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.

From Assessment to Action: An Evidence-To-Reform Translation Framework applied to PISA 2022 data from Uzbekistan

Indira Khadjieva

i.khadjieva@wiut.uz

Westminster International University in Tashkent

Article

Keywords: evidence-informed policymaking, education reform, PISA 2022, hierarchical linear modelling, quantile regression, Uzbekistan, policy translation framework

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10503627/v1

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

Additional Declarations: No competing interests reported.

From Assessment to Action: An Evidence-To-Reform Translation Framework applied to PISA 2022 data from Uzbekistan

Indira Khadjieva
School of Business and Economics,
Westminster International University in Tashkent
Tashkent, Uzbekistan
Corresponding author: Indira Khadjieva, i.khadjieva@wiut.uz

Abstract

International large-scale learning assessments have generated extensive evidence on the determinants of student achievement, yet a persistent gap remains between the production of this evidence and its systematic translation into education reform. Governments possess extensive statistical findings on teacher quality, school climate, and socioeconomic inequalities but lack structured methodologies for converting these findings into prioritised, implementable, and continuously monitored reform strategies. This study addresses that gap by proposing the Evidence-to-Reform Translation Framework (ERTF) – a decision-support methodology that connects empirical evidence to policy appraisal, reform design, implementation monitoring, and adaptive learning within a single analytical process.

The framework is demonstrated empirically using PISA 2022 data from Uzbekistan – the country's first participation, combining two-level hierarchical linear modelling and simultaneous quantile regression applied to an analytical sample of 4,851 students from 196 schools. The null model yields an intraclass correlation coefficient of 0.229, confirming that school-level conditions account for a substantial share of mathematics achievement variation. Inadequately qualified teaching staff emerges as the operative school-level constraint consistently and significantly associated with lower performance in eight of ten plausible values (β= -11.42, $p < 0.05$), and the penalty is roughly four times larger in urban schools, where it is robust in all ten. Formal teacher certification shows no association with achievement in any specification, indicating that credential-holding and instructional competence are distinct.

Quantile regression reveals that teacher support and school belonging have their largest positive associations with performance among the weakest students, while the socioeconomic gradient steepens toward the top of the distribution: school climate raises the floor, family background the ceiling. Rural underperformance proves entirely compositional, disappearing once school-level mean socioeconomic composition is controlled.

The ERTF appraisal process converts these findings into four reform priorities aligned with Uzbekistan's existing Presidential Decree Framework, a nine-indicator monitoring architecture with PISA 2022 baselines and 2030 targets, and two built-in PISA cycle evaluation checkpoints.

Although demonstrated using the Uzbekistan case, the framework's analytical logic is designed to be applicable across diverse education systems and assessment programs. The study contributes to educational effectiveness research, evidence-informed governance, and implementation science by providing a transparent and transferable methodology for converting assessment evidence into sustainable education reform.

Keywords: evidence-informed policymaking; education reform; PISA 2022; hierarchical linear modelling; quantile regression; Uzbekistan; policy translation framework

Introduction

Governments have long recognized that education is not simply a social good but an economic necessity. The connection between what students learn in classroom and what they can contribute to the labor market and civic life is well established, and it has given education reform a place near the top of most national policy agendas. Policymakers now have access to detailed, internationally comparable evidence on what students know and can do, generated through large-scale assessment programs such as the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress International Reading Literacy Study (PIRLS). These assessments have shifted the conversation in education policy from anecdote to data – and from national self-assessment to international benchmarking.

Yet this shift has produced a paradox. Governments now know more about the state of their education systems than at any previous point in history, but the path from knowing to acting remains poorly mapped. PISA results can tell a minister of education that a large share of students is below the minimum proficiency threshold, that teacher qualifications are low relative to comparable systems, and that school climate is associated with performance gaps. What PISA cannot tell that minister in which of these problems to tackle first, how much it will cost, which institutions should lead, and how progress should be measured two years from now. The statistical evidence stops at the door of the reform office.

This gap between evidence and action is not new, it has been documented across policy domains from health to criminal justice (Nutley, Walter, and Davies, 2007; Head, 2010; Cairney, 2016). In education, the problem is particularly acute because the evidence is plentiful and the institutional capacity to act on it varies across countries. High-income OECD systems have decades of experience using PISA data to adjust curriculum frameworks, restructure teacher training, and redesign accountability systems. Lower-middle-income countries participating in PISA for the first time – as Uzbekistan did in 2022 – face a steeper challenge: not only must they interpret unfamiliar data, but they must do so within institutional structures, budget constraints, and governance arrangements that statistical models do not account for.

This situation is increasingly common. Ten low- and middle-income countries including Uzbekistan, Mongolia, El Salvador, Jamaica, Palestine, Kenya, Rwanda, Zambia, Armenia and Ecuador have joined the PISA assessment for the first time since 2020, and each now receives detailed diagnostic evidence about a system it has limited institutional experience of reforming against international benchmarks (OECD, 2025). Each confronts the same difficulty: an abundance of findings and no structured method for deciding which to act on first. Uzbekistan's first cycle offers a case in which that difficulty can be examined directly, but the difficulty itself is general.

Uzbekistan's PISA 2022 results make the challenge concrete. A mean mathematics score of 364.5 points against an OECD average of 472, with 80.3 percent of students below minimum Level 2 proficiency threshold, placed the country 71st out of 81 participating systems (OECD, 2023a). The government had already committed, under Presidential Decree No. DP-5712 (2019) and reaffirmed in Uzbekistan-2030 Strategy (DP-21, 2026), to reaching the top 30 PISA systems by 2030 – a target that implies roughly doubling the share of students at or above Level 2 within eight years. The empirical question is not whether reform is needed but where it should begin and how it should be structured.

This paper responds to that question by proposing the Evidence-to-Reform Translation Framework (ERTF), a structure approach for converting large-scale assessment findings into implementable education reform strategies. The framework is demonstrated empirically using PISA 2022 data from Uzbekistan – a nationally representative sample of 7,293 students in 202 secondary schools - with two-level hierarchical linear modelling and quantile regression applied to an analytical sample of 4,851 students in 196 schools. The empirical analysis identifies the principal school-level and student-level determinants of mathematics achievement. The ERTF then connects those findings to policy prioritization, reform design, implementation planning, and measurable monitoring, treating the statistical results not as conclusions but as the starting point of a structured governance process.

Three research questions guide the analysis. First, how can large-scale assessment evidence be translated systematically into implementable reform strategies? Second, when multiple determinants of student achievement are identified simultaneously, which should receive priority and on what grounds? Third, how can evidence-based priorities be expressed as measurable indicators that allow reform progress to be tracked over time?

Accordingly, the manuscript adds three contributions. It provides the first combined HLM and quantile regression analysis with compositional controls of PISA 2022 data for Uzbekistan. It introduces an explicit evidence appraisal process that evaluates findings against criteria of statistical robustness, educational impact, equity relevance, and implementation feasibility. Therefore, it connects this appraisal to a monitoring architecture that gives reform targets a measurable form. While the empirical demonstration is specific to Uzbekistan, the framework's logic applies wherever governments face the task of converting assessment evidence into reform action.

The paper proceeds as follows. Section 2 reviews the literature on large-scale assessments, evidence-informed policymaking, and policy translation. Section 3 introduces the ERTF. Section 4 describes the data and econometric methods. Section 5 presents empirical results organized around reform priorities. Section 6 demonstrates the operationalization of findings through the framework. Section 7 discusses implications for evidence-informed education governance. Section 8 concludes.

2. Literature Review

2.1. International learning assessments as instruments of education governance

Over the past three decades, the relationship between educational data and policymaking has changed considerably. Governments that once relied on administrative statistics, inspection reports, and national examination results now have access to internationally comparable evidence generated through large-scale assessments, most prominently the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). What matters is less the volume of data than the use to which it is put. These assessments have moved beyond comparative league tables to become governance instruments: tools through which education ministries diagnose systems problems, benchmark their systems against peers, and justify reform priorities to legislatures and the public alike (OECD, 2023a; UNESCO, 2023).

PISA in particular has reshaped national reform agendas in ways that were difficult to anticipate when the program launched in 2000. The publication of results has triggered policy responses in countries as different as Germany, Brazil, and Kazakhstan – revisions to curriculum frameworks, restructuring of teacher education, new accountability mechanisms, and shifts in school funding

formulae. A systematic review of this influence confirms that PISA participation has driven tangible changes in national policy across diverse institutional contexts, though the nature and speed of reform varies considerably depending on political priorities, governance capacity, and existing institutional arrangements (Liu et al. 2025).

Availability of good evidence does not, however, guarantee good policy. Volante and Mattei (2024) validates that governments frequently use assessment evidence selectively, emphasizing findings that align with existing priorities and setting aside those that would require more difficult institutional change. Jerrim (2017) cautions that the statistical associations identified through international assessments should not be treated as direct causal relationships without careful attention to methodology and context. These observations point to a fundamental tension: the production of assessment evidence has expanded rapidly, while the institutional mechanism for translating that evidence into reform have developed more slowly.

2.2 The evidence-to-policy gap and knowledge translation

The gap between generating evidence and using it effectively is not unique to education. Across health, environmental policy, and social protection, researchers have documented the same problem, governments possess extensive empirical knowledge but struggle to convert it into coherent, implementable reform strategies (Nutley, Walter, and Davies, 2007; Head, 2010; Cairney, 2016). The reasons are structural. Policy decisions are shaped not only by evidence but political priorities, institutional constraints, stakeholder interests, budget cycles, and administrative capacity. Evidence informs rather than determines policy and the translation between the two requires deliberate institutional effort.

In education, knowledge translation has received comparatively little attention despite the scale of the assessment evidence now available. Most PISA-based studies conclude with statistical findings and broad recommendations on improving teacher quality, reducing socioeconomic gaps, strengthening school climate without specifying how these recommendations should be prioritised, sequenced, resourced, and monitored. Stewart et al (2022) observe that the value of evidence lies not on its production but in its capacity to support context-sensitive, adaptive decision-making, yet the mechanisms for achieving this remain underdeveloped in the education literature.

Implementation science extends this argument further. Reforms fail not only when evidence is weak but when implementation processes are poorly coordinated, inadequately monitored, or insufficiently adapted to local conditions (Vanlommel & van den Boom-Muilenburg, 2024). This

is particularly relevant in education, where national policies must be translated across regional administrations, school leadership, and classroom practice, each introducing the possibility of distortion or dilution. Monitoring and evaluation are therefore not optional agenda to reform design but integral components of evidence-informed governance (Head, 2010; Owen et al., 2022).

2.3 From determinants of achievement to policy priorities

The empirical literature on the determinants of student achievement provides the substantive foundation on which any evidence-to-reform framework must rest. Five findings from this literature are directly consequential for the present study.

Teacher quality consistently emerges as the most influential school-level predictor of student achievement across PISA, TIMSS, and PIRLS studies (Rivkin, Hanushek, and Kain, 2005; Niu, Xu, and Yu, 2025). The same literature also gives reason for caution about how teacher quality is measured. Rivkin, Hanushek, and Kain (2005) found that observable credentials like certification status, degree level, years of experience explain remarkably little of the variation in teacher effectiveness, a result that has held up well and that has direct implications for the choice of indicator in any monitoring system. Recent distributional analyses suggest, further, that good teaching generates its largest returns among low-performing students, making teacher quality simultaneously an effectiveness and an equity intervention (Shen, 2026; Liu et al., 2024).

School climate encompassing disciplinary order, teacher support, and student belonging independently predicts student achievement after controlling for socioeconomic background (Klieme, Pauli, and Reusser, 2009). Quantile regression evidence confirms that school climate effects are largest at the lower end of the performance distribution, which has direct implications for equity-targeted reform design (Pholphirul, Rukumnuaykit, and Teimtad, 2023).

Socioeconomic background is the strongest student-level predictor of achievement in most systems, but its magnitude varies significantly across contexts. In some systems including Uzbekistan, the SES gradient is unusually compressed, shifting analytical attention toward school-level institutional factors as the primary sources of performance variation (OECD, 2023b). Empirical findings from high-income systems cannot be transferred without verification.

Alongside individual background, the socioeconomic composition of a school's intake exerts an influence of its own, frequently exceeding the individual effect by a considerable margin (Thrupp et al. 2002). Where students are sorted unevenly across schools, raw performance comparisons confound institutional quality with intake, and schools serving disadvantaged populations appear

weaker than their contribution to learning warrants. This has direct implications for accountability design, since indicators built on unadjusted mean scores will systematically misrepresent such schools (Crosnoe, 2009; Willms, 2010).

Geographic disparities in teacher supply, school resources and student composition produce systematically different policy needs across locations (Zhang, Jin, and Fan, 2024; Pholphirul et al., 2023). Whether observed rural-urban gaps reflect differences in school effectiveness or differences in whom schools enroll is an empirical question, and one whose answer determines whether geographic targeting should address school capacity or student disadvantage.

