Studies in Educational Evaluation 88 (2026) 101570

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Curriculum standardization, central exams and achievement inequalities in primary education: A cross-national longitudinal study

Nathalie Vigna *, Giuliana Parente, Moris Triventi

University of Milan, Italy

ARTICLE INFO

Keywords:
Education inequalities
Education policy
Curriculum standardization
PIRLS data
Cross-national comparison
Longitudinal study

ABSTRACT

This study examines whether standardization increases academic performance and reduces achievement inequalities in primary education. We analyze standardization from a multidimensional perspective, considering educational inputs, such as the national curriculum, and outputs, such as central exams.

We use data from the Progress in International Reading Literacy Study (PIRLS) for 2006, 2011, and 2016, covering 43 countries and nearly 700,000 fourth-grade students. For each country and year, we construct a novel index of curriculum standardization and examine the presence of central exams. Using an innovative two-step multilevel meta-regression strategy, we assess the role of standardization in shaping average reading performance and multiple measures of test score inequality.

Results reveal substantial cross-national differences in standardization practices but no consistent relationship between curricular or assessment standardization and reading proficiency or inequality. Although within-country temporal variation is limited, findings remain robust across specifications, challenging the assumption that standardization alone promotes educational equity or effectiveness in primary education.

1. Introduction

Education systems face the dual challenge of promoting excellence while ensuring equity (Brint, 2017). Beyond individual and family characteristics, research increasingly highlights how institutional features shape outcomes and disparities (Skopek et al., 2019; Van de Werfhorst & Mijs, 2010). This shift from micro to system-level explanations has drawn attention to school autonomy, standardization, tracking, teacher quality, and resource allocation as key levers for equity and effectiveness (Chiu & Chow, 2015; OECD, 2016).

Within this constellation of institutional features, standardization has been particularly prominent in secondary education, where it is often linked to various educational outcomes (Almendinger, 1989; Checchi et al., 2014; Van de Werfhorst & Mijs, 2010). Whether similar relationships hold in primary education remains unclear. We therefore ask: How do curriculum standardization and central examinations vary across countries and over time at the primary level? And are higher levels of standardization—understood as uniform content and central examinations—associated with higher average performance and reduced inequality?

We conceptualize standardization as multidimensional, encompassing both inputs (nationally prescribed curricula) and outputs (centralized, uniform assessments; Bishop, 1999). Curriculum standardization reflects the extent to which content, goals, and materials are defined nationally and applied consistently. Assessment standardization refers to the degree to which student performance is evaluated against external standards (Bol & Van de Werfhorst, 2013). Central exams cover the entire student population and can be implemented at either the national or regional level.

Highly standardized systems aim to equalize exposure to content and ensure comparability in evaluations, but they also face a familiar tension between uniformity and flexibility. Standardization can reduce differences in curricular quality and "level the playing field" (Baker, 2012; Stevenson & Baker, 1991), yet rigid frameworks may limit local adaptation, differentiation, and innovation, potentially amplifying inequalities (Remillard, 2005; Zhao et al., 2023). These tensions mirror broader debates about the equity–efficiency trade-off in education (Dubet, 2004; Hanushek & Woessmann, 2006).[1] Indeed, a uniformly

* Corresponding author.
E-mail address: nathalie.vigna@unimi.it (N. Vigna).

[1] In simple terms, *equity* refers to the fairness of educational opportunities and outcomes, ensuring that students from different socioeconomic, cultural, or geographic backgrounds have comparable chances of success. *Efficiency*, in contrast, concerns the system's capacity to maximize overall educational performance, often measured through average student achievement, cost-effectiveness, or international competitiveness.

https://doi.org/10.1016/j.stueduc.2026.101570
Received 31 July 2025; Received in revised form 15 December 2025; Accepted 23 January 2026
Available online 7 February 2026

enforced curriculum can equalize opportunities—especially where regional or socioeconomic gaps are large—by giving all students the same content and expectations, potentially boosting both equity and minimum performance. However, over-standardization can reduce flexibility, limiting teachers' ability to tailor instruction to diverse needs; this risks underserving both high achievers and disadvantaged students and can ultimately lower overall efficiency.

Two gaps motivate our study. First, comparative work often conflates or sidelines the distinct dimensions of curriculum and assessment standardization, with few studies examining their separate contributions to educational outcomes. Second, we lack longitudinal, cross-national evidence on how these dimensions relate to achievement and inequality in primary schooling, a period when foundational skills are formed and early interventions are most consequential (Alexander et al., 2007).

We address these gaps using Progress in International Reading Literacy Study (PIRLS) data (2006–2016) on fourth-grade reading literacy. We separately conceptualize and measure curriculum standardization (inputs) and assessment standardization (outputs), drawing on the PIRLS curriculum questionnaires and information on the presence of national exams in primary education. This design allows us to assess not only whether standardization matters but also which dimension—content or assessment—is more strongly associated with equity and performance.

