.5889	0.0602	0.1393	0.1379
Home language $\neq$ test language	40001	0.9215	0.7033	0.0784	0.1864	0.1829
Non-immigrant	125253	0.9105	0.616	0.0633	0.1495	0.147
First-generation immigrant	9677	0.8993	0.6692	0.0895	0.1972	0.1984
Urban/town	83576	0.9039	0.6605	0.0845	0.1853	0.1838
Rural/small town	51671	0.9185	0.5393	0.038	0.1014	0.099

Key contrasts:

Contrast	Value
TPR gap female–male (95% cluster-bootstrap CI)	–6.2 pts [–7.9, –4.5]
TPR gap rural–urban (95% CI)	–12.1 pts [–13.4, –9.5]
Attribute-blind model (PV1): AUC	0.909
Attribute-blind TPR: female vs male	63.7% vs 61.6%
Attribute-blind TPR: rural vs urban	54.4% vs 65.6%

Note. obs = observed weighted resilience rate; pred = mean predicted probability (calibration-in-the-large). TPR/FPR at the top-15.2% operating point; senate weights.

Robustness

Beyond the definitional and ablation analyses above, conclusions were robust to excluding the eight rotated-module scales with the highest missingness (including ICT resources and the dispositional scales): AUC = 0.904 (vs 0.912; Supplementary Table S5), with the remaining hierarchy unchanged in order. Full per-system results, descriptive statistics, tuning trials, and complete SHAP tables are provided in the Supplementary Information (Tables S1–S5).

Discussion

This study set out to learn, at global scale and under full respect for the survey design, which factors accompany STEM resilience among disadvantaged adolescents, and whether the resulting model, its factor hierarchy, and its errors behave in ways that would justify the uses to which such models are increasingly put. Four findings organize the discussion.

Resilience is predictable, and its strongest malleable correlate is domain-specific self-efficacy

First, STEM resilience is highly predictable from questionnaire-level information alone (AUC = 0.912; PR-AUC = 0.688), with calibration good enough (slope = 0.996) that predicted probabilities can be read at face value, a property rarely reported in the ILSA-ML literature (F. Wang et al., 2026). The largest single contributor was mathematics self-efficacy, ahead of every family, school, and contextual measure, and it ranked first under every outcome definition and across all ten plausible values. This is the pattern social-cognitive theory anticipates for an adaptive system operating under adversity (Bandura, 1997; Masten, 2001), and it extends motivational findings from general-population studies (Eccles & Wigfield, 2020; Liou et al., 2024; Schwerter et al., 2025) to the specific population where the equity stakes are highest. The dependence structure adds practical nuance: the self-efficacy gradient steepens in its upper range, suggesting that consolidating already-moderate competence beliefs may accompany the largest differences in resilience odds. Alongside it, mathematics anxiety and paid work carried negative associations, anxiety's role echoing gendered affect findings in STEM domains (Rozgonjuk et al., 2024), while cognitive activation and opportunity-to-learn measures carried positive ones. The top of the hierarchy is thus disproportionately psychological and instructional rather than structural. Two caveats discipline this reading. Self-efficacy at age 15 partly encodes prior mastery experience, so its association with concurrent proficiency is in part constitutive rather than antecedent (Bandura, 1997); and the model does not depend on it: refitted without both self-efficacy scales, discrimination fell only to 0.903, with socio-economic depth, ICT resources, and stress resistance heading the reshuffled hierarchy. Two associations warn against causal readings: family support for self-directed learning tracked *lower* predicted resilience, most plausibly because support flows compensatorily toward struggling students, and sport frequency's negative association likely reflects time allocation rather than harm. These are predictive contributions, not effects; their value is to focus scarcer causal designs (interventions, longitudinal studies) on the most promising candidates.

Second, the socio-economic gradient does not vanish within the disadvantaged quartile: within-quartile ESCS depth and home ICT resources ranked second and third. Even among the bottom quarter, relative resources matter, consistent with resource-based accounts of the digital divide's role in STEM learning (X. Hu et al., 2018; Kastorff et al., 2025). But the baseline decomposition and the definitional sensitivity analysis jointly qualify what such rankings mean: a system-base-rate-only model already reached AUC = 0.801 against the pooled model's 0.912, and ESCS and ICT resources receded under the fully national-relative outcome definition. These two factors therefore operate substantially as between-system markers: they help predict clearance of a fixed international bar largely because they index where a student lives and learns. Pooled "key determinants" reported without such decompositions risk mistaking geography for psychology.

Protective-factor hierarchies are only weakly universal

Third, the transport experiments answer this study's central theoretical question. A globally trained model retained useful but clearly degraded discrimination in unseen systems (median AUC 0.847, IQR 0.826–0.873, range 0.598–0.993), and system-specific protective-factor hierarchies correlated only modestly with the global one (median $\tau = .30$; 1.5 of 5 top factors shared on average). Resilience processes, as Ungar (2008, 2011) argued from qualitative and cross-cultural evidence, are contextually negotiated: factors that distinguish resilient students in one system need not do so in another. This study's contribution is to quantify this contingency at scale (the divergence exceeds an explicit sampling-noise benchmark) and to identify its limit, the near-universal top ranking of mathematics self-efficacy. For research practice, the implication is direct: pooled SHAP rankings from multi-country ILSA models (e.g., Cheung et al., 2024; J. Liu et al., 2026) should be read as descriptions of the pooled sample, not as universal determinants, unless out-of-system validation is demonstrated. For theory, the finding invites system-level moderation questions (which properties of curricula, stratification, or labor markets reorder protective factors) that the study's design surfaces but cannot answer.