Taken together, these findings establish what is known about the determinants of achievement, but they do not resolve the policy translation problem. Knowing that teacher quality matters does not tell a government whether to prioritize certification reform, pre-service curriculum redesign, salary incentives, or in-service professional development. Knowing that school climate predicts performance does not specify which climate dimensions to target, in which schools, or with what implementation timeline. The gap between empirical findings and operational reform decisions is precisely what the Evidence-to-Reform Translation Framework addresses.

3. The Evidence-to-Reform Translation Framework

3.1 Rationale

The core problem this framework addresses is straightforward: governments now produce more educational evidence than they know how to use. International assessments generate robust findings on teacher quality, school climate, and socioeconomic gradients, but the literature on how to convert those findings into reform decisions remains thin. Most studies conclude with broad recommendations and leave the translation work to policymakers who must navigate institutional constraints, budget cycles, and competing priorities without a structured methodology to guide them.

The Evidence-to-Reform Translation Framework (ERTF) responds to this gap by positioning empirical evidence not as the conclusion of educational research but as the starting point of a broader policy process. It draws on three bodies of scholarship that have developed largely in parallel educational effectiveness research, evidence-informed governance, and implementation science, and connects them into a single decision-support methodology. The framework is not a predictive model and does not estimate the determinants of achievement; it provides governments

with a structured approach for appraising, prioritizing, and operationalizing the findings that assessments generate.

3.2 Architecture of the ERTF

The framework consists of seven stages, illustrated in Figure. 1. These stages do not form linear sequence, they constitute a continuous learning cycle in which implementation outcomes feed back into new evidence generation and policy revision.

Stage 1 & 2– Evidence generation and Econometric Analysis. The cycle begins with the collection of empirical evidence from a large-scale assessment. In this study, that evidence comes from PISA 2022 Uzbekistan a national representative sample of 7,293 students from 202 schools, assessed using internationally standardized procedures. The econometric analysis in Stage 2 is performed on the analytical sample of 4,851 students in 196 schools described in Section 4.3. At this stage the objective is establishing a robust empirical foundation. Data quality, representativeness, methodological transparency are prerequisites for everything that follows, since no policy recommendation can be stronger that the evidence on which it tests.

Stage 3- Evidence appraisal. Not every statistically significant finding deserves equal policy attention and not every available variable can bear the weight of analysis in the first place. Appraisal therefore proceeds in two steps: a validity check, then evaluation against four criteria.

The validity check comes first because it is logically prior to the others. Before a finding is judged on its robustness or its magnitude, the measure that produced it must be shown to capture the construct it names. This is established by inspecting the distribution of the variable, comparing it against external benchmarks where these exist, and considering how respondents are likely to have understood the underlying question. A measure that fails this check yields findings that cannot enter the appraisal, however well-behaved they appear in a regression – a stable coefficient estimation on an invalid measure is still uninformative. The failure is not merely discarded, however. An indicator that cannot support analysis cannot support monitoring either, so a validity failure is itself a policy-relevant result, identifying a gap in the measurement infrastructure that reform must address before progress on the underlying construct can be tracked.

Findings that clear the validity check are then evaluated against four criteria. Statistical robustness asks whether a finding holds across alternative model specifications and datasets in this study, across ten plausible values and multiple robustness checks. Magnitude of effect asks whether the association is large enough to matter in practice, not merely in a regression table. Equity relevance

asks whether addressing this determinant would disproportionally benefit disadvantaged or low-performing students – a central concern in a system where 80.3 percent of students fall below minimum proficiency. Implementation feasibility asks whether policymakers can realistically influence this determinant within existing institutional and financial constraints. A finding that scores well on all four criteria becomes a reform priority; one that fails feasibility, however statistically significant, is acknowledged in the evidence base but not elevated to action.

Stage 4 – Policy prioritization. Appraised evidence is converted into an ordered set of reform priorities. Education reforms are implemented under conditions of resource scarcity, so, governments cannot address very significant determinant simultaneously. This stage asks which interventions offer the greatest combined return on the four appraisal criteria, and in what sequence they should be pursued. The output is not a ranked list of regression coefficients but a portfolio of reform priorities calibrated to the country's institutional capacity, financial constraints, and equity objectives.

Stage 5 – Reform design. Each priority is translated into operational reform strategies specifying what interventions are required, which institutions are responsible, how implementation should be sequenced, and what resources are needed. The ERTF does not prescribe unique policy solutions; empirical evidence constraints the choice of interventions but does not determine it. A government with strong central administrative capacity might centralize teacher certification through a national pathway; one with weaker central capacity might operate through regional education authorities. The framework is flexible on instrument design while being specific on the empirical basis for reform selection.

Stage 6 – Monitoring architecture. One of the most consistent weaknesses in education reform is the absence of measurable implementation frameworks. Reforms are announced, funded, and launched, and then evaluated only years later, if at all. The ERTF addresses this by requiring that each reform priority be accomplished by measurable indicators, baseline values drawn from the assessment data, annual targets, designated responsible institutions, specified data sources, reporting frequency, and evaluation criteria. This monitoring architecture is not an administrative formality; indeed, it is what makes adaptive policy learning possible. Without it Stage 7 cannot function.

Stage 7 – Adaptive policy learning. The final stage transforms monitoring results into policy revision. Implementation data generate new evidence about what is working, what is not, and why

that feeds back into Stage 3 for re-appraisal. This closes the cycle and prevents reform from becoming a static exercise. In practice, this means that the PISA 2025 results, expected in late 2026, will serve as the first empirical test of whether the reforms initiated in response to PISA 2022, are producing measurable gains.

3.3 Theoretical contribution

The ERTF contributes to three bodies of literature. Within educational effectiveness research, it extends the standard practice of identifying determinants of achievement by providing an explicit methodology for converting those findings into reform priorities. Where most PISA-based studied stop at statistical conclusions, the ERTF treats those conclusions as inputs to a governance process which is consistent with the iterative model of evidence-informed change proposed by Vanlommel and van den Boom-Muilenburg (2024).

Within evidence-informed policymaking, the framework advances the argument of Nutley, Walter, and Davies (2007) and Stewart et al (2022) that evidence utilization requires deliberate institutional mechanisms – not simply better evidence production. The four appraisal criteria provide one such mechanism: an explicit, reproducible decision procedure that reduces reliance on ad hoc expert judgement and increases transparency of how statistical findings are converted into policy priorities. The validity check strengthens this procedure in a way the empirical application made necessary. Frameworks for evidence use commonly assume that available indicators measure what their labels claim; the Uzbekistan case shows that this assumption can fail precisely for the variable a government would be most inclined to monitor, which is why validity screening is built in as a prior step rather than left implicit.

Within implementation science, the framework responds to the observation that many reforms fail not because the underlying evidence is wrong but because implementation is poorly monitored and insufficiently adapted to local conditions. By integrating monitoring architecture and adaptive learning as structural components. The ERTF positions evaluation as the engine of continuous improvement rather than the final chapter of a reform cycle.

Although demonstrated using PISA 2022 data from Uzbekistan, the framework's logic applies wherever governments face the task of converting assessment evidence into reform action. It is designed to be applicable to evidence generated through TIMSS, PIRLS, national learning

assessments, or any dataset that combines student achievement data with contextual information about schools, teachers, and learning environments.

4. Methodology

4.1 Research design

This study adopts a sequential explanatory design that combines advanced econometric analysis with structured policy translation. The research proceeds in two complementary stages. The first identifies the determinants of student achievement using data from PISA 2022 Uzbekistan, applying hierarchical linear modelling and quantile regression. The second translates the resulting empirical evidence into reform priorities through the Evidence-to-Reform Translation Framework (ERTF). The overall research design is summarized in Figure. 2.

What distinguishes this design from conventional PISA studies is its treatment of econometric analysis as the foundation for a policy rather than its conclusion. Statistically significant findings are not immediately converted into policy recommendation. They are first evaluated against four explicit criteria relating to statistical robustness, effect magnitude, equity relevance, and implementation feasibility, as set out in Section 3.2. Only findings that satisfy multiple criteria progress to policy prioritization. This integrated design allows the study to contribute simultaneously to educational effectiveness research and evidence-informed governance scholarship.

4.2 Data source

The empirical analysis draws on the PISA 2022 international database for Uzbekistan, publicly released by the OECD in December, 2023. PISA is a triennial cross-sectional assessment administered to nationally representative samples of 15-year-old students, measuring reading, mathematics, and science literacy across 81 participating nations. In 2022, the total international sample comprised approximately 690,000 students.

Uzbekistan participated in PISA for the first time in 2022. The national sample comprises 7,293 students from 202 schools, constructed to be nationally representative through a two-stage stratified cluster sampling design in which schools were selected as primary sampling units and students sampled within schools. The Uzbekistan dataset is extracted from the PISA 2022 international student file (STU22) and school file (SCH22) using the ISO country code UZB.

Mathematics is the focus of this study because it constitutes the major assessment domain in the 2022 cycle and is widely reorganized as a core indicator of educational quality and future economic competitiveness (Hanushek and Woessmann, 2015). The analysis draws on three data instruments: the cognitive assessment, which provides the dependent variable; the student background questionnaire, which provides student-level predictors; and the school questionnaire completed by principals, which provides school-level predictors.

4.3 Sample and missing data

The analytical sample is restricted to students with complete data on all variables included in the econometric models, yielding a working sample of 4,851 students from 196 schools. Table 2 sets out the derivation. Of the 7,293 students in the national sample, 5,007 have complete data on all student-level predictors; the further reduction to 4,851 reflects missing school-level information, principally from the four schools whose principals did not complete the teacher-staffing items.

Most attrition occurs at the student level and is attributable to PISA's rotated questionnaire design rather than to non-response. Mathematics self-efficacy, disciplinary climate and teacher support were administrated to subsets of the sample, and 862 students are missing precisely these three items. Missingness of this kind arises rather than from non-response. by design and is unrelated to student characteristics conditional on form assignment.

The consequences for representativeness nonetheless require statement. Students in the analytical sample score 375.7 points on average, against 342.6 among those excluded and 364.5 in the full national sample. Two points follow. Estimates describe a relatively stronger subgroup rather than the national cohort; and because the weakest students are disproportionately excluded, associations estimated at the lower end of the achievement distribution are more likely to be attenuated than inflated. Section 7.3 discusses this further.

All descriptive statistics and single-level estimates apply the PISA final student weight (W_FSTUWT). Weighted estimation on the full national sample reproduces the mathematics mean published by the OECD to within half a point, confirming correct application. Replicate weights for balanced repeated replication are supplied in PISA 2022 as W_FSTURWT1 through W_FSTURWT80. These are not used in the models reported below, since neither the multilevel nor the quantile estimator accommodates them; the implications are noted in Section 4.5 and Section 7.3.

4.4 Variables

Dependent variable. Mathematics achievement is measured using the ten plausible values provided by PISA 2022 (PV1MATH-PV10MATH). Plausible values are multiply imputed draws from each student's estimated proficiency distribution rather than a single point estimate, and their use is mandated by the OECD to avoid underestimating population variance (Von Davier, Gonzalez, and Mislevy, 2009). All models are estimated separately on each of the ten plausible values and coefficients averaged, with the robustness of each estimate assessed by the number of plausible values in which it attains conventional significance, and the full per value output appears in the Supplementary Information. **Student-level variables (Level 1).** Socioeconomic status is measured through the PISA ESCS index – a composite of parental education, parental occupation, and home educational resources. Gender is coded as a binary indicator for male students. Mathematics self-efficacy (MATHEFF) captures students' beliefs about their capacity to perform specific mathematical tasks. Teacher support (TEACHSUP) measures students' perceptions of their mathematics teachers' attentiveness and responsiveness. Sense of belonging (BELONG) measures social integration at school. Disciplinary climate (DISCLIM) measures classroom order and focus in mathematics lessons. All PISA indices are standardized to a mean of zero and standard deviation of one across OECD countries, so coefficients are interpretable as score-point differences per one standard deviation change. Table 2 presents descriptive statistics for all variables.

School-level variables (Level 2). Five school-level predictors are drawn from the principal questionnaire. Inadequately qualified teaching staff (UNQUALSTAFF, SC017Q02NA) is a binary variable equal to one if the principal reports instruction hindered by qualification inadequacy. Staff quantity shortage (STAFFSHORT, SC017Q01NA) is a parallel binary indicator. School location (RURAL, SC001Q01TA) distinguishes rural from urban schools. Student-teacher ratio (STRATIO) captures class size pressure, School-level mean socioeconomic composition is constructed as the within-school mean of student ESCS and enters the extended specification described below.