Our contributions are fourfold: (1) we introduce a continuous index of curriculum standardization based on the PIRLS curriculum questionnaires, moving beyond binary centralization indicators and considering it alongside the presence of central examinations; (2) we focus on primary education, often overlooked in institutional research yet crucial for the early equalization of opportunities; (3) we adopt a multi-outcome perspective, examining average performance and multiple measures of inequality (e.g., socio-economic gradients, percentile gaps); and (4) we implement an innovative multilevel meta-regression approach to repeated cross-sectional international data, leveraging PIRLS's hierarchical structure and within-country changes in standardization levels.

The article proceeds as follows. First, we present the theoretical framework and hypotheses, situating our study within prior literature and its limitations. Next, we describe the data, variables, and methods. We then report descriptive patterns and core model results. Finally, we discuss the findings and conclude.

2. Theoretical framework

2.1. Conceptualizing standardization

Standardization is a key institutional feature in education systems and has been the focus of extensive theoretical and empirical work. It is commonly defined as the extent to which educational inputs and outputs are uniform across schools and regions, ensuring that students are exposed to similar learning experiences and evaluated according to common standards (Allmendinger, 1989; Bishop, 1999). In contrast to broader concepts such as centralization—which refers to where decisions are made—standardization captures the degree of consistency in what is taught and how student performance is assessed (Van de Werfhorst & Mijs, 2010).

Building on these distinctions, scholars have proposed a multidimensional approach to standardization encompassing both input standardization (e.g., curriculum, teacher training, instructional materials) and output standardization (e.g., central examinations, national assessments, uniform evaluation criteria; Bol & Van de Werfhorst, 2013). This conceptualization allows for the empirical investigation of how specific mechanisms of standardization—rather than broad education system types—affect student achievement and educational inequality.

2.2. Why standardization might be beneficial

Curriculum standardization refers to the establishment of a common, centrally defined body of knowledge and skills that students are expected to learn. A standardized curriculum can reduce variation in what is taught across schools and mitigate the influence of local discretion or teacher bias. It fosters instructional coherence and can enhance the predictability and transparency of educational expectations, particularly for disadvantaged students (Baker, 2012; Stevenson & Baker, 1991).

Assessment standardization, typically implemented through national or external exams, reinforces curriculum goals and ensures consistent criteria for evaluating student achievement. Central exams reduce variation in grading standards, limit teacher discretion, and provide more objective signals of student competence (Jürges et al., 2005). These features have the potential to increase equity—by ensuring comparability across contexts—and accountability, by aligning school practices with shared performance expectations (Ayalon & Livneh, 2013; Van de Werfhorst & Mijs, 2010).

Moreover, standardization may influence inequality through its effects on teacher behavior, instructional practices, and the allocation of learning time. Teachers in standardized systems are more likely to align their teaching with curriculum objectives and external assessments, which can reduce opportunities for informal tracking within classrooms and help keep disadvantaged students on pace with their peers (Horn, 2009; Schmidt & Prawat, 2006).

In sum, following these lines of reasoning, one can expect that more standardized curricula and the presence of central exams will be positively associated with mean reading achievement (Hypothesis 1), will reduce interindividual variation in student outcomes (Hypothesis 2), and will attenuate the strength of the relationship between ascriptive characteristics (e.g., gender, socioeconomic background) and achievement (Hypothesis 3).

2.3. Empirical literature review

Empirical research generally supports the idea that standardization—at least in secondary education—can promote educational equality without necessarily compromising performance. A key early study by Gamoran (1996) on the Scottish education system found that aligning school curricula with national exams in secondary education reduced the impact of socioeconomic status on student achievement and increased homogeneity in school-level performance. Similarly, Montt (2011), using PISA data on 15-year-old students, showed that countries with more standardized curricula and assessments in upper secondary education tend to display lower levels of inequality in student achievement, particularly among low-performing students.

Evidence from PISA further suggests that curriculum and assessment standardization are associated with narrower socioeconomic gradients and higher overall achievement (Bishop, 1997, 1999; Bol & Van de Werfhorst, 2013). Countries such as Finland and Canada, which combine a national curriculum with comprehensive assessment systems and strong teacher support, often score highly in both equity and performance, although they also have many other societal and school-level characteristics that may foster educational achievement and reduce inequalities.

Some evidence also suggests that standardization in secondary education contributes to reducing gender inequalities, particularly in STEM fields. By limiting variation in exposure to content and grading, standardized curricula and exams can reduce gendered patterns of subject choice and performance (Ayalon & Livneh, 2013; Han & Buchmann, 2016). However, the robustness of positive effects from central examinations, based on cross-sectional international survey data, has been questioned due to the likely presence of important unobserved variables at the country level (Jürges & Schneider, 2004).