Accurate on average, uneven at the margins

Fourth, the fairness audit (to date the first of any ILSA-based classifier) found high and similar subgroup AUCs and near-exact subgroup calibration, yet materially uneven error rates at a single operating threshold: resilient girls were missed 6 percentage points more often than resilient boys, and resilient rural students 12 points more often than urban ones. The attribute-blind re-analysis revealed that these two gaps have different anatomies. The gender gap was largely attribute-driven (the model had learned to use gender itself against girls' predictions), and removing protected attributes eliminated the gap at trivial accuracy cost, a case where the answer to Yu et al.'s (2021) question (should prediction models include protected attributes?) is clearly no. The rural gap persisted under blinding: it is carried by base rates and correlated covariates, the textbook consequence of one threshold across groups with unequal base rates (Hardt et al., 2016), and remediable only by threshold policy (group-aware or base-rate-adjusted operating points, or abstention bands), not by feature hygiene. The distinction matters because the plausible policy uses of resilience models (flagging students for enrichment, targeting programs, monitoring equity) are threshold uses. A model that is accurate on average could still direct recognition and resources away from resilient students who are already least visible to their systems. Such audits are inexpensive to conduct and report, and should become standard in this literature, as they are becoming in institutional learning analytics (Baker & Hawn, 2022; Deho et al., 2022; Kizilcec & Lee, 2022).

Methodological contributions

The design-ablation results discipline both criticism and complacency about current practice. In this application, cluster leakage from random splits inflated AUC only trivially (+ 0.001) and unweighted estimation shifted it moderately (– 0.005, while changing the estimand); the choices that mattered most were the outcome definition (up to 0.076 AUC across definitions, with different top-5 factors), the validation frame (pooled 0.912 vs median transported 0.847), and the baseline against which accuracy is judged (geography alone: 0.801). The general lesson for ML with ILSA data is therefore not merely “use

the weights” (though one should, and this study shows it costs nothing) but that labels, validation logic, and baselines, the three choices least often justified in this literature, dominate the conclusions. This study also demonstrates a practical template others can adopt wholesale: PV-parallel pipelines with Rubin-style pooling (whose between-PV component proved reassuringly small, $\tau = .90$ rank stability, an empirical answer to how much PV shortcuts matter for ML interpretation); school-grouped and country-grouped cross-validation, with the grouped scheme validated against strict leave-one-system-out; senate weighting for equal-system pooled estimands; sampling-noise benchmarks for cross-group comparisons of importance rankings; cluster-bootstrap uncertainty for pooled metrics; and prevalence-matched, attribute-blind-contrasted fairness auditing.

Implications for policy and practice

Three implications follow; each tied to a specific result. First, for intervention design: the top ranking and cross-definition robustness of mathematics self-efficacy, together with anxiety’s negative and cognitive activation’s positive associations, make competence-belief-supportive mathematics instruction the strongest candidate this correlational evidence can nominate for intervention studies with disadvantaged students, consistent with experimental and longitudinal evidence on self-efficacy-building practice (Bandura, 1997; Eccles & Wigfield, 2020). Second, for system-level analytics: ministries importing “global determinants of resilience” should first check transportability against their own data (the per-system results, Supplementary Table S1, show where the global model is least informative) and should prefer locally validated or locally re-fitted models for any operational use. Third, for equitable use of prediction: any threshold-based use should be preceded by an error-parity audit with an attribute-blind contrast, because different gaps require different remedies: feature blinding closed the gender gap here, while the rural gap requires threshold policy; without such audits, resilience-recognition programs would under-serve girls and rural students and would thereby reproduce the inequity such programs aim to reduce.

Limitations

This study has several limitations. First, the data are cross-sectional; all reported relationships are predictive associations, and some (compensatory family support; sport time; self-efficacy’s mastery-experience component) almost certainly reflect reverse or third-variable processes. Second, predictors are self- or principal-reports subject to reference-group and response-style effects that may themselves vary across systems, potentially contributing to hierarchy divergence; and because mathematics was the 2022 major domain, affect measures are mathematics-specific: no science self-efficacy or science-anxiety scales were available, which may understate science-side psychological factors in the dual-domain outcome. Third, resilience thresholds are conventions; arbitrariness was mitigated with three alternative definitions (with converging conclusions about self-efficacy and diverging ones about resource variables), but other operationalizations exist (Rudd et al., 2021). Fourth, the public file constrained the design: Costa Rica lacked ESCS entirely, rotated questionnaire modules produced high missingness for some scales (addressed by native missing-value handling and a robustness re-analysis),

and 12 small systems lacked support for stable within-system SHAP estimation. Fifth, country-grouped transport estimates approximate strict leave-one-system-out validation (validated on 11 systems, $r = .84$); transported-performance heterogeneity is described, not explained, and per-system estimates for very-low-prevalence systems are unstable at the extremes. Sixth, hyperparameters were selected on one outer fold's training partition rather than by fully nested search; the untouched-fold check (0.913 vs 0.912) indicates negligible optimism, but nested tuning would be cleaner. Finally, findings concern 15-year-olds' proficiency-based resilience; other STEM outcomes (aspirations, persistence, attainment) may behave differently (Nitzan-Tamar & Kohen, 2022; Zhao & Perez-Felkner, 2022).

Directions for future research

Four extensions follow directly. Explaining transport heterogeneity with system-level features (tracking age, spending, stratification, cultural dimensions) in a multilevel meta-learning design; testing the self-efficacy finding in intervention and longitudinal designs targeted at disadvantaged students; extending the fairness audit to intersectional subgroups and to mitigation strategies (group-aware thresholds; distributionally robust training); and repeating the design across cycles (PISA 2025) for temporal external validation.

Conclusions

Among 143,573 disadvantaged adolescents in 79 education systems, STEM resilience (solid proficiency in both mathematics and science despite socio-economic disadvantage) was rare (12.5% population-weighted), highly system-dependent, and strongly predictable from student, family, and school factors, foremost among them mathematics self-efficacy, whose top ranking survived every definitional, methodological, and cross-system stress test imposed. Beyond that near-universal core, both predictive accuracy and protective-factor hierarchies transported only imperfectly across systems, and a single-threshold application of the model would have systematically under-recognized resilient girls (an attribute-driven gap, removable by blinding) and resilient rural students (a base-rate-driven gap, requiring threshold policy). Design-consistent estimation proved practically costless, whereas outcome definitions, validation logic, and baselines materially changed conclusions. Machine learning can serve equity in STEM education, but only if models are validated in the systems where they will be used, audited for uneven subgroup errors, and interpreted as associational rather than causal evidence.