Teacher certification. The school questionnaire records the number of fully certified mathematics teachers rather than a certification rate, so a school-level rate was derived by dividing certified staff by total staff, combining full-time and part-time counts. Applying the validity check specified in section 3.2 shows this measure to be unusable in its continuous form. The distribution is bimodal rather than merely skewed: 37 % of schools report that no teaching staff hold full certification and 19 % report that all do, leaving under half the sample at intermediate values. Schools at the two

poles employ 56 and 73 teachers on average against a sample mean near 63, which excludes a small-school arithmetic artefact and points instead to inconsistent interpretation of the item by principals. Certification therefore enters the models as a three-category variable - none, partial, full - with schools reporting no certified staff as the reference category. Section 7.3 considers what this permits and rules out.

4.5 Estimation

The empirical analysis proceeds through four models, each estimated in Stata 17 on all ten plausible values.

Model 1: null model. Students are nested within schools in the PISA data, violating the independence assumption of ordinary least squares regression. Ignoring this clustering produces incorrect standard errors and misleading conclusions about school-level effects. Two-level hierarchical linear modelling (HLM) addresses this by simultaneously estimating within-school (Level 1) and between-school (Level 2) effects. The unconditional null model first partitions total variance into within-school and between-school components through the intraclass correlation coefficient (ICC):

$$Y_{ij} = \gamma_{00} + \varepsilon_{ij} + r_{0j}$$

Where

Y_{ij} is the mathematics plausible value for the student i in school j

γ_{00} is the grand mean mathematics score

ε_{ij} is the residual for student i in school j

r_{0j} is the school level residual

The key diagnostic statistic derived from Model 1 is the Intraclass Correlation Coefficient (ICC), which expresses the proportion of total score variance attributable to differences between schools:

$$ICC_{school} = \frac{\tau^2}{\tau^2 + \sigma^2}$$

where τ^2 is the between-school variance and σ^2 is the within-school variance. A high ICC signals that school identity is an important determinant of individual student outcomes — and thus that multilevel modelling is not merely appropriate but necessary.

Model 2: OLS Baseline with Clustered Standard Errors. The OLS baseline model includes student-level predictors only, with standard errors clustered at the school level. It provides a

transparent description of student-level associations and a comparison point for assessing omitted variable bias in the HLM. The estimated equation is:

$$Y_{ij} = \beta_{0j} + \beta_1 ESCS_{ij} + \beta_2 Gender_{ij} + \beta_3 MATHEFF_{ij} + \beta_4 DISCLIM_{ij} + \beta_5 TEACHSUP_{ij} + \beta_6 BELONG_{ij} + \varepsilon_{ij}$$

where Y_{ij} is the (averaged) plausible value mathematics score for student i in schoolj; β_1 through β_6 are partial regression coefficients representing the expected change in mathematics score associated with a one-unit increase in the respective predictor, holding all other predictors constant; and ε_{ij} is the error term.

Model 3: Two-Level HLM

The primary analytical model is a two-level HLM in which students (Level 1) are nested within schools (Level 2). The HLM is the appropriate specification for PISA data because it simultaneously models within-school and between-school variation, produces unbiased standard errors under the clustering structure of the sample, and permits the investigation of school-level predictors as explanatory variables for school-level score differences. A single-level OLS model applied to clustered data would underestimate standard errors, inflating Type I error rates (Raudenbush & Bryk, 2002).

The full two-level HLM is specified as follows:

$$Level\ 1$$

$$Y_{ij} = \beta_{0j} + \beta_1 ESCS_{ij} + \beta_2 Gender_{ij} + \beta_3 MATHEFF_{ij} + \beta_4 DISCLIM_{ij} + \beta_5 TEACHSUP_{ij} + \beta_6 BELONG_{ij} + \varepsilon_{ij}$$

$$Level\ 2$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} STAFFSHORT_j + \gamma_{02} UNQUALSTAFF_j + \gamma_{03} CERT_j + \gamma_{04} RURAL_j + \gamma_{05} STRATIO_j + \gamma_{06} ESCS_j + u_{0j}$$

Combining these two levels yields the reduced-form mixed model equation:

$$Y_{ij} = \gamma_{00} + \gamma_{01} STAFFSHORT_j + \gamma_{02} UNQUALSTAFF_j + \gamma_{03} CERT_j + \gamma_{04} RURAL_j + \gamma_{05} STRATIO_j + \gamma_{06} ESCS_j + \beta_1 ESCS_{ij} + \beta_2 Gender_{ij} + \beta_3 MATHEFF_{ij} + \beta_4 DISCLIM_{ij} + \beta_5 TEACHSUP_{ij} + \beta_6 BELONG_{ij} + u_{0j} + \varepsilon_{ij}$$

Both models are estimated by maximum likelihood with robust standard errors clustered at the school level. Restricted maximum likelihood, which is often preferred for variance component estimation, cannot be used when probability weights are applied and is therefore unavailable here.

The PISA final student weight is applied at Level 1; PISA supplies no school-level weight suited to multilevel estimation, and weighting at a single level may bias estimates where inclusion probabilities differ systematically across levels.

Model 4: Quantile Regression. Quantile regression, introduced by Koenker and Bassett (1978) and applied to education data by Koenker and Hallock (2001), estimates the conditional quantile function:

$$Q_\tau(PVMATH_{ij}|X_{ij}, Z_j) = \alpha(\tau) + X_{ij}\beta(\tau) + Z_j\gamma(\tau)$$

where, $Q_\tau\ denotes\ the\ \tau - th\ conditional$ quantile of mathematics performance; $PVMATH_{ij}$ is the mathematics plausible value for student i in school j; X_{ij} is a vector of student-level predictors comprising ESCS, gender, mathematics self-efficacy, disciplinary climate, teacher support and sense of belonging; $Z_j$is a vector of school-level predictors: UNQUALSTAFF, STAFFSHORT, CERT, RURAL, school mean ESCS and STRATIO; $\beta(\tau)$ is a vector of student-level coefficients that vary across quantiles τ; $\gamma(\tau)$ is a vector of school-level coefficients that vary across quantiles τ; $\alpha(\tau)$ is the quantile-specific intercept.

Five quantiles are estimated (τ = 0.10, 0.25, 0.50, 0.75, 0.90) using simultaneous quantile regression (sqreg) with 100 bootstrap replications for standard error estimation.

4.6 Robustness checks

Four checks support the primary results, reported in full in Supplementary Information.

First, the dependent variable is replaced with reading and science plausible values as alternative outcomes, testing weather school-level associations are domain-general or specific to mathematics. Second, models are estimated separately for rural and urban schools to examine geographic heterogeneity in the school-level associations. Third, mathematics self-efficacy (MATHEFF) is excluded to assess whether its inclusion suppresses instructional quality coefficients through mediation. The sensitivity model recovers 598 additional observations previously excluded due to METHEFF missingness (n increases to 5,449 students). The multilevel form of this specification did not converge and the check is reported as a single-level model with clustered standard errors. Fourth, variance inflation factors are computed for the corresponding single -level specification to assess multicollinearity.

Across all models, coefficient stability is assessed by the number of plausible values in which an estimate attains conventional significance, following the robustness criterion specified in Section 3.2.

4.7 Operationalizing the ERTF

Each determinant surviving the validity gate is evaluated against four appraisal criteria before being considered for policy recommendation. This explicit appraisal process reduces reliance on ad hoc expert judgement and increases the transparency of how empirical findings are converted into reform priorities.

Table 1. ERTF evidence appraisal criteria

Step	Criterion	**Definition**	**Application in this study**
Validity gate	**Measurement Validity**	The variable measures what its name claims, judged by inspecting the distribution, comparing against external benchmarks, and considering how respondents are likely to have read the question	Applied to every school-level measure. Teacher certification did not survive it. Schools of similar size sit either at zero or at full coverage, suggesting principals read the item differently. The variable was recoded to permit estimation but excluded from priority consideration.
Appraisal criteria	**Statistical Robustness**	The finding is consistent across the HLM and quantile regression specifications and stable across all ten plausible values.	The coefficient must attain $p<0.05$ in at least 8 of the 10 plausible values and remain directionally consistent across the estimated quantiles.
	Effect Magnitude	The association is large enough to be educationally meaningful, not merely statistically significant.	Effect size benchmarked against OECD-average associations and the within-Uzbekistan distribution of mathematics scores.
	Equity Relevance	Addressing the determinant would disproportionately benefit disadvantaged or low-preforming students.	Assessed from the quantile regression evidence: variables with larger effects at Q10 than at Q90 score higher on this criterion.
	Implementation Feasibility	Government can realistically influence the variable within existing institutional and financial constraints.	Assessed against Uzbekistan's existing Presidential Decree Framework and current Ministry of Preschool and School Education capacity.

Note: HLM (Hierarchical Linear Modelling); QR (quantile regression); Q10 and Q90 denote the 10^{th} and 90^{th} conditional quantiles of the mathematics achievement distribution. A determinant is elevated to a reform priority only when it satisfies the statistical-robustness threshold and scores favorably on the remaining three criteria.

Source: Developed by the author, drawing on Nutley et al. (2007) and Stewart et al. (2022) on evidence utilization, and on Vanlommel and van den Boom-Muilenburg (2024) on implementation.

4.8 Ethical considerations

This study uses publicly available, fully anonymised secondary data released by the OECD. The PISA 2022 database contains no individual student or school identifiers, and all data were collected under the ethical standards of the OECD Data Collection and Reporting Framework. No direct contact with human participants was involved. Ethical approval was not required for this study as the analysis uses publicly available secondary data from the OECD PISA 2022 international database.

5. Results

5.1 Overview

The empirical analysis proceeds in five stages consistent with the analytical strategy described in Section 4. Descriptive statistics establish the baseline profile of the sample. Hierarchical linear modelling identifies student-level and school-level determinants of mathematics achievement while accounting heterogeneous effects across the achievement distribution. Evidence appraisal applies the ERTF criteria to determine which findings warrant policy priority. Policy prioritization translates the appraised evidence into four strategic reform priorities. Unlike most PISA-based studies, the objective of this section is not only to report statistical relationships but to demonstrate how empirical evidence can be systematically transformed into practical policy guidance.

5.2 Descriptive statistics

Uzbekistan's fifteen-year-olds scored a weighted average of 364.5 points in mathematics in PISA 2022, which matches the figure the OECD published for the country and confirms that the sampling wights have been applied correctly. The score sits 108 points below the OECD average pf 472. More telling than the average is the shape of the distribution: 80.3 percent of students fall below Level 2, the threshold the OECD treats as the minimum needed to participate fully in economic and civic life. When four in five are below that line, the average conceals more than it reveals, which is why the distributional analysis in Section 5.7 matters as much as the mean effects reported in Section 5.5

Table 2 describes the 4,851 students and 196 schools that enter the models. This sample is not a miniature of the national one. Students with complete data on every model variable average 375.7 points, while those excluded average 342.6, a difference of 33 points or roughly half a standard deviation. The reason lies mainly in PISA’s rotated questionnaire design rather than in refusal to answer: self-efficacy, disciplinary climate and teacher support went to subsets of students, and 862 students are missing exactly those three items. Three quarters of the analytical sample falls below Level 2, against four fifth nationally. Readers should therefore treat what follows as describing a somewhat stronger group of students than the country as a whole, and should note the quantile estimates at the bottom of the distribution are more likely to understate than to exaggerate the associations reported there. Section 7.3 returns to this.

Socioeconomic status in the sample averages -0.651 on the ESCS index, well below the OECD mean of zero. What matters more the analysis is how unevenly students are distributed across schools. School-level mean ESCS ranges from -2.24 to +0.50, with a standard deviation of 0.457, so the distance between a school at the tenth percentile of intake and one at the ninetieth exceeds a full standard deviation of individual socioeconomic status. Uzbek schools differ considerably in whom they enroll, and Section 5.5 and 5.6 show that this sorting drives much of what initially looks like variation in school quality.

Among school principals, 29.1 percent report that instruction suffers because teaching staff are inadequately qualified, and 25.0 percent report that there are simply too few of them. Keeping these two apart turns out to matter a great deal. Rural schools make up 67.9 percent of the sample by school and about 64 percent by student once weights are applied. Rural students average 358.9 points against 374.7 for urban students, and they attend schools whose intake is considerably

poorer (mean ESCS -0.83 against -0.35). Section 5.6 asks whether the performance gap follows from the schools or from the students who attend them.

Teacher certification needs longer explanation. PISA's school questionnaire asks principals how many of their mathematics staff are fully certified, not what share of them are, so a rate was constructed by dividing certified staff by total staff and combining full-time with part-time counts. The resulting variable is bimodal. Around 37 percent of schools report that none of their teaching staff hold full certification and around 19 percent report that all of them do, leaving under half the sample anywhere in between. This is not an artefact of school size: schools at both extremes employ around 65 and 73 teachers respectively, against a sample average near 63. A more plausible reading is that principals understood 'fully certified by the appropriate authority' in quite different ways, perhaps because the category translates awkwardly into Uzbekistan's own credentialing system. Whatever the cause, treating the variable as continuous would mean assuming that a move from 40 to 41 percent certification carries the same weights as a move from 0 to 1 percent, which the distribution does not support. Certification enters the models as three categories instead - none, partial, and full – and Section 7.3 discusses what this measurement problem does and does not allow to conclude.