2.4. Limitations of existing research

Despite the breadth of research on standardization, several limitations remain. First, most existing studies are cross-sectional and do not account for longitudinal dynamics or changes over time. This limits the ability to determine whether the association between standardization and national-level achievement outcomes reflects a causal effect or is instead driven by other unobserved country characteristics (Van de Werfhorst & Mijs, 2010). As Jürges et al. (2005) point out, we must consider the possibility that countries or federal states with central examinations place greater importance on education and academic achievement. In this case, both high average student achievement and the presence of central examinations might simply reflect the higher value placed on education by the electorate in these states or regions.

Second, most empirical studies have focused on secondary education and subjects such as mathematics and science, where performance gaps are more pronounced. Far less attention has been paid to early education and literacy, even though inequalities that emerge during the primary school years often compound over time and shape subsequent educational trajectories (Kulic et al., 2019).

3. Analytical strategy

3.1. Data sources

We used data from the PIRLS, a repeated cross-sectional survey conducted every five years by the International Association for the Evaluation of Educational Achievement (IEA). PIRLS aims to assess the reading literacy of fourth-grade students worldwide. In particular, we focused on the 2006, 2011, and 2016 waves, due to the availability and consistency of key information on both student achievement and curricular features across these cycles. These three waves provide the most comparable data in terms of assessment design, background questionnaires, and curriculum indicators, enabling robust longitudinal and cross-national analyses.

PIRLS employs a two-stage stratified cluster sampling design. In the first stage, schools are randomly selected with probabilities proportional to size within defined strata (e.g., regions or school types). In the second stage, intact fourth-grade classrooms are randomly selected within each sampled school, and all students in the selected class are eligible to participate. Sampling weights are provided at each level to ensure national representativeness and adjust for differential participation. PIRLS test scores capture students' reading literacy performance, assessed through reading passages and comprehension questions designed to evaluate skills in retrieving, interpreting, and critically reflecting on written texts.

The full sample comprises about 680,000 students across 122 country-year units from 47 countries that participated in at least two of the three waves. Table A1 in Appendix A reports the total available observations by country and year. For the multivariate analyses, to avoid non-representative samples due to excessive missingness, we retained 89 country-year units (≈430,000 students). We estimated models only for country-year samples with less than 30 % missing data on key individual variables, applying this threshold at the country-year level. Table A2 in Appendix A details the proportion of missing values for the individual variables and indicates which country-year samples were dropped.

Reading achievement is estimated under a matrix-sampling design: Each student answers only a subset of items, and proficiency is inferred using an item-response model. To reflect measurement uncertainty, PIRLS supplies five plausible values per student—multiple imputations of latent reading proficiency—which are analyzed in parallel and combined following PIRLS guidelines. Individual scores range from about 5 to about 830 points, with most students scoring between 300 (5th percentile) and 650 (95th percentile). The score scale was established in 2001 and maintains comparability of results across educational systems and over time.

We relied on data from several sources. First, the individual-level data include reading test scores (plausible values), student background characteristics (gender, language spoken at home), and responses to the home questionnaire completed by parents (parental education). Second, country-level data are drawn from the PIRLS curriculum questionnaires completed by national research coordinators. We also supplemented these data with additional time-varying contextual indicators, such as economic development and educational spending, derived from external sources (see below).

3.2. Variables and operationalization

The study investigated three types of educational outcomes: (a) average achievement (effectiveness), (b) interindividual inequality, and (c) group-based inequality.

(a) Average achievement was measured by the mean reading score, based on the five plausible values provided by PIRLS.
(b) Interindividual inequality was captured by two measures: the interquartile range, calculated as the difference between the 75th and 25th percentiles of the score distribution; and educational poverty, defined as the proportion of students scoring below 400 points on the reading test. This lowest benchmark, used in PIRLS reports on achievement levels, indicates that students can retrieve explicitly stated information and make straightforward inferences when reading the simplest texts (approximately corresponding to the bottom 10 % of the international distribution).
(c) Group-based inequality was assessed by calculating average score differences between socio-demographic categories, focusing on gender (boys vs. girls) and parental education. Parental education is a binary variable distinguishing between students with at least one tertiary-educated parent (high parental education) and students with no tertiary-educated parent (low parental education). The distribution of these variables in each country and year is shown in Table A3 of Appendix A.

We used two key independent variables that capture curriculum standardization and central exams. The first is an index of curriculum standardization, constructed from the PIRLS national curriculum questionnaires. The index includes five items related to the organization and content of the reading curriculum: (1) whether there is a national curriculum covering reading instruction; (2) whether reading is treated as a separate curriculum area; (3) whether the curriculum includes defined goals and objectives; (4) whether it prescribes instructional methods; and (5) whether it prescribes instructional materials. Each item is coded as 1 (yes), 0.5 (varies), or 0 (no), and the sum is standardized to produce a continuous index per country-year.[2] Table 1 shows the variables used to construct the index, the specific wording of each question, and the weight assigned to each response.[3]

This information originally came from the curriculum datasets provided by PIRLS. However, the original data presented several inconsistencies, mainly due to different interpretations of the questions by

[2] The option "Varies" was primarily used to indicate the existence of state- or province-level curricula in federal countries (Canada, Australia, Germany, and the United States). If the state or provincial curricula also specified goals and objectives, the option "Varies" was likewise chosen for the item on goals and objectives.