Abbreviations

AUC

area under the receiver operating characteristic curve

BRR

balanced repeated replication

CI

confidence interval

CV

cross-validation
ESCS
PISA index of economic, social and cultural status
ICT
information and communication technology
ILSA
international large-scale assessment
LOCO
leave-one-country-out
ML
machine learning
OECD
Organisation for Economic Co-operation and Development
PISA
Programme for International Student Assessment
IQR
interquartile range
PR-AUC
area under the precision–recall curve
PV
plausible value
RQ
research question
SD
standard deviation
SE
standard error
SHAP
Shapley additive explanations
STEM
science, technology, engineering, and mathematics
TPR/FPR
true-/false-positive rate.

Declarations

Ethics

The study analyzed publicly available, fully anonymized secondary data; no additional ethical approval or participant consent was required.

Data availability

The datasets analysed during the current study are publicly available in the OECD PISA 2022 database.

Code availability

The analysis code is available from the author upon request.

Competing interests

The author declares that he has no competing interests.

Funding

This research received no external funding.

Author Contribution

O.A. conceived the study, conducted the analysis, interpreted the results, and wrote and approved the final manuscript.

References

1. Agasisti T, Avvisati F, Borgonovi F, Longobardi S (2018) Academic resilience: What schools and countries do to help disadvantaged students succeed in PISA. OECD Education Working Papers No. 167. OECD Publishing, Paris <https://doi.org/10.1787/e22490ac-en>
2. Agasisti T, Avvisati F, Borgonovi F, Longobardi S (2021) What school factors are associated with the success of socio-economically disadvantaged students? An empirical investigation using PISA data. *Social Indicators Research* 157:749–781 <https://doi.org/10.1007/s11205-021-02668-w>
3. Baker RS, Hawn A (2022) Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32:1052–1092 <https://doi.org/10.1007/s40593-021-00285-9>
4. Bandura A (1997) *Self-efficacy: the exercise of control*. W. H. Freeman, New York
5. Bernardo ABI, Cordel MO II, Lucas RIG, Teves JMM, Yap SA, Chua UC (2021) Using machine learning approaches to explore non-cognitive variables influencing reading proficiency in English among Filipino learners. *Education Sciences* 11(10):628 <https://doi.org/10.3390/educsci11100628>
6. Bronfenbrenner U, Morris PA (2006) The bioecological model of human development. In: Lerner RM, Damon W (eds) *Handbook of child psychology: Vol. 1. Theoretical models of human development*, 6th edn. Wiley, Hoboken, pp 793–828
7. Caspi A, Gorsky P (2024) STEM career expectations across four diverse countries: Motivation to learn mathematics mediates the effects of gender and math classroom environments. *International Journal of STEM Education* 11:52 <https://doi.org/10.1186/s40594-024-00511-5>

8. Cheung K-c, Sit P-s, Zheng J-q, Lam C-c, Mak S-k, leong M-k (2024) A machine-learning model of academic resilience in the times of the COVID-19 pandemic: Evidence drawn from 79 countries/economies in the PISA 2022 mathematics study. *British Journal of Educational Psychology* 94(4):1224–1244 <https://doi.org/10.1111/bjep.12715>
9. Deho OB, Zhan C, Li J, Liu J, Liu L, Le TD (2022) How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology* 53(4):822–843 <https://doi.org/10.1111/bjet.13217>
10. Delprato M, Sandoval-Hernández A (2025) Academic resilience in the Latin America region post COVID-19 pandemic: An explainable machine learning analysis of its determinants and heterogeneity using alternative definitions. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2509.24830>
11. Eccles JS, Wigfield A (2020) From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology* 61:101859 <https://doi.org/10.1016/j.cedpsych.2020.101859>
12. Gabriel F, Signolet J, Westwell M (2018) A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *International Journal of Research & Method in Education* 41(3):306–327 <https://doi.org/10.1080/1743727X.2017.1301916>
13. Gamazo A, Martínez-Abad F (2020) An exploration of factors linked to academic performance in PISA 2018 through data mining techniques. *Frontiers in Psychology* 11:575167 <https://doi.org/10.3389/fpsyg.2020.575167>
14. García-Crespo FJ, Fernández-Alonso R, Muñoz J (2021) Academic resilience in European countries: The role of teachers, families, and student profiles. *PLOS ONE* 16(7):e0253409 <https://doi.org/10.1371/journal.pone.0253409>
15. Gardner J, Brooks C, Baker R (2019) Evaluating the fairness of predictive student models through slicing analysis. In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19)*. ACM, New York, pp 225–234 <https://doi.org/10.1145/3303772.3303791>
16. Gardner J, Yu R, Nguyen Q, Brooks C, Kizilcec R (2023) Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, New York, pp 1664–1684 <https://doi.org/10.1145/3593013.3594107>
17. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems* 29. Curran Associates, Red Hook, pp 3315–3323
18. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: Data mining, inference, and prediction*, 2nd edn. Springer, New York <https://doi.org/10.1007/978-0-387-84858-7>
19. Hu J, Peng Y, Ma H (2022) Examining the contextual factors of science effectiveness: A machine learning-based approach. *School Effectiveness and School Improvement* 33(1):21–50 <https://doi.org/10.1080/09243453.2021.1929346>