Table 2. Descriptive statistics – analytical sample (n=4,851 students,196 schools)

Variable	Mean	SD	Min	Max	Expected sign
Dependent variable					
Mathematics Score (PV1)	375.67	66.93	156.9	625.0	—
Level 1 — Student-level (weighted, n = 4,851)					
ESCS index	−0.651	1.004	−4.135	2.489	+
Gender (male=1)	0.500	0.500	0	1	+
Mathematics self-efficacy	−0.409	1.365	−3.731	2.435	+
Disciplinary climate	0.287	1.114	−2.493	1.851	+
Teacher support	0.404	1.175	−2.910	1.556	+
Sense of belonging	0.121	1.025	−3.257	2.785	+

Level 2 — School-level (n = 196 schools)					
Inadequately qualified staff	0.291	0.455	0	1	−
Teaching staff shortage	0.250	0.434	0	1	−
Certified teachers (%)	25.29	40.33	0	100	+
Rural location	0.679	0.468	0	1	−
Student-teacher ratio	13.49	5.78	1.0	28.6	−
School-mean ESCS	−0.698	0.457	−2.244	0.495	+

Source: PISA 2022 international database (OECD, 2023) and author's calculations

5.3 How much variation lies between schools

Before introducing any predictor, the null model divides total variance in mathematics scores into a within-school and a between-school part. Averaged over the ten plausible values, between-school variance is 1,040.2 and within-school variance 3,504.7, giving an intraclass correlation of 0.229 (Supplementary Table S1). The estimate barely moves across plausible values, ranging from 0.212 to 0.246. Full variance components for each plausible value are reported in Supplementary Table S2.

Just under a quarter of the variation in achievement, then, lies between schools rather than among students within them. That is comfortably above the five percent conventionally taken as the point where multilevel modelling becomes necessary (Raudenbush and Bryk, 2002), and it confirms that single-level regression would produce misleadingly precise standard errors for any school-level predictor. It is also below the OECD-average between-school share of roughly thirty percent. Uzbek schools differ from one another less than schools in many OECD systems do, and more than three quarters of the variation in scores remains inside them.

At first glance this might seem to weaken the case for school-level reform. It does not. The variation that sits within schools is largely the variation in students' home circumstances, and education ministries cannot legislate household income and parental education. What they can influence is who teachers, under what conditions, and with what support, and those conditions operate at the school level. A between-school share of 23 percent is the part of the problem that policy can reach directly, which is why the analysis concentrates there.

One further number belongs here. The grand mean in the analytical sample is 372.8 rather than the national 364.5, which is the selection effect described in Section 5.2 showing up again.

5.4 Student-level associations

Table 3 reports an ordinary least squares baseline with student characteristics only and standard errors clustered by school. Every coefficient carries the expected sign, and all but disciplinary climate clear the one percent level.

Self-efficacy is among the strongest predictors at 7.92 points per standard deviation, in line with what PISA studies have found repeatedly (Lie et al., 2024; Niu et al., 2025). Boys outscore girls by 11.23 points on average, a gap that Section 5.7 shows to be far from evenly distributed.

Socioeconomic status is worth 8.37 points per unit, sense of belonging 6.89 and teacher support 5.56. Disciplinary climate trails the rest at 2.76.

Together these variables account for 9.4 percent of the variation in scores. That figure is low, two things explain it. Uzbekistan's socioeconomic gradient is compressed by international standards, so ESCS carries less explanatory weigh that it would in a wealthier system, and this specification omits everything about the schools themselves. Section 5.5 adds them.

Table 3. OLS baseline results – averaged across ten plausible values (n = 4,851 obs., 196 schools)

Variable	Coefficient	Interpretation
ESCS	8.37 ***	One-unit SES increase → 8.4 score points
Gender (male=1)	11.23 ***	Male advantage in mathematics
Mathematics self-efficacy	7.92 ***	Strongest student-level predictor
Teacher support	5.56 ***	
Sense of belonging	6.89 ***	
Disciplinary climate	2.76 **	
Constant	374.09 ***	
R^2	0.094	9.4 % of variance explained

Note. Standard errors clustered at school level (196 clusters); weighted by W_FSTUWT. n = 4,851. *** p < 0.001, ** p < 0.01. Source: PISA 2022 international database (OECD, 2023) and author's calculations

Estimates for each plausible value appear in Supplementary Table S3.

5.5 What schools contribute

Table 4 reports two multilevel models. The first adds school characteristics to the student -level predictors; the second adds school-level mean socioeconomic composition. The second is the

specification the analysis relies on. Leaving composition out biases the remaining school-level coefficients, since intake is correlated with almost everything else a principal report; the extended model also accounts for 29.2 percent of between-school variance against 20.3 percent for the first; and separating institutional effects form intake differences is precisely the question a ministry needs answered before allocating resources.

The individual ESCS coefficient falls from 8.37 in the OLS baseline to 3.46 once schools are modelled, an attenuation of about 59 percent. School-mean ESCS, meanwhile, carries a coefficient of 24.49 and is significant in all ten plausible values. Socioeconomic advantage in Uzbekistan works far more through which school a student attends than through how that student compares with classmates. This is a sorting story rather than a story about individual endowments, and it carries an unusually direct policy implication: if the channel is school composition, the school-level policy can act on it, whereas household circumstances lie beyond a ministry's reach. Self-efficacy (7.44), belonging (6.15) and teacher support (4.38) all survive the addition of school controls, so instructional climate is doing work independently of who the students are.

Of the school-level variables, only inadequately qualified teaching staff is reliably associated with achievement. Its coefficient is -11.42 and it reaches significance in eight of ten plausible values, which meet the robustness criterion set out in Table 1. The two plausible values falling short of conventional significance are identified in Supplementary Table S6. Schools whose principals report that qualification problems hinder instruction score about eleven points lower, holding student composition constant. Coefficients for all ten plausible values are reported in Supplementary Tables S4 and S5.

Staff shortage, by contrast, is positive and nowhere near significance, and the student-teacher ratio is indistinguishable from zero. It is the quality of teachers that constrains performance here, not

how many of them there are. The two conditions call for different instruments, so the distinction is not academic.

Neither certification category differs from schools reporting no certified staff at all. Schools where every teacher is formally certified perform no better than schools where none is, once principals' assessments of qualification adequacy are taken into account. This echoes Rivkin, Hanushek, and Kain (2005), who found that observable credentials explain little of the variation in teacher effectiveness. Some caution is needed, though. A null result drawn from a variable with the measurement problems described in Section 5.2 cannot bear too much weight. What can be said that principals' report of whether qualification problems actually impede teaching carry information about student outcomes that formal credential counts do not.

Table 4. Two-level hierarchical linear models (averaged across ten plausible values)

	Model 3 (base)	Model 3b (extended)
Level 1 — Student-level fixed effects		
ESCS	3.47***	3.46***
Gender (male=1)	9.37***	9.37***
Mathematics self-efficacy	7.44***	7.44***
Teacher support	4.38***	4.38***
Sense of belonging	6.15***	6.15***
Disciplinary climate	1.60	1.60
Level 2 — School-level fixed effects		
Inadequately qualified teacher	−9.35 (2/10)	−11.42 (8/10)
Teaching staff shortage	7.59	7.71
Certification: partial vs. none	−1.30	−1.12
Certification: full vs. none	1.91	3.18
Rural location	−13.52 (10/10)	−1.21 (0/10)
Student-teacher ratio	0.10	−0.03
School-mean ESCS	—	24.49 (10/10)

	Model 3 (base)	Model 3b (extended)
Constant	379.40	390.16
Variance components		
Between-school variance (τ^2)	828.9	736.8
Within-school variance (σ^2)	3,275.1	3,275.3
Conditional ICC	0.202	0.184
Reduction in τ^2 from null model	20.3%	29.2%
Observations / schools	4,851 / 196	4,851 / 196

Note. Maximum likelihood with robust standard errors clustered at school level. Figures in parentheses give the number of plausible values (of ten) in which the coefficient reaches $p < 0.05$. n = 4,851 students, 196 schools.

Source: Author's Calculation in Stata 17.

5.6 Rural and urban schools

Rural students score 358.9 against 374.7 for urban students, and the base multilevel model puts the rural coefficient at -13.52, significant in every plausible value. On its own, that looks like clear evidence of weaker rural schools. Subsample estimates for each plausible value appear in Supplementary Table S7.

The extended model tells a different story. Once school-mean ESCS enters, the rural coefficient falls to -1.21 and fails to reach significance in any of the ten plausible values. Rural schools serve much poorer intakes than urban ones (mean ESCS −0.83 against −0.35), and comparing schools with similar intakes leaves no rural–urban difference worth speaking of. The rural deficit is real in the sense that rural students do score lower; it is not, however, evidence that rural schools are less effective. They educate poorer children. Figure 3 shows the coefficient before and after the control applied.

Location does not matter for how teacher qualification operates, though. Estimating the model separately for each group gives a coefficient on inadequately qualified staff of -19.57 in urban

schools (1,792 students in 63 schools), significant in all ten plausible values, against -4.65 in rural schools)3,059 students in 133 schools), significant in none. The penalty is roughly four times larger in cities and detectable inly there. This also explains why the pooled estimate in the first model looked weak: averaging a large urban effect with a negligible rural one produces a middling coefficient that struggles for significance.

These two results fit together, though the fit is worth spelling out. Schools with comparable intakes perform alike wherever they are, but the circumstances under which a principal's report of qualification problems translates into measurable learning loss are concentrated in urban schools. Uzbekistan's cities contain both selective specialist schools and severely overcrowded multi-shift ones, so the same questionnaire item may be separating genuinely different teaching enviornments there, while rural schools face more uniform constraints that leave the item with less to distinguish.

5.7 Where in the distribution the effects fall

Average effects can mislead when four fifth of students sit below a proficiency threshold. Table 5 estimates the model at five points of the conditional distribution to see whether the determinants of achievement work differently for weeks and strong students. Average effects can mislead when four fifths of students sit below a proficiency threshold. Table 5 estimates the model at five points of the conditional distribution to see whether the determinants of achievement work differently for weak and strong students. Figure 4 shows the result.

They do, and the pattern favors the students Uzbekistan most needs to reach. Teacher support is largest in the lower and middle of the distribution, averaging 5.62 at the tenth percentile and 6.36 at the median, before dropping to 3.39 at the ninetieth. Sense of belonging declines steadily from 7.79 to 4.70 across the same range. Both patterns run in the same direction in all ten plausible

values. Since roughly eighty percent of student fall below Level 2, this is where school climate does its most useful work: it lifts the floor rather than the ceiling.

Socioeconomic status behaves the opposite way, rising from 4.80 at the tenth percentile to 11.15 at the ninetieth. Plausible-value-specific quantile coefficients appear in Supplementary table S10 Family background barely separates the weakest students from one another but distinguishes sharply among the strongest. The point is not that socioeconomic inequality can be set aside. It is that policies aimed at it will show their largest returns among students who are already doing well, which is not where the urgency lies. For the majority below Level 2, instructional climate is both the more tractable lever and the better-targeted one.

The gender gap follows the same logic as socioeconomic status. Boys hold no measurable advantage at the tenth percentile, where the coefficient is small and insignificant, but lead by 18.93 points at the ninetieth. Uzbekistan's aggregate gender gap in mathematics is generated almost entirely among the high performers.

Two further results are worth recording. The rural coefficient is negative and significant at every quantile, so the gap documented in Section 5.6 is spread evenly across the distribution rather than concentrated among the particular students. Certification, again, is insignificant everywhere.

Table 5. Simultaneous quantile regression of mathematics achievement at five conditional quantile

Variable	Q10	Q25	Q50	Q75	Q90
Student-level predictors					
Socioeconomic Status (ESCS)	4.80***	5.16***	7.97***	9.57***	11.15***
Gender (male = 1)	4.34	5.95**	9.21***	14.89***	18.93***
Mathematics self-efficacy	3.27***	4.61***	6.79***	10.14***	12.36***
Disciplinary climate	2.90**	4.64***	4.09***	3.48**	1.87
Teacher support	6.47***	5.20***	6.78***	4.42***	2.56

Variable	Q10	Q25	Q50	Q75	Q90
Sense of belonging	8.40***	7.18***	5.68***	6.40***	4.47**
School-level predictors					
Inadequately qualified teacher	−10.38**	−9.36***	−7.18***	−8.87***	−13.58***
Teaching staff shortage	10.49***	6.18**	7.01**	12.18***	22.76***
Certification: partial vs. none	−4.48	0.62	−0.57	−3.07	−2.26
Certification: full vs. none	−0.07	2.59	1.86	4.62	−2.46
Rural location	−8.68***	−12.22***	−11.18***	−12.93***	−12.14***
Student-teacher ratio	0.03	0.14	0.02	0.00	0.18
Model					
Constant	301.76	336.02	378.43	424.85	463.35
Pseudo R^2	0.043	0.042	0.051	0.068	0.089

Note. Estimated in Stata 17 using simultaneous quantile regression (sqreg) with 100 bootstrap replications; n = 4,851 students in 196 schools. Coefficients derive from the first plausible value; averages across all ten plausible values for teacher support and sense of belonging appear in Appendix Table S10 and confirm the gradients shown here.