[3] We also explored an alternative index based on factor analysis using polychoric correlations, which is appropriate for categorical items. However, the combination of very high inter-item correlations and limited response variability produced near-singular correlation matrices and unstable or indeterminate factor solutions across waves. Consequently, we discontinued this approach and retained the additive specification.

Table 1
The items used to build the indexes of curriculum standardization and curriculum evaluation (NB: the specific wording of the questions may vary across years).

INDEX OF CURRICULUM STANDARDIZATION	
Does your country have **a national curriculum that covers reading instruction** at the fourth grade of primary/elementary school?	Yes= 1, Varies= 0.5, No= 0
How is reading addressed in the curriculum? Reading is presented a **separate curriculum area**?	Yes= 1, Varies= 0.5, No= 0
What does the language/reading curriculum prescribe? **Goals and objectives**	Yes= 1, Varies= 0.5, No= 0
What does … prescribe? **Instructional processes or methods**	Yes= 1, Varies= 0.5, No= 0
What does … prescribe? **Materials (e.g., textbooks, instructional materials)**	Yes= 1, Varies= 0.5, No= 0

national experts and some errors. The data were therefore checked and, in some cases, responses were recoded based on the answers provided to the open questions in the curriculum questionnaires and/or in the PIRLS reports (Martin et al., 2007, Mullis, 2012, Mullis et al., 2017)[4].

Second, we included an indicator for the presence of central examinations at the primary school level, based on the item "How is the implementation of the language/reading curriculum evaluated?" A positive response indicating the presence of "national or regional assessments" was coded as 1, and the absence of such assessments was coded as 0. The variable on the presence of central examinations was extensively checked and cleaned using information from the PIRLS Encyclopedia (Martin et al., 2007; Mullis, 2012; Mullis et al., 2017). We retained positive responses only for examinations administered at the national or regional level and covering the entire primary school population, rather than a sample.[5]

Country-level controls were obtained from external sources: share of tertiary-educated adults (ages 25–64) in the country (Barro & Lee, 2015; Lee & Lee, 2016; data retrieved from Our World in Data, https://ourworldindata.org); Human Development Index (United Nations Development Programme, 2024; data retrieved from Our World in Data, https://ourworldindata.org); share of public expenditure on education (UNESCO Institute for Statistics, 2025; Tanzi and Schuknecht (2000); data retrieved from Our World in Data, https://ourworldindata.org); Gini coefficient (World Bank Poverty and Inequality Platform, World Bank, 2024; data retrieved from Our World in Data, https://ourworldindata.org); proportion of students enrolled in private schools (UNESCO Institute for Statistics 2025.; data retrieved from World Bank Open Data, https://data.worldbank.org); and share of the population born in another country (United Nations Department of Economic and Social Affairs, Population Division, 2024; data retrieved from Our World in Data, https://ourworldindata.org).[6]

3.3. Statistical methods

We adopted a two-step, three-level hierarchical meta-regression approach to estimate the relationship between standardization and student outcomes. In this strategy, students constitute level 1, country-year combinations constitute level 2, and countries constitute level 3. This hierarchical structure follows common recommendations in the methodological literature on applying multilevel models to repeated cross-sectional survey data (Schmidt-Catran & Fairbrother, 2016). All analyses were conducted using Stata 19, and the full code is publicly available.[7]

In the first step, we estimated country-year–specific measures of achievement and inequality. Table 2 summarizes the five educational outcomes, their theoretical estimands (quantities of interest), and the corresponding empirical estimands, methods, and model formulas (Lundberg et al., 2021). Average student achievement is measured as the intercept from an empty OLS model predicting student test scores. Educational poverty is measured as the probability of being low-achieving (score < 400), given by the intercept from an empty linear probability model. Achievement dispersion is captured by the interquartile range, calculated as the difference between the intercepts from two intercept-only quantile regressions at the 0.75 and 0.25 quantiles. Group-based inequalities are estimated using regressions of individual test scores on socio-demographic dummies; specifically, the gender gap and parental education gap are represented by the coefficients on the female and tertiary-educated-parent indicators, respectively.

We ran these individual-level regressions separately for each country-year. To account for the complex survey design, the estimates were computed using the five plausible values of test scores and were combined. Standard errors were clustered at the school level to account for potential intra-class correlation of residuals among students within the same school. This approach provided correct standard error estimates under clustered sampling without explicitly modelling school-level effects (Angrist & Pischke, 2009; Cameron & Miller, 2015). Finally, we applied country-level weights to adjust the sample to reflect population figures.[8]

In the second step, the parameters estimated in step one (e.g., average achievement, inequality measures, and group gaps) were treated as dependent variables in multilevel meta-regression (MMR) models (Thompson et al., 2001). MMR is an extension of meta-regression techniques in which effect sizes (our step-one parameters, constituting level 2) are nested within a higher grouping variable (country, level 3), and may therefore be correlated. Dependencies among the estimated parameters within the same country were accounted for by including random effects at the country level. In this

[4] We focused on four main types of doubtful cases: (1) logical contradictions (i.e., cases where the response to the first item indicated that no national curriculum existed, but one or more follow-up items on curriculum content were coded as 1 or 0.5); (2) federal states, (i.e., countries where educational policy varies across subnational entities, such as Germany, Canada, the United States, Australia, and Belgium); (3) countries with responses coded as "Varies" for one or more items, prompting verification of whether this was appropriate given the national policy context; and (4) countries where the index varied in one direction at the beginning of the study period and then shifted in the opposite direction later. The complete list of changes made, as well as the compiled dataset, is provided in Appendix B.