20. Hu X, Gong Y, Lai C, Leung FKS (2018) The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. *Computers & Education* 125:1–13 <https://doi.org/10.1016/j.compedu.2018.05.021>
21. Huang H, Tang X, Wang M-T (2026) Be curious and persistent: Joint roles of curiosity and persistence in adolescents' STEM achievements. *International Journal of STEM Education* 13:32 <https://doi.org/10.1186/s40594-026-00616-z>
22. Huang Y, Zhou Y, Chen J, Wu D (2024) Applying machine learning and SHAP method to identify key influences on middle-school students' mathematics literacy performance. *Journal of Intelligence* 12(10):93 <https://doi.org/10.3390/jintelligence12100093>
23. Immekus JC, Jeong T, Yoo JE (2022) Machine learning procedures for predictor variable selection for schoolwork-related anxiety: Evidence from PISA 2015 mathematics, reading, and science assessments. *Large-scale Assessments in Education* 10:30 <https://doi.org/10.1186/s40536-022-00150-8>
24. Jin S, Fang G, Cheung K-c, Sit P-s (2022) Factors associated with academic resilience in disadvantaged students: An analysis based on the PISA 2015 B-S-J-G (China) sample. *Frontiers in Psychology* 13:846466 <https://doi.org/10.3389/fpsyg.2022.846466>
25. Kastorff T, Müller M, Greiff S (2025) Digital media use and availability can mitigate the digital divide: Meta-analytic insights from PISA 2022. *Computers & Education* 238:105409 <https://doi.org/10.1016/j.compedu.2025.105409>
26. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems* 30. Curran Associates, Red Hook, pp 3146–3154
27. Kizilcec RF, Lee H (2022) Algorithmic fairness in education. In: Holmes W, Porayska-Pomsta K (eds) *The ethics of artificial intelligence in education*. Routledge, London, pp 174–202 <https://doi.org/10.4324/9780429329067-10>
28. Lezhnina O, Kismihók G (2022) Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA. *International Journal of Research & Method in Education* 45(2):180–199 <https://doi.org/10.1080/1743727X.2021.1963226>
29. Liou P-Y, Jang J, Myoung E (2024) Synergistic effects of students' mathematics and science motivational beliefs on achievement, and their determinants. *International Journal of STEM Education* 11:50 <https://doi.org/10.1186/s40594-024-00509-z>
30. Liu H, Chen X, Liu X (2022) Factors influencing secondary school students' reading literacy: An analysis based on XGBoost and SHAP methods. *Frontiers in Psychology* 13:948612 <https://doi.org/10.3389/fpsyg.2022.948612>
31. Liu J, Tahri D, Aziku M (2026) Predicting STEM competencies with machine learning: Identifying key determinants among 522,802 adolescents. *International Journal of STEM Education* 13:1 <https://doi.org/10.1186/s40594-025-00590-y>

32. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1):56–67 <https://doi.org/10.1038/s42256-019-0138-9>
33. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* 30. Curran Associates, Red Hook, pp 4765–4774
34. Masci C, Johnes G, Agasisti T (2018) Student and school performance across countries: A machine learning approach. *European Journal of Operational Research* 269(3):1072–1085 <https://doi.org/10.1016/j.ejor.2018.02.031>
35. Masten AS (2001) Ordinary magic: Resilience processes in development. *American Psychologist* 56(3):227–238 <https://doi.org/10.1037/0003-066X.56.3.227>
36. Nitzan-Tamar O, Kohen Z (2022) Secondary school mathematics and entrance into the STEM professions: A longitudinal study. *International Journal of STEM Education* 9:63 <https://doi.org/10.1186/s40594-022-00381-9>
37. Ocumpaugh J, Baker R, Gowda S, Heffernan N, Heffernan C (2014) Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology* 45(3):487–501 <https://doi.org/10.1111/bjet.12156>
38. OECD (2009) PISA data analysis manual: SPSS, 2nd edn. OECD Publishing, Paris <https://doi.org/10.1787/9789264056275-en>
39. OECD (2011) Against the odds: disadvantaged students who succeed in school. OECD Publishing, Paris <https://doi.org/10.1787/9789264090873-en>
40. OECD (2023a) PISA 2022 database. OECD, Paris. <https://www.oecd.org/en/data/datasets/pisa-2022-database.html>
41. OECD (2023b) PISA 2022 results (volume I): the state of learning and equity in education. OECD Publishing, Paris <https://doi.org/10.1787/53f23881-en>
42. OECD (2023c) PISA 2022 results (volume II): learning during – and from – disruption. OECD Publishing, Paris <https://doi.org/10.1787/a97db61c-en>
43. OECD (2024) PISA 2022 technical report. OECD Publishing, Paris <https://doi.org/10.1787/01820d6d-en>
44. Opoku RA, Pei B, Xing W (2025) Unveiling accuracy–fairness trade-offs: Investigating machine learning models in student performance prediction. *Journal of Learning Analytics* 12(2):125–139 <https://doi.org/10.18608/jla.2025.8543>
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830
46. Rozgonjuk D, Täht K, Soobard R, Teppo M, Rannikmäe M (2024) The S in STEM: Gender differences in science anxiety and its relations with science test performance-related variables. *International Journal of STEM Education* 11:45 <https://doi.org/10.1186/s40594-024-00504-4>

47. Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
<https://doi.org/10.1002/9780470316696>
48. Rudd G, Meissel K, Meyer F (2021) Measuring academic resilience in quantitative research: A systematic review of the literature. *Educational Research Review* 34:100402
<https://doi.org/10.1016/j.edurev.2021.100402>
49. Rutkowski L, Gonzalez E, Joncas M, von Davier M (2010) International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher* 39(2):142–151
<https://doi.org/10.3102/0013189X10363170>
50. Sandoval-Hernández A, Białowolski P (2016) Factors and conditions promoting academic resilience: A TIMSS-based analysis of five Asian education systems. *Asia Pacific Education Review* 17:511–520 <https://doi.org/10.1007/s12564-016-9447-4>
51. Schwerter J, Lauermann F, Doeblner P, Fokkema M (2025) Putting the pieces of the puzzle together in modeling gendered educational choices. *International Journal of STEM Education* 12:38
<https://doi.org/10.1186/s40594-025-00558-y>
52. Sevilla MP, Luengo-Aravena D, Farías M (2023) Gender gap in STEM pathways: The role of secondary curricula in a highly differentiated school system, the case of Chile. *International Journal of STEM Education* 10:58 <https://doi.org/10.1186/s40594-023-00450-7>
53. Steyerberg EW, Harrell FE Jr (2016) Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology* 69:245–247
<https://doi.org/10.1016/j.jclinepi.2015.04.005>
54. Ungar M (2008) Resilience across cultures. *British Journal of Social Work* 38(2):218–235
<https://doi.org/10.1093/bjsw/bcl343>
55. Ungar M (2011) The social ecology of resilience: Addressing contextual and cultural ambiguity of a nascent construct. *American Journal of Orthopsychiatry* 81(1):1–17 <https://doi.org/10.1111/j.1939-0025.2010.01067.x>
56. von Davier M, Gonzalez E, Mislevy RJ (2009) What are plausible values and why are they useful? In: IERI monograph series: issues and methodologies in large-scale assessments, vol 2. IEA-ETS Research Institute, Hamburg, pp 9–36
57. Vooren M, Haelermans C, Groot W, Maassen van den Brink H (2022) Comparing success of female students to their male counterparts in the STEM fields: An empirical analysis from enrollment until graduation using longitudinal register data. *International Journal of STEM Education* 9:1
<https://doi.org/10.1186/s40594-021-00318-8>
58. Wang F, Li Z, King RB (2026) The application of machine learning approaches in international large-scale educational assessments: A scoping review. *Studies in Educational Evaluation* 89:101608
<https://doi.org/10.1016/j.stueduc.2026.101608>
59. Wang X, Perry LB, Malpique A, Ide T (2023) Factors predicting mathematics achievement in PISA: A systematic review. *Large-scale Assessments in Education* 11:24 <https://doi.org/10.1186/s40536-023-00174-8>

60. Ye W, Strietholt R, Blömeke S (2021) Academic resilience: Underlying norms and validity of definitions. *Educational Assessment, Evaluation and Accountability* 33:169–202
<https://doi.org/10.1007/s11092-020-09351-7>
61. Yoo JE (2018) TIMSS 2011 student and teacher predictors for mathematics achievement explored and identified via elastic net. *Frontiers in Psychology* 9:317
<https://doi.org/10.3389/fpsyg.2018.00317>
62. Yu R, Lee H, Kizilcec RF (2021) Should college dropout prediction models include protected attributes? In: *Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S '21)*. ACM, New York, pp 91–100 <https://doi.org/10.1145/3430895.3460139>
63. Zhao T, Perez-Felkner L (2022) Perceived abilities or academic interests? Longitudinal high school science and mathematics effects on postsecondary STEM outcomes by gender and race. *International Journal of STEM Education* 9:42 <https://doi.org/10.1186/s40594-022-00356-w>
64. Zheng J-q, Cheung K-c, Sit P-s (2024) Identifying key features of resilient students in digital reading: Insights from a machine learning approach. *Education and Information Technologies* 29(2):2277–2301 <https://doi.org/10.1007/s10639-023-11908-0>

Figures

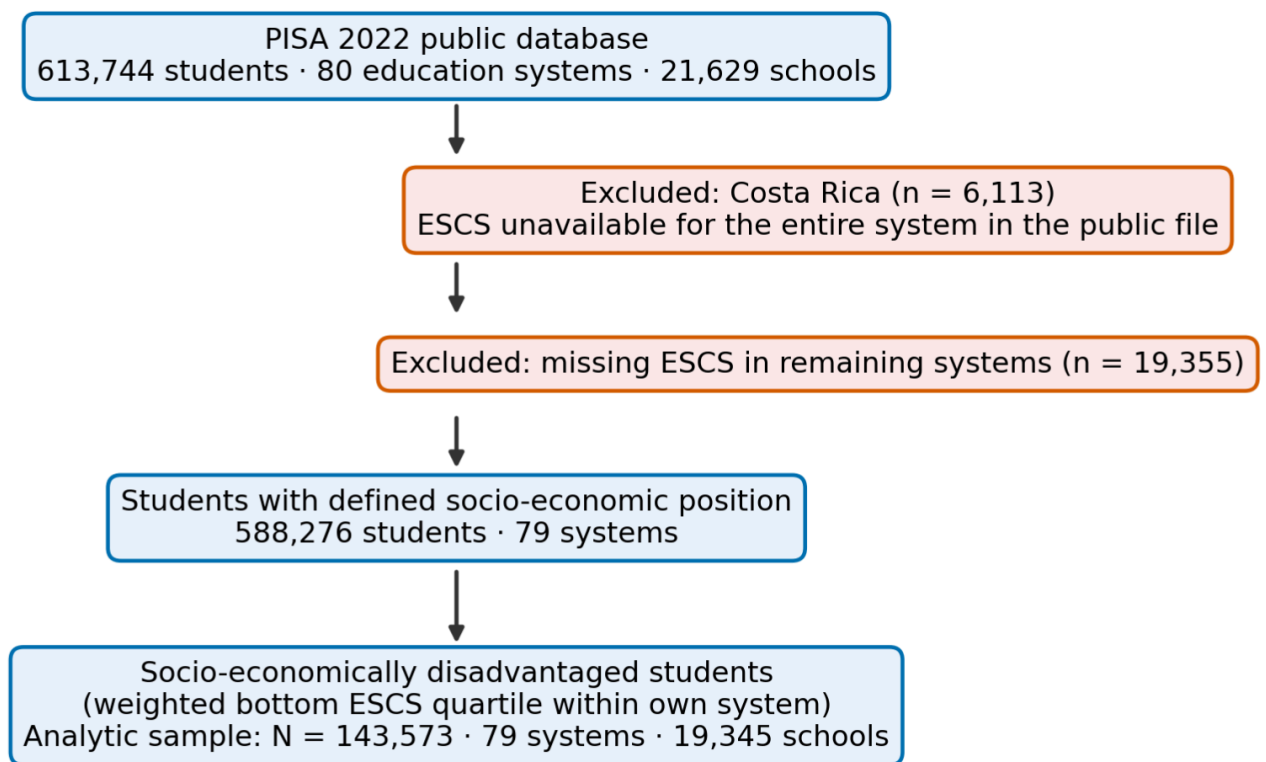


Figure 1

Construction of the analytic sample from the PISA 2022 public database.