*** $p < 0.01$, ** $p < 0.05$.
Source: PISA 2022 international database (OECD, 2023) and author's calculations

5.8 Robustness

Replacing the mathematics outcome with reading and science plausible values reproduces the climate findings and, if anything, strengthens them. Teacher support carries coefficients of 6.89 in reading and 5.11 in science, belonging carries 7.88 and 7.41. All four are significant in every plausible value. Full domain results are reported in Supplementary table S8. Whatever these variables are capturing, it is not specific to mathematics. The qualification penalty behaves as a content-knowledge account would predict, being largest in mathematics at -9.35, smaller in reading at -7.64 and smallest in science at -6.35, though it reaches conventional significance only discontinuously in the pooled models, which is why the compositional and subsample estimates in

Sections 5.5 and 5.6 carry the argument. The rural coefficient is negative and significant in all ten plausible values in both additional domains.

Dropping self-efficacy, the largest single source of item non-response, expands the sample to 5,449 students. Supplementary Table S9 reports the expanded-sample estimates. Estimated by OLS with clustered standard errors, nothing of substance changes: teacher support 6.10, belonging 7.95, inadequately qualified teacher -10.13, rural location -11.59. the coefficients run somewhat larger than in the multilevel models because OLS does not separate out between-school variance and because self-efficacy is no longer absorbing part of the climate effect. The findings do nor depend on the complete-case restriction.

Variance inflation factors from the corresponding OLS specification run from 1.03 to 1.52 with a mean of 1.20, so multicollinearity is not distorting the estimates. Every model was estimated separately on each of the ten plausible values and the results averaged; coefficient ranges appear in the appendix, and the direction of each substantive finding holds in all ten. Supplementary Table S11 reports the full variance inflation factors.

5.9 Evidence appraisal using the ERTF

The models reported above identify several statistically significant associations. Not all of them warrant policy attention, and one does not survive appraisal at all. Table 6 sets out the appraisal matrix, applying the validity gate and the four criteria specified in section 3.2.

Teacher certification fails at the gate. As Section 5.2 described, the derived certification rate is bimodal, with schools of comparable size clustered at zero and at complete coverage, indicating that principals interpreted the source item inconsistently. The categorical recoding permits the variable to enter the models, and its coefficients are uniformly non-significant, but a null result from an unsound measure supports no inference about certification itself. The finding is therefore excluded from appraisal. It is not, however, discarded: an indicator that cannot support analysis

cannot support monitoring either, and Section 6 treats the reconstruction of a valid certification measure as a reform objective in its own right.

Inadequately qualified teaching staff clears the gate and scores well on all four criteria. The measure is a direct principal report of whether qualification problems impede instruction, with no missing observations and a straightforward interpretation. The coefficient attains significance in eight of ten plausible values in the primary specification, meeting the robustness threshold, and reaches all ten in urban schools. An association of roughly eleven score points is educationally substantial is a system where the national mean sits 108 points below the OECD average. Equity relevance is established less by the quantile gradient than by the concentration of the penalty in urban schools serving heterogenous intakes. Feasibility is high: teacher qualification is directly addressable through existing instruments, and Uzbekistan already has the legislative vehicle in place.

Teacher support and sense of belonging clear the gate and core highly throughout. Both are significant in the multilevel model and across four of five quantiles, with their largest associations among the weakest students. Both replicate in all ten plausible values in reading and science, which is the strongest evidence of robustness anywhere in the analysis. Both fall within the reach of education authorities through professional development, making them feasible as well as effective.

Disciplinary climate presents a more mixed case. It is significant in the OLS baseline and across the lower and middle quantile but only marginal in the multilevel specification, and it disappears entirely in urban schools. It clears the robustness threshold in the distributional analysis while falling short of it in the primary model, and its effect magnitude is modest. It is retained as a secondary consideration rather than priority.

School-level socioeconomic composition presents the reverse difficulty. The evidence is unambiguous: the coefficient is the largest at Level 2, significant in all ten plausible values, and it accounts entirely for the observed rural performance gap. Yet composition is not directly modifiable. A ministry cannot legislate the socioeconomic profile of a school's catchment. What follows is not that the finding is unusable but that it operates differently from the others – as an addition that reform design must accommodate, and as a correction that accountability systems must apply, rather than as a lever to be pulled. Individual socioeconomic status follows the same logic, and is classified as a background condition motivating equity-targeted design rather than a reform target.

Teaching staff shortage and the student-teacher ratio clear the gate but fail on the other criteria. Neither is significant in the multilevel models. Staff shortage is positive and significant across the quantiles, which is unlikely to reflect a benefit of understanding and more likely indicates that schools reporting a need for teachers are those experiencing growth or excess demand, characteristics correlated with stringer intakes. The variable does not measure what its name suggests, and neither variable supports a reform priority.

Table 6. ERTF evidence appraisal matrix

Determinant	Validity	Robustness	**Magnitude**	**Equity**	**Feasibility**	**Appraisal outcome**
Determinants elevated to reform priorities						
Inadequately qualified teaching staff	Pass	**High (8/10)**	**High**	**High**	**High**	**Priority 1**
Teacher support	Pass	**High(10/10)**	**High**	**High**	**High**	**Priority 2**
Sense of belonging	Pass	**High (10/10)**	**High**	**High**	**High**	**Priority 2**
School mean ESCS	Pass	**High (10/10)**	**High**	**High**	Low	**Priority3 (diagnostic, not lever)**
Determinants failing the validity gate						
Teacher certification coverage	Pass	—	—	—	—	**Priority 4 (indicator reconstruction)**

Determinant	Validity	Robustness	**Magnitude**	**Equity**	**Feasibility**	**Appraisal outcome**
Determinants retained but not elevated						
Disciplinary climate	Pass	Mixed	Moderate	Moderate	High	Secondary consideration
Individual socioeconomic status	Pass	High (10/10)	High	Moderate	Low	Background condition
Rural school location	Pass	High (10/10) before control; null after	Moderate	Moderate	Moderate	Subsumed within Priority 3
Determinants not supporting a priority						
Teaching shortage staff	Questionable	Low	Low	Low	Moderate	Not prioritized
Student-teacher ratio	Pass	Low	Low	Low	High	Not prioritized

Note: Appraisal follows the two-step procedure specified in Section 3.2 and summarized in table 1

5.10 Policy prioritization

Following evidence appraisal, four strategic reform priorities emerge.

Priority 1 - Teacher qualification adequacy, concentrated in urban schools. Inadequately teaching staff is the binding school-level constraint on mathematics achievement, at roughly elven score points net of intake compositions. The subsample analysis specifies where the constraint operates: the penalty is approximately four times larger in urban schools, where it is robust across all ten plausible values, and statistically undetectable in rural schools. This runs against the intuition that qualification problems concentrate in remote areas, and it warrants care in interpretation, The likeliest explanation is not that rural teaching is better but that rural schools are more uniformly constrained, so that a principal's report of qualification adequacy distinguishes little between them, whereas urban system span selective specialist institutions and severely overcrowded multi-shift schools, and the same report separates genuinely different teaching environments. Whatever the mechanism, the deployment implication is clear: interventions targeting teacher qualification should be directed first at urban schools, where the association with learning outcomes is strongest and most reliably measured.

Priority 2 - School Climate as an equity intervention. Teacher support and sense of belonging carry their largest associations among the weakest students, and both replicate across mathematics, reading, and science in every plausible value. Where roughly for students in five fall below minimum proficiency, an intervention whose returns concentrate at the bottom of the distribution reaches most of the population that reform must serve. Climate investment raises the floor; socioeconomic advantage, which steepens toward the top of the distribution, raises the ceiling. For a system whose principal difficulty is the sheer proportion of students below Level 2, the former is the better-targeted instrument.

Priority 3 – Composition-adjusted resource allocation and diagnostics. School-level socioeconomic composition is the largest school-level association in the analysis, and it accounts entirely for the apparent rural performance deficit. Two consequences follow. Resource allocation should track intake composition rather than location, since geography turns out to be a proxy for disadvantage rather than an independent source of it, and a rural school serving a comparatively advantaged intake has less claim on compensatory resources than an urban school serving as disadvantaged one. Equally, any accountability framework built on unadjusted means scores will systematically mispresent school serving disadvantaged populations, understating their contribution to learning and misdirecting both support and scrutiny. Composition-adjusted indicators are a precondition for the monitoring architecture the framework requires.

Priority 4 – Rebuilding the teacher qualification indicator. The appraisal established that Uzbekistan's certification data cannot support a valid school-level measure. This is not a marginal difficulty. Teacher certification is among the most obvious indicators a ministry would adopt to track the progress on teacher quality, and the 2030 targets under the existing decree framework presuppose that such an indicator exists. An analysis that cannot construct one from national data

has identified a gap in monitoring capacity that will constrain the reform cycle irrespective of how the substantive priorities are pursued. Establishing a certification measure with a clear definition, consistent application across school, and administrative verification is therefore a reform priority in its own right, and a precondition for the adaptive learning stage of the framework.

Table 6 summarizes the appraisal, and Section 6 turns these priorities into an operational reform package.

6. Operationalizing the ERTF

6.1 Connecting evidence to reform priorities

The evidence appraisal in Section 5.9 produces a ranked set of findings, it does not produce a reform package. Converting those findings into operational reform requires three additional steps: aligning each priority with existing legislative instruments, sequencing reforms across time horizons that reflect what the system can absorb, and establishing measurable indicators through which progress can be tracked. This section works through those steps for Uzbekistan.

The connection between evidence and policy is not mechanical. And the framework does not assume that a statistically significant coefficient automatically implies a specific intervention — it assumes that empirical findings constrain the choice of interventions while leaving implementation design to the judgement of policymakers who understand the institutional context. What the framework provides is the basis for that judgement: which constraints are empirically binding, how large their effects are, for whom, and where they are most severe.

6.2 Reform priorities and legislative alignment

Four reform priorities emerge from the appraisal process, each aligned with existing Presidential decrees rather than requiring new legislative instruments. What this section adds is the instrument in each case and the reasoning behind the sequencing the empirical basis for each priority is set out in Section 5.10. Table 7 summaries the alignment.

Priority 1 - Teacher qualification adequacy. Presidential Decree No DP-79 (2023), which established a salary supplement for certified teachers, is the instrument already in place. The empirical contribution here is to specify where it should be concentrated. The penalty is approximately four times larger in urban schools, where it is robust across all ten plausible values,

and statistically undetectable in rural ones, the supplement and any accompanying support should be directed there first, where the association with achievement is strongest and most reliably measured. This inverts the deployment a geographic-disadvantage account would suggest, and is the study's most specific contribution to existing policy.

Priority 2 – School climate. No dedicated instrument exists, but none is needed beyond the in-service professional development structures already operating. The teacher support and sense of belonging indices from the student questionnaire can be folded into national assessment cycles as trackable measures, and schools in the lowest quantile on both should receive priority access to development focused on teacher-student interaction and classroom belonging. Because the returns concentrate among the weakest students, this reaches most of the population below Level 2.

Priority 3 – Composition-adjusted allocation and diagnostics. Cabinet of Ministers Resolution No. 646 (2021), which set a precedent for differentiated teacher incentives by location, provides a basis that extends naturally from location to composition. The Ministry of Preschool and School Education should introduce a composition-adjusted performance measure and direct support toward schools performing below expectation once intake is **accounted** for, rather than toward those with the lowest unadjusted scores. Since school composition proves to be the operative factor behind the rural gap, allocation that tracks intake rather than geography follows directly.

Priority 4 — Rebuilding the teacher qualification indicator. This priority is addressed not through a decree but through the ministry's statistical reporting framework. National certification data cannot currently support a valid school-level measure, establishing - one with an unambiguous definition, consistent application across schools and administrative verification – is a precondition for the monitoring architecture described below. Until it exists, teacher quality can be tracked only through the principal-report measure on which Priority 1 relies.