[5] The cleaning process particularly targeted countries that reported inconsistent responses across different years, in order to verify whether the observed variation reflected actual policy changes or resulted from coding errors. Several entries were corrected. A detailed description of the cleaning procedure and the list of revisions are provided in Appendix C.

[6] Importantly, because these data are available only at the country level in international databases, Flemish Belgium and French Belgium were both assigned values for Belgium. Similarly, England and Northern Ireland were assigned UK values. If data were missing for a country in a particular year, they were imputed using linear interpolation and extrapolation. However, some countries lacked data for certain variables in all years and could not be included in the analysis. This applies to Azerbaijan, Chinese Taipei, Georgia, Hong Kong, Oman, Quatar, Saudi Arabia and Singapore.

[7] The full Stata code is available at: https://osf.io/qs27c/overview?view_only=cfc3b140aa1947329ad67c56a148491e

[8] We used the user-written -pv- command in Stata, specifically designed to handle the complex structure of large-scale international assessment datasets like PIRLS, including plausible values, sampling weights, and stratified multistage designs. Weights could not be applied in the first set of the analysis based on the interquartile regression because the Stata command *iqreg*, which we used to estimate the interquartile range in the complex setting of multiple plausible values, does not support any weighting option.

Table 2
First-step estimation: outcomes, estimands, methods, and formulas.

Educational outcome	Theoretical estimand	Method	Equation/Formula	Empirical estimand
Effectiveness				
Average achievement	Mean literacy test score (T)	OLS linear regression	$T_{scy} = \beta_{0,cy} + \varepsilon_{scy}$	Intercept from empty model: $\widehat{\beta}_{0,cy} = \overline{T}_{cy}$
Interindividual inequality				
Educational poverty	Probability of being low achiever (L)	Linear probability model	$L_{scy} = \pi_{0,cy} + u_{scy}$	Intercept from empty model: $\widehat{\pi}_{0,cy} = \overline{L}_{cy}$
Achievement dispersion	Interquartile range in test scores (IQR)	Interquartile range regression	$Q_{0.75}(T_{scy}\|c,y) = \alpha_{0.75,cy}$; $Q_{0.25}(T_{scy}\|c,y) = \alpha_{0.25,cy}$	Intercept from empty model: $\widehat{\mathrm{IQR}}_{cy} = \widehat{\alpha}_{0.75,cy} - \widehat{\alpha}_{0.25,cy}$
Group-based inequality				
Gender inequality	Gender gap in test scores	OLS linear regression	$T = \beta_{0cy} + \beta_{1cy}\mathrm{Gender}_{scy} + \beta_{2cy}\mathrm{ParentalEd}_{scy} + \beta_{3cy}\mathrm{LanguageHome}_{scy} + \varepsilon_{scy}$	Regression coefficient of female dummy: β_{1cy}
SES inequality	Gap in test scores between students with at least one tertiary educated parent and the others	OLS linear regression		Regression coefficient of parental tertiary education dummy: β_{2cy}

Note: s = student, c=country, y = year.

second step, the within-country-year sampling error from step one enters as a known level-1 variance, that is, $\mathrm{Var}(e_{cy}) = \widehat{\mathrm{Var}}(\widehat{\theta}_{cy})$. Consequently, observations are inverse-variance weighted, giving more weight to more precise first-step estimates (Cameron & Miller, 2015; StataCorp, 2025). With random effects, weights reflect the sum of the known sampling variance and the estimated between-group variance component(s), as in standard meta-analysis.

The key predictors in this second step are the curriculum standardization index and the dummy variable indicating the presence of central exams in primary education. The second-step model is expressed as follows:

$$\widehat{\theta}_{cy} = \beta_0 + \beta_1 \mathrm{Stand\_Index}_{cy} + \beta_2 \mathrm{Central\_Exams}_{cy} + u_c + u_{cy} + \widehat{\varepsilon}_{scy}$$

where $\widehat{\theta}_{cy}$ represents the estimated parameters from the first step of the analysis. *Stand_index*$_{cy}$ and *Central_Exams*$_{cy}$ are the main time-varying macro-level predictors of interest, capturing curriculum standardization and the presence of central exams, respectively. They are included at the second level of the model, as they vary by country-year; u_{cy} is the country-year level random intercept (level 2), u_c is country-level random intercept (level 3), $\widehat{\boldsymbol{\varepsilon}}_{scy}$ is the residual error term derived from the first-step estimates.