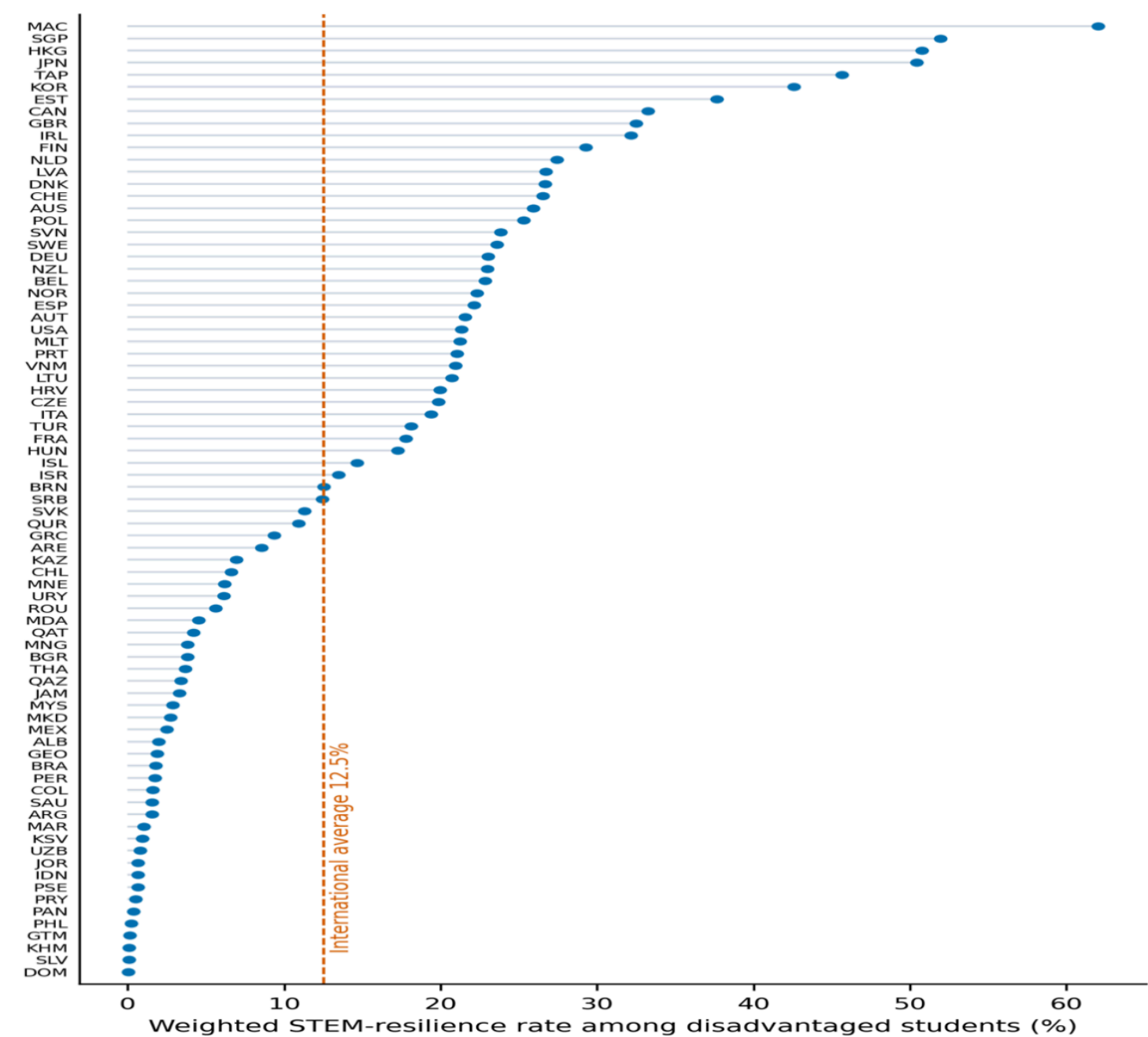


Figure 2

Weighted STEM-resilience rates among disadvantaged students in 79 education systems. Rates are PV-averaged and computed with final student weights; dashed lines mark the population-weighted international rate (12.5%) and the equal-system average (15.2%)

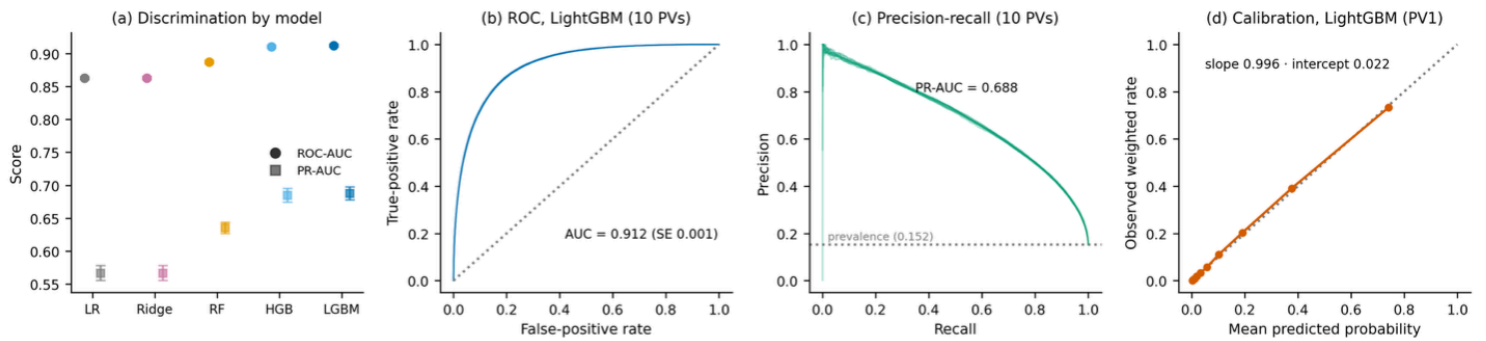


Figure 3

Primary model performance: (a) ROC-AUC and PR-AUC across classifiers; (b–c) LightGBM ROC and precision–recall curves across plausible values; and (d) calibration of out-of-fold predictions.