Table 7. Reform priorities – evidence base & Legislative alignment

Priority	Empirical basis	Policy implication	Existing Instrument	Horizon
1. Teacher qualification adequacy	Inadequately qualified staff $\beta = -11.42$, significant in 8 of 10 plausible values; -1957 in urban schools (10/10) against -4.65 in rural schools (0/10)	Concentrate qualification interventions on urban schools, where the association with achievement is strongest and most reliably measured	DP-79 (2023), certification salary supplement	2026–2027

Priority	Empirical basis	Policy implication	Existing Instrument	Horizon
2. School climate investment	Teacher support β = 4.38 and belonging β = 6.15 in the primary model; both largest at the lower quantiles and significant in all ten plausible values across mathematics, reading and science	Direct in-service development on teacher–student interaction and belonging to schools in the lowest quartile on both indices	Existing in-service professional development structures	2027–2028
3. Composition-adjusted allocation	School-mean ESCS β = 24.49, significant in all ten plausible values; rural coefficient falls from −13.52 to −1.21 and loses significance once composition is controlled	Allocate compensatory resources by intake composition rather than location, and adjust school performance measures for intake before use in accountability	Resolution No. 646 (2021), precedent for differentiated allocation	2027–2028
4. Qualification indication reconstruction	Certification measure fails the validity gate: roughly a third of schools report no certified staff and a fifth report full coverage, in schools of comparable size	Establish a certification measure with an unambiguous definition, consistent application and administrative verification before it is used for monitoring	Ministry statistical reporting framework	2026–2027

Note: All coefficients from PISA 2022 Uzbekistan analysis. Timeline reflects sequencing by implementation readiness rather than policy importance.

6.3 Monitoring architecture

The monitoring architecture operationalize Stage 5 of the ERTF by translating reform priorities into measurable indicators with PISA 2022 baselines, annual targets, responsible institutions, and data sources. Table 8 presents the nine-indicator framework.

Three features of this architecture reflect the ERTF's design principles. First, all baseline values are drawn from the PISA 2022 Uzbekistan dataset or from verified motional statistics, no assumed or projected baselines are used. Second, annual targets are calibrated to the government's 2030 PISA top-30 foal established under DP-5712 (2019) and reaffirmed under the Uzbekistan-2030 Strategy (DP-21, 2026) – the framework connects institutional accountability to a declared national target. Third, the indicators are designed to be trackable through administratively available data between PISA cycles, so that progress can be assessed annually rather than only at three-year PISA intervals.

Two indicators are deliberately incomplete. Neither the teacher certification measure nor a composition-adjusted performance measure has no baseline, because the analysis established that no valid baseline can be constructed from existing data. Rather than assign a figure the data cannot support, the architecture records indicator development itself as the 2027 milestone, with tracking beginning once a defensible measure exists. This is a smaller concession than it appears. An indicator built on an unsound measure would produce apparent progress that reflects changes in reporting practice rather than in teacher quality, which is worse than acknowledging the gap.

Table 8. Nine-indicator monitoring framework – PISA 2022 baselines and 2030 targets

Indicator	Baseline (PISA 2022)	2027	2028	2030	Data Source	Responsible institution
Learning Outcomes						
Mean mathematics score	364.5	390	420	460	PISA/national assessment	Ministry of Preschool and School Education
Students below Level 2, mathematics	80.3%	72%	60%	45%	PISA/national assessment	Ministry of Preschool and School Education
Teacher quality						
Schools reporting instruction hindered by unqualified staff	29.1%	22%	15%	8%	PISA school questionnaire, ministry registry	Regional educational departments
Validated teacher certification indicator	None available	Indicator established	Baseline recorded	Target set	Ministry statistical reporting framework	Ministry of Preschool and School Education
School climate						
Mean teacher support index	0.40	0.48	0.56	0.65	PISA/national student survey	Ministry of Preschool and School Education
Mean sense of belonging index	0.12	0.20	0.28	0.35	PISA/national student survey	Ministry of Preschool and School Education
Equity and composition						
Between-school share of achievement variance (ICC)	0.229	0.21	0.19	0.17	PISA/national assessment	Ministry of Preschool and School Education
Dispersion of school-mean socioeconomic composition	0.457	0.43	0.40	0.36	PISA/national assessment	Ministry of Preschool and School Education
Composition-adjusted school performance measure	None available	Measure established	In operational use	Embedded in allocation	Ministry statistical reporting framework	Ministry of Preschool and School Education

Note: Baselines for learning outcomes, school climate and equity indicators are computed from the PISA 2022 Uzbekistan dataset using the final student weight. Source: PISA 2022 international database (OECD, 2023) and author's calculation

6.4 Adaptive policy learning

The final stage of the ERTF – adaptive policy learning – closes the cycle by connecting monitoring outcomes to policy revision. Two checkpoints are built into the framework for Uzbekistan. The PISA 2025 results, expected in late 2026, provide the first empirical test of whether the immediate priorities have begun to move the qualification and climate indicators in the predicted direction. The PISA 2028 results, available in late 2029, provide the last diagnostic before the target year. If the trajectory implied by the nine-indicator framework is not being achieved at either checkpoint, the reform sequencing should be recalibrated using the updated evidence.

The design also responds to a constraint that reform proposals often ignore. The TALIS 2024 finding that 83.4 percent of teachers in Uzbekistan report too many simultaneous reform initiatives,

the highest proportion among 55 participating systems, signals that additional new programs risk amplifying this fatigue. The ERTF therefore does not propose new institutions or new legislative instruments. It provides the first econometrically grounded specification of which existing constraints blind most tightly on student performance in Uzbekistan, and argues that the greatest reform return will come from deepening the implementation of what Presidential decrees have already mandated, particularly DP-79 rather than from introducing additional initiatives.

6.5 Transferability

The specific indicators and targets in Tables 7 and 8 are country-specific, they reflect Uzbekistan's PISA 2022 baseline, its decree framework, and its 2030 reform commitment. The operationalization process itself, however, is generalizable. Any government applying the ERTF to its own PISA, TIMSS, or PIRLS data would follow the same sequence; identify empirically robust findings through HLM and quantile regression; appraise those findings against the four ERTF criteria; align appraised priorities with existing legislative instruments; and construct a monitoring architecture with assessment-derived baselines and target-calibrated annual milestones. The Uzbekistan demonstration is intended as a worked example rather than a template, the indicators, timelines, and legislative alignments will differ across contexts, but the analytical logic connecting them remains constant.

7. Discussion

7.1 Interpreting the empirical findings in context

The null HLM result ICC=0.229 establishes the analytical foundation for everything that follows. That 22.9 percent of total mathematics score variance in Uzbekistan lies between schools, rather than between students within the same school, confirms that school-level institutional conditions matter independently of individual student background. This figure is below the OECD-average school share but well above the 5% modelling threshold; the actionable levers nonetheless sit at Level 2.

The identification of UNQUALSTAFF as the operative school-level constraint (β = -11.42, p <0.05) is the central empirical finding. Staff quantity shortage, certification coverage, student-teacher ratio are all non-significant in the full model which distinguishes the quality of teacher preparation from its quantity and credentialling as the binding institutional constraint. This finding parallels Rivkin, Hanushek, and Kain (2005), who showed that observable teacher credentials

explain little of the variation in teacher effectiveness, and extends their US-based results to a Central Asian context where constraint is not simply credential holding but principal-reported instructional hindrance. The geographic concentration of this penalty β = -19.57 in urban schools, significant in all ten plausible values, against β = -4.65 in rural schools, significant in none, indicates that the same reported constraint carries substantially different consequences depending on institutional setting, and that pooling the two dilutes an effect which is robust in one of them.

The negative RURAL coefficient in the primary HLM (β = -13.52, $p<0.01$) initially appears to establish a rural school-quality deficit. The extended HLM resolves this. Once school-level mean ESCS is added as a compositional control, the RURAL coefficient collapses to -1.21 and becomes entirely non-significant across all ten plausible values. Rural schools in the PISA 2022 Uzbekistan sample serve substantially more socioeconomically disadvantaged students (mean ESCS_school = -0.83) than urban schools (mean_ESCS_school = -0.35), and once like-for-like schools are compared, no genuine rural-urban performance differential remains. The apparent rural deficit is therefore compositional rather than institutional. This finding has direct implications for resource allocation: compensatory resources should follow intake composition rather than location, since geography operates here as a proxy for disadvantage rather than an independent source of it.

The school climate findings from the quantile regression add an equity dimension that the HLM cannot reveal. Teacher support (TEACHSUP) carries a coefficient of 5.62 at Q 10 among the weakest students declining to 3.39 at Q90. School belonging (BELONG) shows a parallel and monotonic pattern, from 7.79 at Q10 to 4.70 at Q90. This declining gradient confirms that school climate improvements are equity-enhancing rather than equity-neutral: their largest returns accrue precisely among the 80.3 % of students currently below the PISA Level 2 threshold. Shen (2026) finds the same pattern for Shanghai-China using PISA 2012 data, Liu et al. (2024) document it across ten European countries over two PISA cycles - the equity-enhancing distributional property of instructional climate effects appears to be a general feature of school-level interventions rather than a context-specific finding.

Socioeconomic status shows the opposite distributional pattern steepening from β = 4.80 at Q10 to β = 11.15 at Q90. Family background is a modest predictor among the weakest students but a powerful differentiator at the top of the distribution. This asymmetry implies that policies targeting socioeconomic inequality will generate their largest returns among high-performing students - the opposite of where Uzbekistan's reform urgency lies. The implication is not that socioeconomic

inequality can be ignored but that school-level instructional interventions offer a more tractable and equitably distributed lever for reducing the 80.3% below Level 2.

7.2 The ERTF as an advance in evidence-informed governance

The more fundamental contribution of this study is methodological. The transition from the statistical results in Section 5 to the reform priorities in Section 5.10 is not automatic, it is mediated by an explicit appraisal process that evaluates each finding against four criteria before elevating it to policy priority status. This process reveals consequential distinctions that regression tables alone cannot make. UNQUALSTAFF, TEACHSUP and BELONG score high on all four criteria and become Priority 1 and Priority 2 interventions. ESCS scores high on robustness and magnitude but low on feasibility, it becomes a background condition that motivates equity targeting rather than a direct reform lever. STAFFSHORT is significant across the achievement distribution but with a positive sign that almost certainly reflects excess demand in growing schools rather than any benefit of understaffing, and a naive reading would have supported an indefensible recommendation. CERTTEACH is excluded despite its apparent policy relevance. Its coefficients are well behaved, but underlying measure is bimodal in a way that indicates inconsistent interpretation rather than genuine variation, and it is excluded from appraisal rather than read as evidence that certification does not matter.

This explicit appraisal process addresses what Nutley, Walter, and Davies (2007) identify as the central challenge of evidence utilization: the gap between knowing what matters and knowing what to do about it. The existing PISA literature is extensive on the former and limited on the latter. Vanlommel and van den Boom-Muilenburg (2024) argue that evidence-informed educational change requires not better evidence production but better institutional mechanisms for using existing evidence. The ERTF provides one such mechanism, and the Uzbekistan application demonstrates that it produces prioritization decisions that differ meaningfully from a naive reading of the regression coefficients.

The monitoring architecture in Stage6 of the ERTF addresses a second gap. Stewart et al. (2022) observe that the evidence-informed policy movement has focused heavily on evidence generation while comparatively little attention has been paid to how implementation should be tracked after reforms are introduced. The nine-indicator monitoring framework developed in Section 6 of this paper with baselines drawn from PISA 2022 data and annual targets calibrated to the 2030 trajectory, make this tracking concrete rather than aspirational. PISA 2025 results, expected in late

2026, will provide the first empirical test of whether the reforms introduced in response to PISA 2022 are producing measurable gains on these indicators.

7.3 Limitations

Four limitations frame the interpretation of these results.

The cross-sectional design prevents causal interpretation - the associations estimated through HLM and quantile regression are robust across specifications but cannot establish that reducing UNQUALSTAFF would raise scores by the implied magnitude. Longitudinal data linking PISA cycles to administrative records on certification uptake would enable the difference-in-differences analysis needed for causal claims.

The analytical analysis is not nationally representative. Complete-case restriction removes roughly a third of students, and those excluded score 33 points below those retained. The missingness arises principally from PISA's rotated questionnaire design rather than from refusal, which makes it more tractable than non-response would be, but it leaves the estimates describing a somewhat stronger group of students than the national cohort. The direction of the resulting bias is known: because the weakest students are disproportionally absent, associations estimated at the lower end of the distribution are more likely to be weakened than inflated, and the equity findings in Section 5.7 should accordingly be read as conservative. Multiple imputation would retain more of the sample and is the obvious extension.

Estimation involved two compromises that a fuller treatment would avoid. Sampling weights are applied at the student level only, since PISA supplies no school-level weight suited to multilevel estimation, and weighting at a single level may bias estimates where inclusion probabilities differ systematically between levels. The quantile models are unweighted altogether, because, the estimator does not accept probability weights, and their bootstrap standard errors do not reflect the replication design. The quantile results are accordingly read alongside the weighted multilevel estimates rather than in place of them. The consistency of the two gives some reassurance, though not a substitute for correctly weighted distributional estimation.

The PISA 2022 school questionnaire does not capture several structural factors specific to Uzbekistan. Textbook revision frequency, teacher administrative burden, and curriculum instability may independently affect performance, and their omission limits the completeness of the school-level model.