In a second specification (Model 2), we included country-level, time-varying control variables obtained from external sources, as described in the variables section. In a third specification (Model 3), we added country fixed effects to Model 2. This model introduces a full set of country dummies, thereby controlling for all time-invariant, country-specific unobserved factors. Consequently, the estimated associations between policy variables and outcomes rely solely on within-country changes over time, which helps mitigate omitted-variable bias due to stable national characteristics. Unlike random intercept models, which exploit both within- and between-country variation and treat country effects as random, fixed-effects models remove all between-country variation by design.

Finally, as a robustness check, we replicated the analysis on a restricted sample of Western countries to enhance institutional comparability.

4. Empirical findings

4.1. Descriptive trends: mean achievements and achievement inequalities in reading

Fig. 1 shows the trends of our outcomes of interest—various indicators of achievement levels and inequality—in the sample of countries examined (see Table A4 and Table A5 in Appendix A for the complete list of estimated statistics for the full sample and by countries). To avoid biases due to an unbalanced sample over time, these trends were estimated using data only from the 28 countries available across the entire study period.

Average reading scores increased slightly, from 512 points in 2006–527 points in 2016. Score dispersion remained fairly stable over time, around 100 points, while the proportion of children scoring below 400 points decreased from 12 % in 2006 to 9 % in 2016. This benchmark, the lowest used in PIRLS reports on achievement levels, indicates that students can retrieve explicitly stated information and make straightforward inferences when reading the simplest texts.

The score advantage for children with at least one parent with tertiary education remained relatively stable (46 points in 2006, 43 in 2011, and 46 in 2016), while the female advantage increased slightly, from 13 points in 2006 to 14 points in 2016.

However, these overall trends, computed on pooled data, mask substantial heterogeneity between countries in both performance and inequalities. The average score fell below 400 points in several countries, such as South Africa, Kuwait, and Qatar in 2006, and Morocco in 2011 and 2016, while it exceeded 570 points in Singapore and Russia in 2016. Trends also varied considerably across countries: For example, the average score decreased by 22 points in Flemish Belgium between 2006 and 2016, but increased by 20 points in Slovenia.

Inequalities likewise show substantial variation. The advantage of students with at least one parent with tertiary education was 76 points in Indonesia in 2006, decreasing to 56 points in 2011, whereas it was only 54 points in Bulgaria in 2006 and increased to 60 points in 2016.

While many institutional and cultural factors could explain the differences in levels and trends between countries, one hypothesis is that these may be at least partly related to variation in the standardization of the education system. In the main analysis, we therefore test whether differences in score inequality across countries can be understood in terms of varying levels of standardization, specifically considering curriculum standardization and the presence of central exams.

4.2. Curriculum standardization and central exams in primary education

We next provide a comprehensive mapping of variation in curriculum standardization and the presence of central exams across countries and over time. As described above, the curriculum standardization index reflects both the existence of a standardized reading curriculum in a given year and the degree of specificity it entails. Fig. 2 presents the values of the curriculum standardization index (ranging from −2.9 to 1.9, after standardization) for each country and year.

Our analyses can rely on consistent variation over time, as the level of curriculum standardization changes in 33 of the 47 countries during