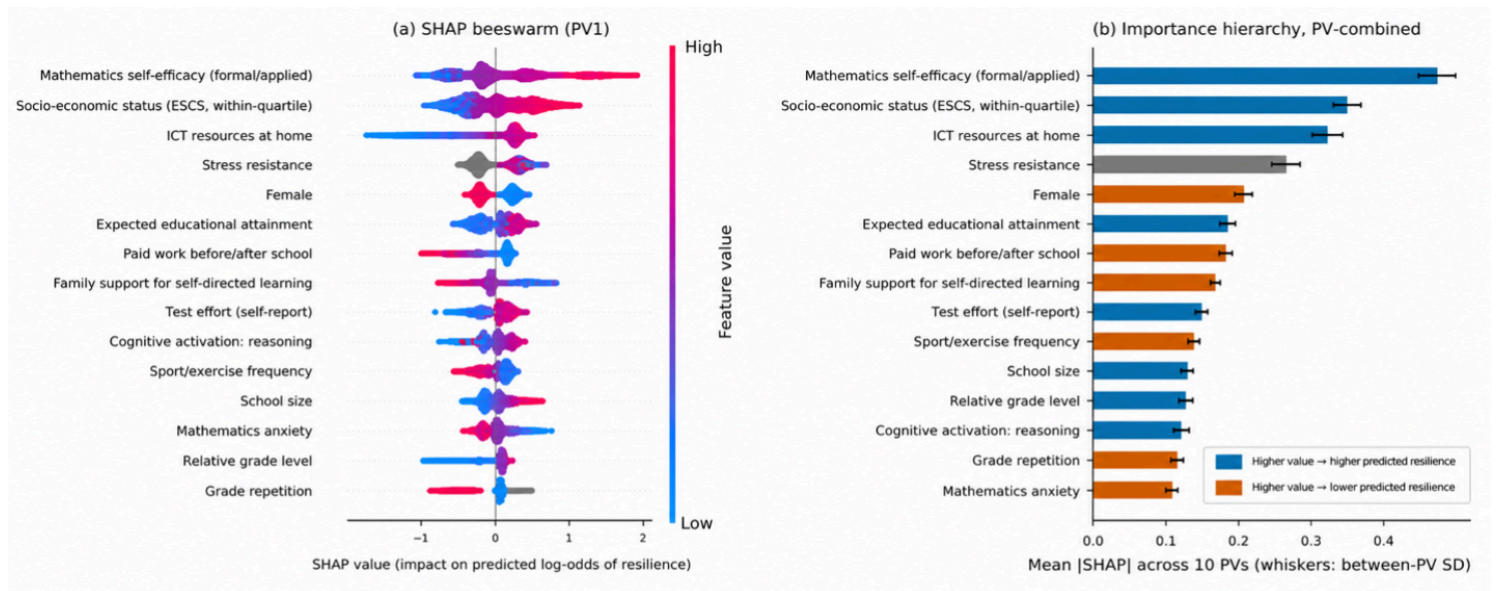


Figure 4

Protective and risk factors: (a) SHAP beeswarm for PV1; and (b) mean absolute SHAP values across all plausible values, with colors indicating association direction.

Dependence of predicted STEM resilience on selected factors (PV1)