The certification measure deserves separate discussion. Its failure at the validity gate means the analysis can say nothing about whether formal credentials predict achievement in Uzbekistan. The null coefficients reported in Table 4 should not be read as evidence that certification is irrelevant.

The ERTF remains partially conceptual in the implementation and adaptive learning stages - future research should examine whether governments that apply the framework produce better implementation outcomes than those that rely on conventional approaches, and what institutional conditions support its adoption.

8. Conclusion

This study addressed a challenge that sits at the heart of contemporary education governance: the persistent gap between producing robust empirical evidence and translating it systematically into reform. International large-scale assessments now generate more evidence on the determinants of student achievement than in any previous period, yet the pathway from regression coefficients to implementable policy remains poorly mapped. The Evidence-to-Reform Translation Framework proposed here offers a structural approach for navigating that pathway.

Applied to PISA 2022 data from Uzbekistan, the framework produces findings that go beyond what statistical analysis alone can support. The hierarchical linear models confirm that 22.9 percent of total mathematics score variance lies between schools, a figure that establishes the institutional rather than purely individual character of Uzbekistan's learning challenge. The primary binding constraint at the school level is inadequately qualified teaching staff ($\beta = -11.42$, significant in eight of ten plausible values), with the penalty roughly four times larger in urban schools, where it is robust across all ten. The quantile regression adds an equity dimension invisible to average-effect models, teacher support and sense of belonging generate their largest return among the 80.3 percent of students below the minimum proficiency threshold, establishing school climate investment as equity-enhancing rather than equity-neutral. The extended HLM resolves the apparent rural performance deficit as a compositional artefact, redirecting attention from location to disadvantage wherever it is concentrated.

One finding emerged from the appraisal rather than estimation. The teacher certification measure produced stable and well-behaved coefficients that, read conventionally, would have supported a substantive conclusion about credentials and achievement. Inspection of its distribution showed instead that the variable did not measure what its name claimed. Screening for measurement

validity before appraisal turned what would have been a quiet error into an explicit finding, and into a reform priority: an indicator that cannot support analysis, cannot support monitoring either, and Uzbekistan's 2030 targets presuppose on that does not yet exist.

The ERTF operationalisation in Section 6 demonstrates that these findings connect directly to Uzbekistan's existing legislative framework without requiring new institutions or new budget lines. Presidential decree No. DP-79 (2023), which established the teacher certification bonus, is the empirically indicated instrument; the contribution of this study is to specify where its implementation should be concentrated. The qualification penalty is roughly four times larger in urban schools, where it is robust across all ten plausible values, and statistically undetectable in rural ones, which locates the immediate return to certification investment in urban districts. The nine-indicator monitoring framework makes the 2030 reform trajectory measurable annually using administrative data, with PISA 2025 and PISA 2028 results serving as built-in empirical checkpoints for realignment.

Three contributions follow from this study. The paper extends educational effectiveness research by moving from identifying determinants to specifying how they should inform reform decisions. It operationalizes evidence-informed governance by connecting statistical findings to explicit appraisal criteria, legislative alignment, and a monitoring architecture. It contributes to implementation science by embedding adaptive learning within the reform cycle, recognizing with Vanlommel and van den Boom-Muilenburg (2024), that sustainable educational improvement depends on continuous institutional learning rather than the one-time launch of well-designed programs.

The ERTF is designed to be transferable. Its appraisal logic applies wherever governments face the task of converting assessment evidence into reform action - TIMSS, PIRLS, national assessments, or any dataset combining achievement data with contextual information on schools, teachers, and learning environments. The Uzbekistan case demonstrates the framework in a specific institutional context; the analytical process it embeds is generalizable. As education systems worldwide pursue improvements in quality and equity under fiscal constraints, the ability to translate evidence systematically into prioritized, monitored, and adaptable reform strategies will become increasingly consequential. This framework offers a concrete methodology for that task, together with a reminder that the first question to ask of any indicator is whether it measures what it claims.

Data availability
The datasets analysed in this study are publicly available from the OECD PISA 2022 international database at https://www.oecd.org/pisa/data/ (OECD, 2023). The Stata do-files used to construct the analytical dataset and estimate all reported models are available from the author upon request.

Competing interests
The author declares no competing interests.

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

Author contributions
The study was designed, conducted and written by a single author, who constructed the dataset, carried out the analysis, interpreted the findings and prepared the manuscript. The author accepts responsibility for the integrity of the data and the accuracy of the analysis.

Ethics approval
This study used publicly available, fully anonymised secondary data released by the OECD. The PISA 2022 database contains no individual student or school identifiers. No direct contact with human participants was involved, and ethical approval was not required.

Informed consent
This article does not contain any studies with human participants performed by the author.

Reference

References

Cairney P (2016) The politics of evidence-based policy making. Palgrave Macmillan, London

Crosnoe R (2009) Low-income students and the socioeconomic composition of public high schools. Am Sociol Rev 74(5):709–730

Hanushek EA, Woessmann L (2015) The economic impact of educational quality. In: McGrath S, Gu Q (eds) Handbook of international development and education. Edward Elgar, Cheltenham, pp 6–19

Head BW (2010) Reconsidering evidence-based policy: key issues and challenges. Policy Soc 29(2):77–94. https://doi.org/10.1016/j.polsoc.2010.03.001

Jerrim J, Lopez-Agudo LA, Marcenaro-Gutierrez OD, Shure N (2017) What happens when econometrics and psychometrics collide? An example using the PISA data. Econ Educ Rev 61:51–58. https://doi.org/10.1016/j.econedurev.2017.09.007

Klieme E, Pauli C, Reusser K (2009) The Pythagoras study: investigating effects of teaching and learning in Swiss and German mathematics classrooms. In: Janik T, Seidel T (eds) The power of video studies in investigating teaching and learning in the classroom. Waxmann, Münster, pp 137–160

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50. https://doi.org/10.2307/1913643

Koenker R, Hallock KF (2001) Quantile regression. J Econ Perspect 15(4):143–156. https://doi.org/10.1257/jep.15.4.143

Li J, Xue E, Guo S (2025) The effects of PISA on global basic education reform: a systematic literature review. Humanit Soc Sci Commun 12:106. https://doi.org/10.1057/s41599-025-04403-z

Liu X, Hansen KY, Valcke M, De Neve J (2024) A decade of PISA: student perceived instructional quality and mathematics achievement across European countries. ZDM Math Educ 56(5):859–891. https://doi.org/10.1007/s11858-024-01630-7

Niu J, Xu H, Yu J (2025) Identifying multilevel factors on student mathematics performance for Singapore, Korea, Finland, and Denmark in PISA 2022: considering individualistic versus collectivistic cultures. Humanit Soc Sci Commun 12:151. https://doi.org/10.1057/s41599-025-04466-y

Nutley SM, Walter I, Davies HTO (2007) Using evidence: how research can inform public services. Policy Press, Bristol

OECD (2023) PISA 2022 Database. Organisation for Economic Co-operation and Development, Paris. https://www.oecd.org/pisa/data/. Accessed 28 January 2026

OECD (2023a) PISA 2022 results (volume I): the state of learning and equity in education. OECD Publishing, Paris. https://doi.org/10.1787/53f23881-en

OECD (2023b) PISA 2022 results (volume II): learning during – and from – disruption. OECD Publishing, Paris. https://doi.org/10.1787/a97db61c-en

OECD (2024) PISA 2022 technical report. OECD Publishing, Paris. https://doi.org/10.1787/01820d6d-en

OECD (2025a) New PISA results coming soon. Organisation for Economic Co-operation and Development, Paris.

https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/announcements/New_PISA_results_coming_soon.pdf. 12 February 2026

OECD (2025b) Results from TALIS 2024: the state of teaching. OECD Publishing, Paris. https://doi.org/10.1787/90df6235-en

Pholphirul P, Rukumnuaykit P, Teimtad S (2023) Teacher shortages and educational outcomes in developing countries: empirical evidence from PISA-Thailand. Cogent Educ 10(2):2243126. https://doi.org/10.1080/2331186X.2023.2243126

Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, 2nd edn. Sage, Thousand Oaks

Rivkin SG, Hanushek EA, Kain JF (2005) Teachers, schools, and academic achievement. Econometrica 73(2):417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Shen T (2026) Teaching effectiveness on secondary mathematics: evidence from PISA—Shanghai-China. Educ Res Policy Pract. https://doi.org/10.1007/s10671-025-09412-y

StataCorp (2021) Stata statistical software: release 17. StataCorp LLC, College Station

Stewart R, Dayal H, Langer L, van Rooyen C (2022) Transforming evidence for policy: do we have the evidence generation house in order? Humanit Soc Sci Commun 9:116. https://doi.org/10.1057/s41599-022-01132-5

Thrupp M, Lauder H, Robinson T (2002) School composition and peer effects. Int J Educ Res 37(5):483–504. https://doi.org/10.1016/S0883-0355(03)00016-8

UNESCO (2023) Global education monitoring report 2023: technology in education — a tool on whose terms? UNESCO, Paris. https://doi.org/10.54676/UZQV8501

Vanlommel K, van den Boom-Muilenburg SN (2024) How can we understand and stimulate evidence-informed educational change? A scoping review from a systems perspective. J Educ Change 25(3):605–634. https://doi.org/10.1007/s10833-024-09506-z

Volante L, Mattei P (2024) The politicization of PISA in evidence-based policy discourses. Policy Futures Educ 22(8):1554–1569. https://doi.org/10.1177/14782103241227309

von Davier M, Gonzalez E, Mislevy R (2009) What are plausible values and why are they useful? IERI Monogr Ser 2:9–36

Willms JD (2010) School composition and contextual effects on student outcomes. Teach Coll Rec 112(4):1008–1037

Zhang F, Jin C, Fan C (2024) Factors influencing students' reading literacy in Morocco: a multilevel analysis. Humanit Soc Sci Commun 11:1056. https://doi.org/10.1057/s41599-024-03575-4

Legislative and policy instruments

Republic of Uzbekistan (2019) On approval of the concept for developmentof the public education system of the Republic of Uzbekistan until 2030. Presidential Decree No. PF-5712, 29 April 2019. National Database of Legislation, Tashkent. https://lex.uz/en/docs/6968779 Accessed 01 March, 2026

Republic of Uzbekistan (2021) On measures to attract qualified teachers and educators to schools with low education level. Cabinet of Ministers Resolution No. 646, 15 October 2021. National Database of Legislation, Tashkent. https://lex.uz/docs/7465994 Accessed 01 March, 2026

Republic of Uzbekistan (2023) On measures to effectively organize the activities of the Ministry of Preschool and School Education and organisations in its system. Presidential Decree No. PF-79, 26 May 2023. National Database of Legislation, Tashkent. https://lex.uz/docs/7187245 Accessed 01 March, 2026

Republic of Uzbekistan (2026) On additional measures to consistently continue reforms and bring them to a new stage within the framework of priority directions of the country's development until 2030. Presidential Decree No. PF-21, 16 February 2026. National Database of Legislation, Tashkent. https://lex.uz/docs/8052981 Accessed 01 March 2026