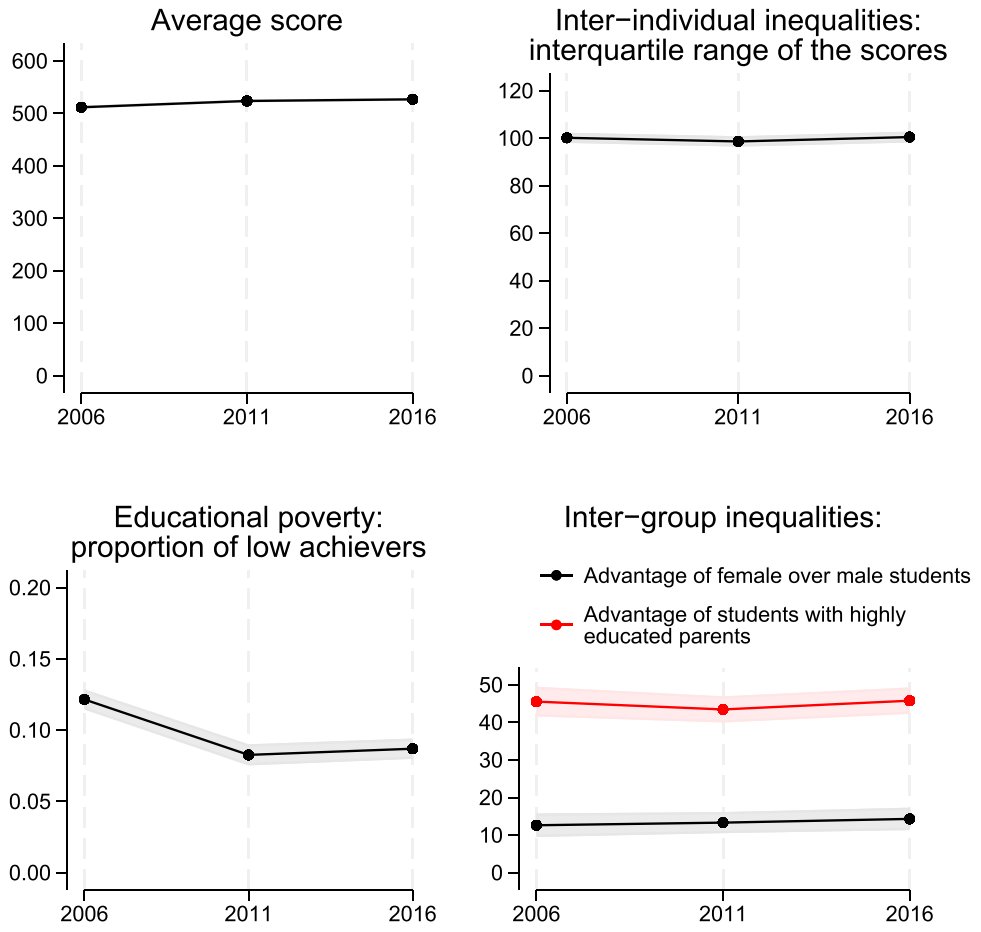

Fig. 1. *Average score and several measures of score inequality in the pulled sample in 2006, 2011, 2016.* Note: Only countries available throughout the entire study period are included. Each country sample is given the same weighting. 95 % confidence intervals are shown.

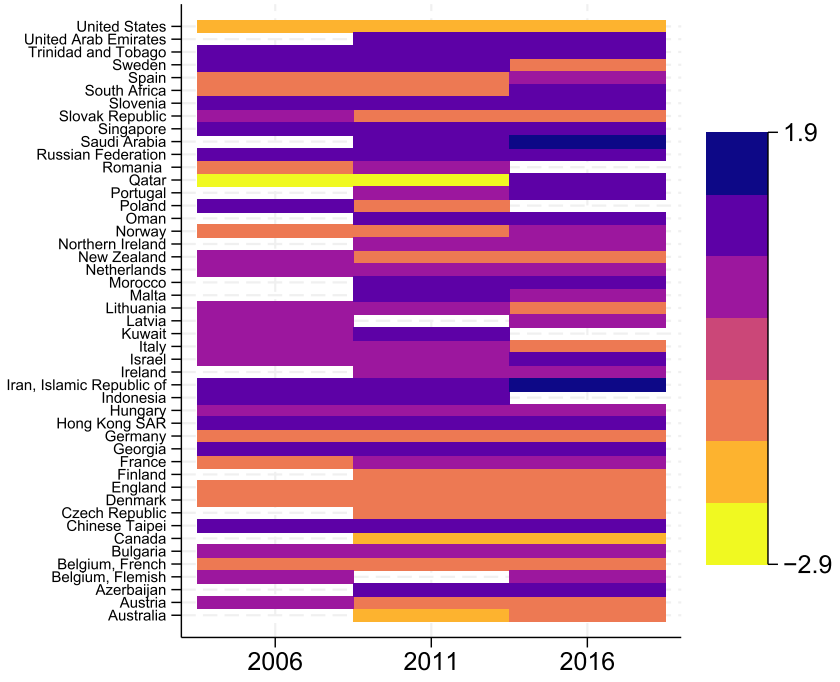

Fig. 2. The index of curriculum standardization in each country and year (2006, 2011, 2016).

the period studied. About half of the variance in the index occurs within countries, reducing the risk that only a small fraction of the total variation is captured over time.[9] At the start of the study period, a cluster of countries—particularly in Western Europe (e.g., Germany, France, and England) and North America (e.g., Canada, the United States)—exhibited relatively low levels of curriculum standardization. In contrast, East Asian and post-Soviet countries (e.g., Chinese Taipei, Singapore, Georgia, and Azerbaijan) consistently displayed higher levels across all three survey waves. Over time, some countries—such as Israel, Qatar, and South Africa—notably increased their standardization index by 2016, suggesting ongoing reforms to centralize and align curriculum content. Conversely, other nations—such as Sweden, Lithuania, and Austria—showed reversals or declines, possibly reflecting decentralization or greater curricular flexibility. Overall, while most countries maintain relatively stable levels over time, these trends highlight the persistence of regional models of curriculum governance and the limited convergence in primary education policy across countries.

Moving to the second variable capturing output standardization, Fig. 3 shows the presence or absence of central exams in each country and year. Across the 2006–2016 PIRLS cycles, central examinations in primary education exhibit moderate cross-national variation and limited temporal change. Indeed, only 9 of the 47 countries studied altered their status over time. Nevertheless, 46 % of the variance in this variable occurs at the within-country level (see Figure A2 in Appendix A for a visualization of the total and residual distribution once country fixed effects are absorbed).