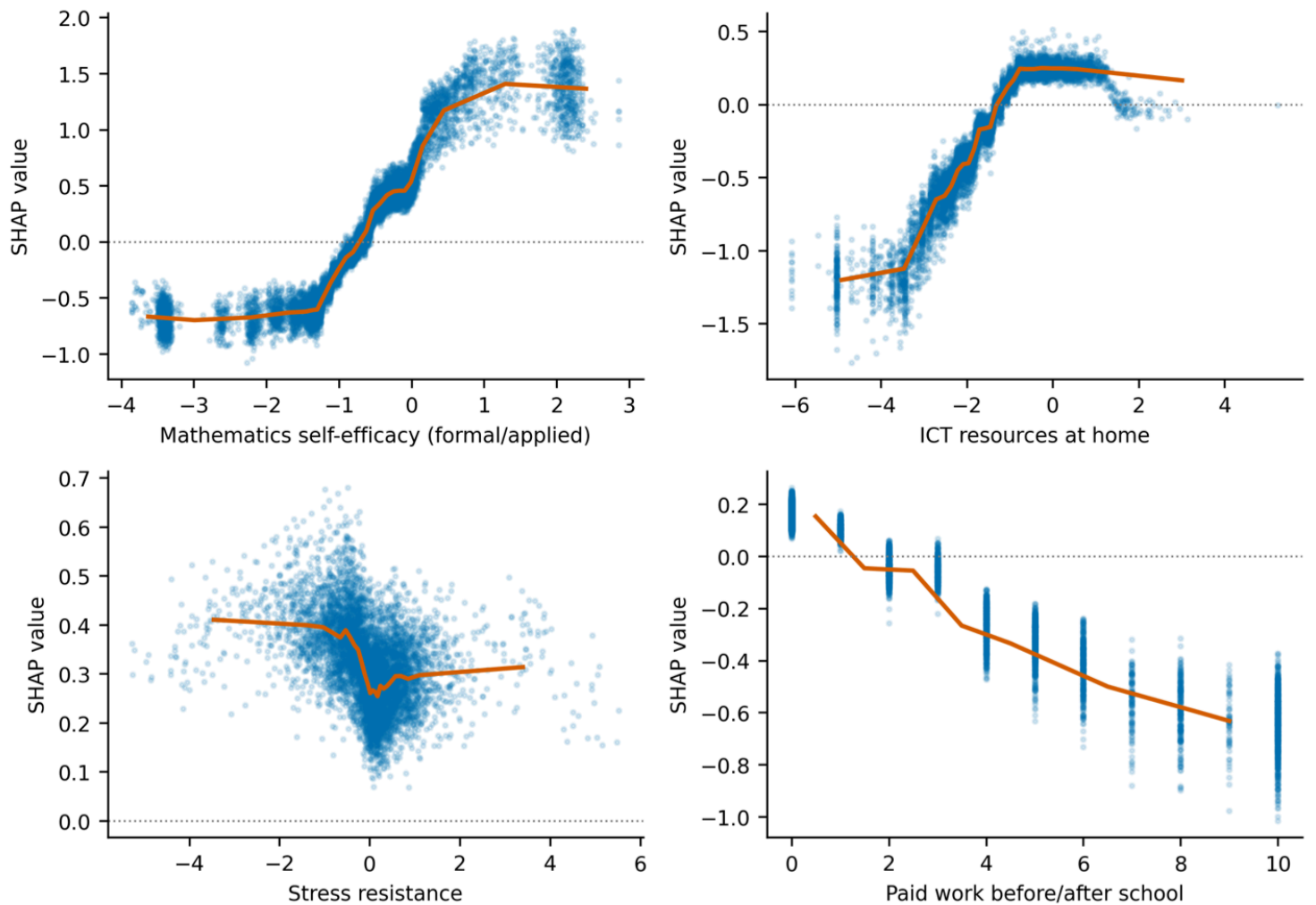


Figure 5

SHAP dependence plots for four leading predictors. Panels show PV1 SHAP values against mathematics self-efficacy, home ICT resources, stress resistance, and paid work; the line is a 25-bin mean.

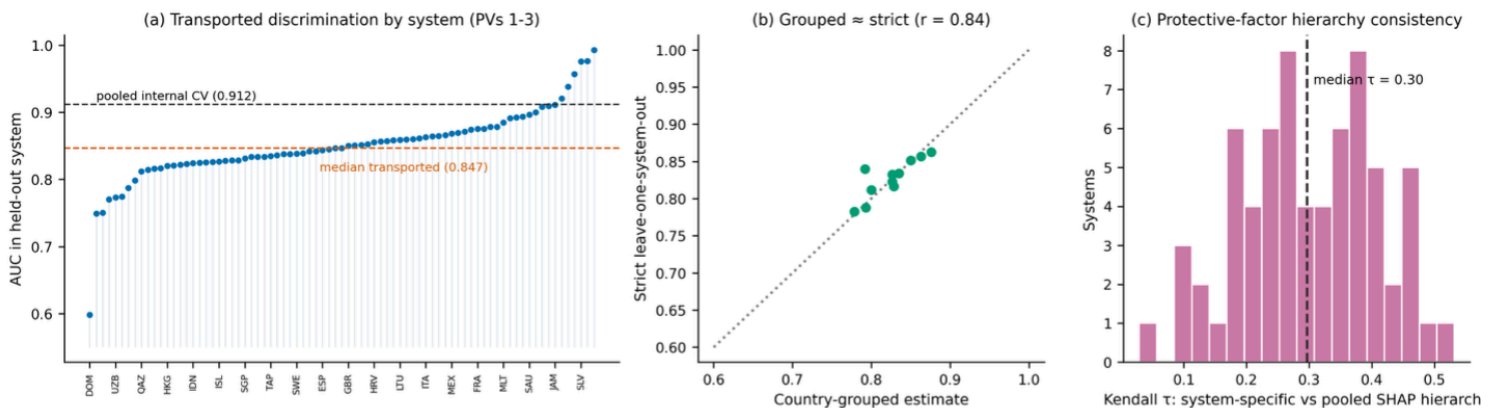


Figure 6

Transportability: (a) held-out-system AUC; (b) agreement between grouped and leave-one-system-out estimates; and (c) similarity between system-specific and pooled SHAP rankings.

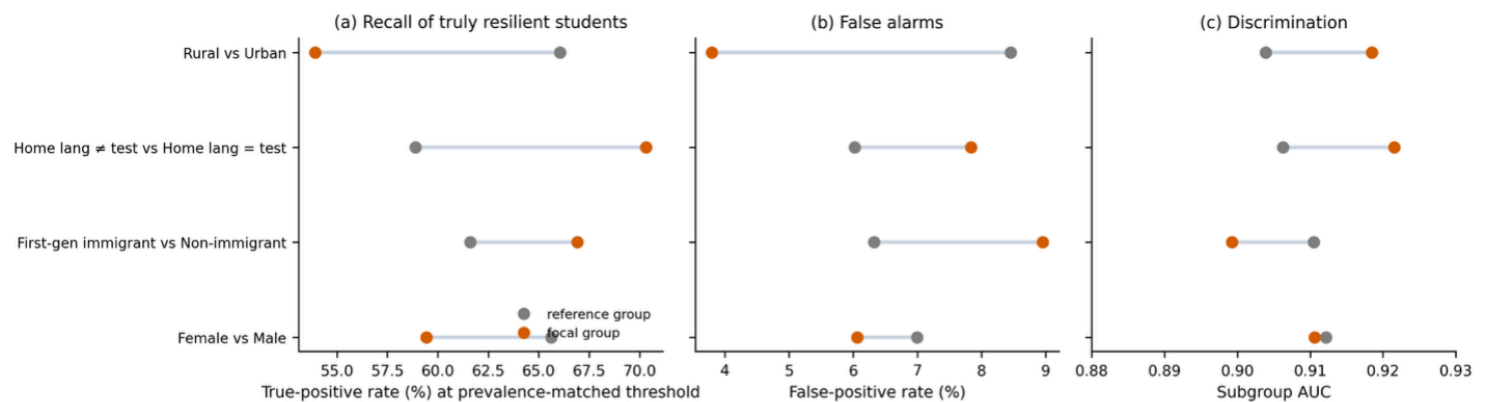


Figure 7

Fairness audit at the prevalence-matched threshold. Rubin-pooled over ten PVs: (a) true-positive and (b) false-positive rates and (c) subgroup AUC, for gender, immigrant background, home language, and school location.

Supplementary Files

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

- [Additionalfile1Supplementarymaterials.docx](#)
- [Supplementarymaterials.docx](#)