Figures

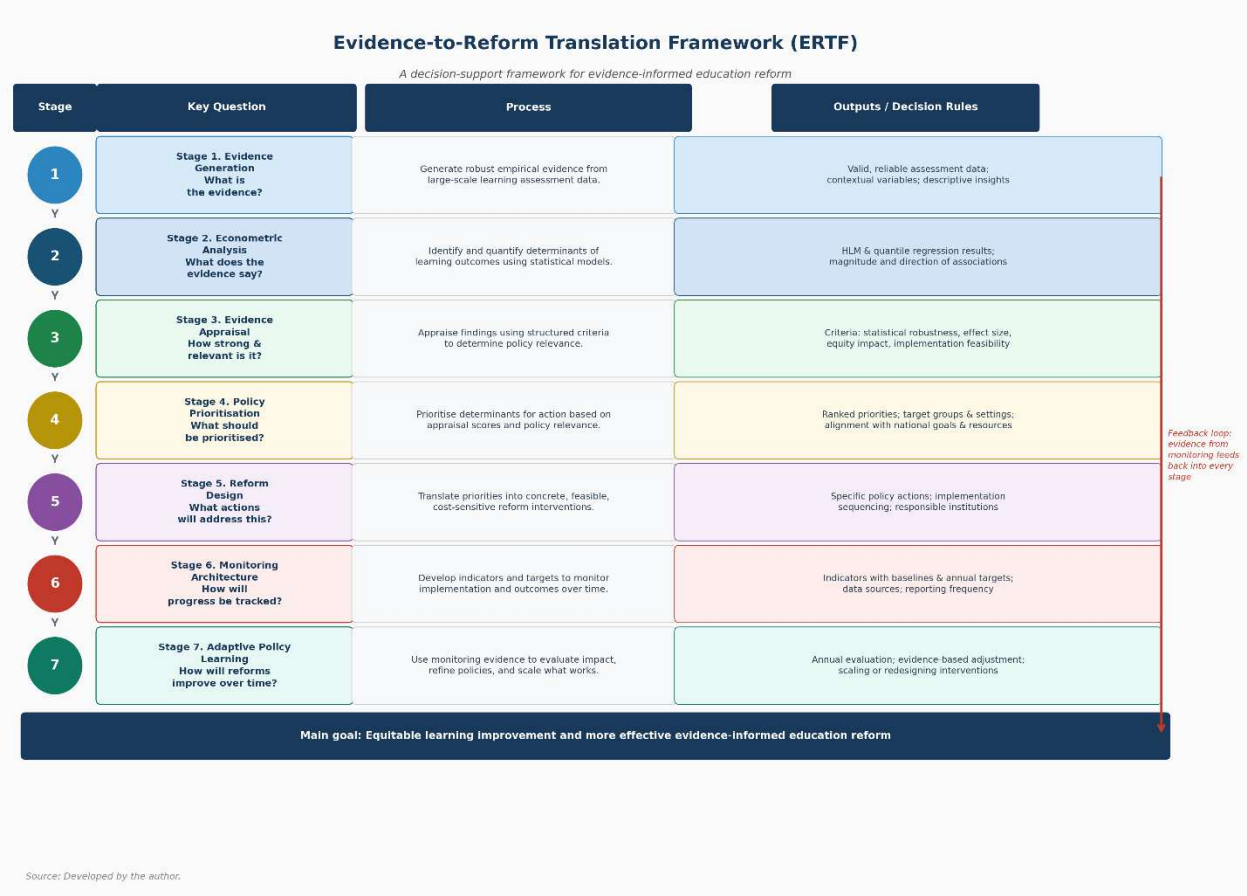


Fig. 1 The Evidence-to-Reform Translation Framework (ERTF). The framework consists of seven interconnected stages that transform large-scale assessment evidence into an adaptive cycle of evidence-informed education reform. Each stage builds on the outputs of the previous stage while generating information that supports subsequent policy decisions. The feedback loop indicates that monitoring evidence from Stages 6 and 7 feeds back into earlier stages to strengthen future reform decisions. Source: Developed by the author.

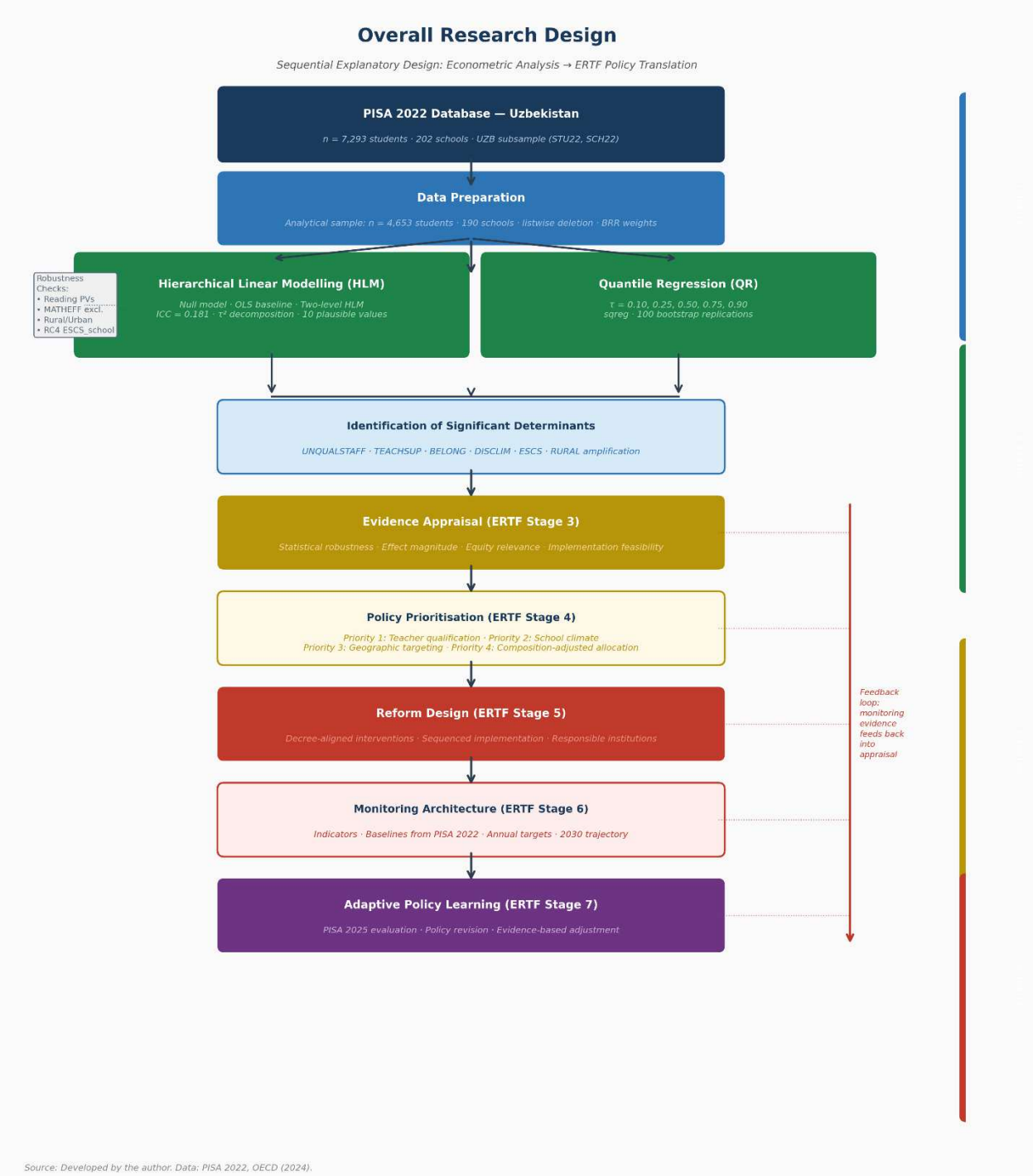


Fig. 2 Overall research design. The diagram illustrates the sequential analytical pathway from the PISA 2022 database through data preparation, hierarchical linear modelling, quantile regression, evidence appraisal, policy prioritisation, and education reform recommendations. Source: Developed by the author.

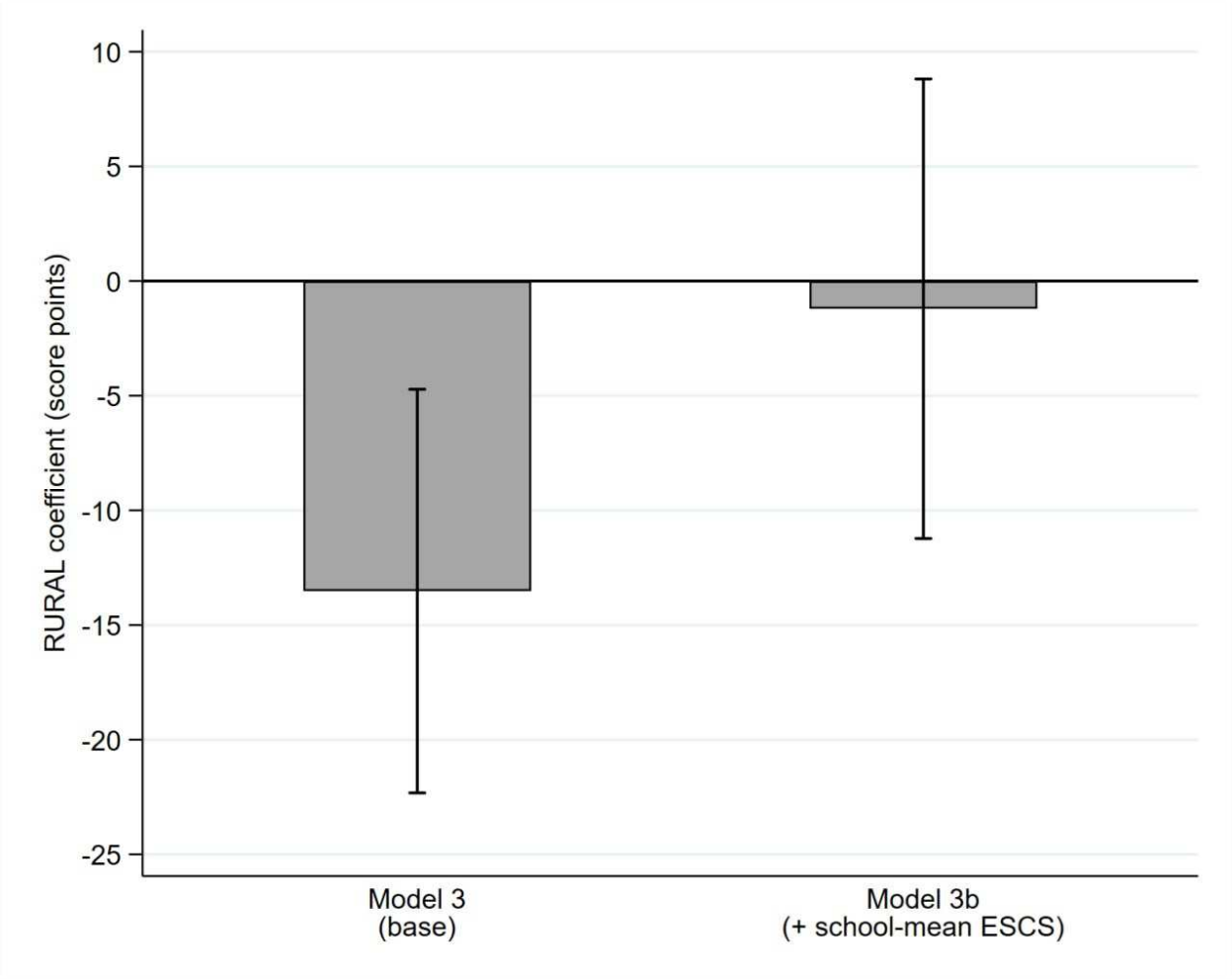


Fig.3 The rural coefficient before and after compositional control. Bars show the coefficient on rural school location in the base multilevel model and in the extended model including school-level mean socioeconomic composition, with approximate 95% confidence intervals averaged across ten plausible values. The rural performance deficit does not survive the addition of the compositional control, indicating that it reflects differences in student intake rather than in school effectiveness.
Source: PISA 2022 international database (OECD, 2023) and author's calculations.

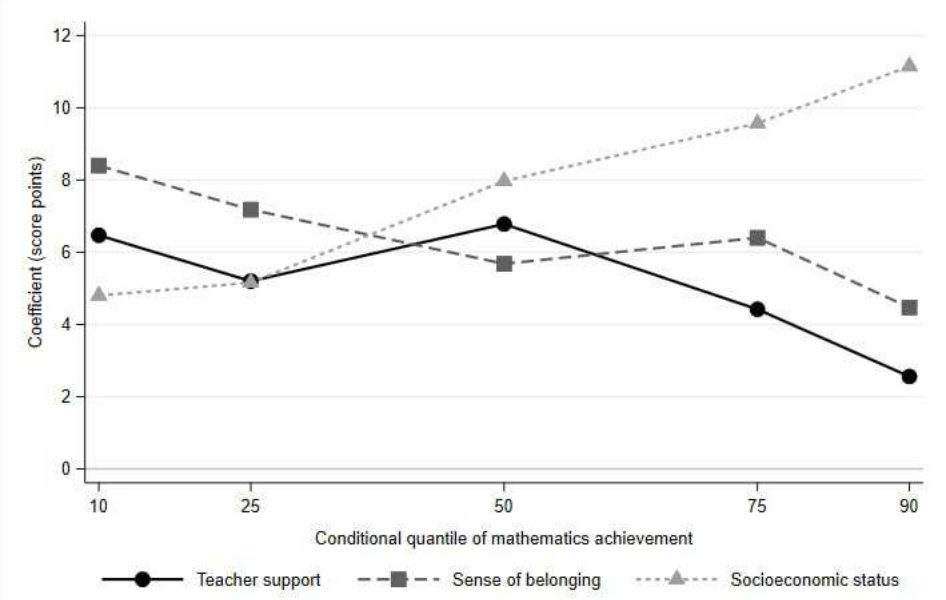


Fig.4 Quantile regression coefficients across the mathematics achievement distribution. Teacher support and sense of belonging carry their largest associations in the lower and middle distribution, while the socioeconomic gradient steepens toward the upper quantiles. School climate is associated with larger differences among the weakest students; socioecnomic advantage differentiates most among the strongest. Coefficients are from the first plausible value and correspond to those reported in Table 6; ten-value avergaes show the same pattern. Source: PISA 2022 international database (OECD, 2023) and author's calculations.

Supplementary Files

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

- SupplementaryInformationERTFUzbekistanPISA2022.docx