Several educational systems introduced such exams during this period, reflecting ongoing institutional adjustments and policy reforms. For example, the Slovak Republic did not have national assessments at the primary level in 2006 or 2011 but implemented them by 2016. The French-speaking community of Belgium also introduced central exams between 2006 and 2011, aligning with broader efforts to strengthen national curriculum oversight. These changes indicate a modest expansion of standardized assessment tools in primary education in some contexts. Only South Africa moved away from central exams between 2011 and 2016, as the national examinations introduced in 2010 were suspended after 2015.

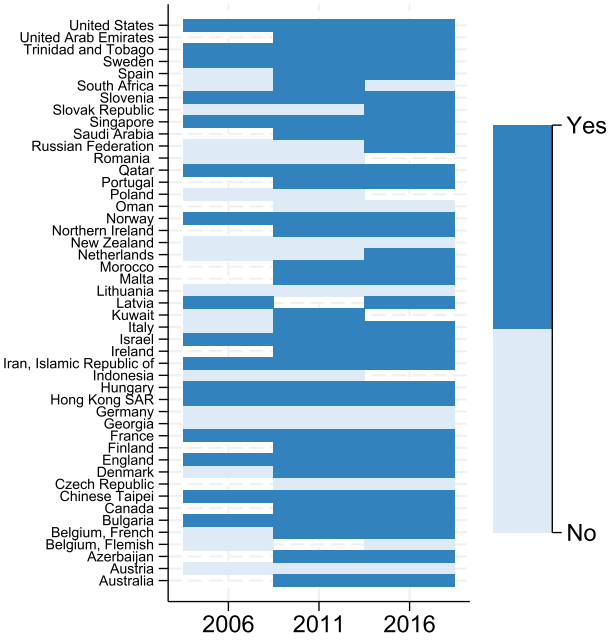

Fig. 3. The presence of central exams in primary education across countries and years (2006, 2011, 2016).

Geographically, the use of central exams varies considerably. In Western Europe, countries such as France, Ireland, and England had already introduced standardized assessments at the primary level by 2006, while others, including Italy and the French-speaking part of Belgium, introduced them later. Some countries, such as Austria and the Flemish part of Belgium, never used them. The Nordic countries—Finland, Sweden, and Norway—maintained central exams throughout the study period. In Eastern Europe and the post-Soviet region, Slovenia and Bulgaria consistently implemented central exams. However, many other countries in this area—such as Lithuania, Georgia, Poland, Romania, and the Czech Republic—reported no central exams in any wave. In the Middle East and North Africa, central exams are prevalent, with countries such as Qatar, Saudi Arabia, and Iran reporting them consistently. In East Asia, high-performing systems such as Singapore, Chinese Taipei, and Hong Kong SAR consistently employed central assessments, reflecting broader trends of curriculum coherence and national monitoring. In Anglophone countries, central exams are also common, being implemented throughout the study period in England, Canada, Australia, and the United States.

Taken together, these patterns reflect both continuity and ongoing reform dynamics. While central exams are increasingly a common feature of primary education systems worldwide, their implementation is neither universal nor static. Interestingly, when considered jointly, the two dimensions of input and output standardization are only moderately correlated. Fig. 4 presents a strip plot showing the distribution of countries (dots) across the two dimensions of standardization for the three years included in the analysis.

While both curriculum standardization (input) and the use of central examinations (output) are often viewed as complementary aspects of educational standardization, their empirical relationship is only moderate. In other words, countries that adopt a centralized and prescriptive national curriculum do not necessarily implement centralized assessments, and vice versa. This observation underscores the institutional decoupling that can occur between the formal specification of what should be taught and the mechanisms used to evaluate what students have learned. Some countries pursue curriculum coherence without central exams, relying instead on teacher- or school-level assessments to monitor student progress. Conversely, others implement national assessments without a fully prescriptive curriculum, using standardized testing primarily for system monitoring or accountability rather than to enforce curricular uniformity.

4.3. Effects of curriculum standardization and central exams on student achievement and inequality

We now examine the effects of curriculum standardization and the presence of central exams on both average achievement and achievement inequality. As outlined in the methods section, we employed multilevel meta-regression techniques to estimate the relationship between our measures of input and output standardization and a set of outcome indicators—capturing effectiveness and inequality—calculated at the country-year level. Figs. 5–7 show the results of this analysis. For each outcome, we estimated three models of increasing restrictiveness. The green coefficient corresponds to a random intercept model that leverages both between-country and within-country variation. The black coefficient reflects a random intercept model that additionally includes macro-level control variables—namely, the Human Development Index, the Gini coefficient, public expenditure on education, the proportion of students enrolled in private schools, and the share of non-native population—to account for time-varying structural differences across countries. Finally, the orange coefficient is derived from a model with

[9] See Figure A1 in Appendix A for a visualization of the total and within-country (residual) distribution of the curriculum standardization index after accounting for country fixed effects.

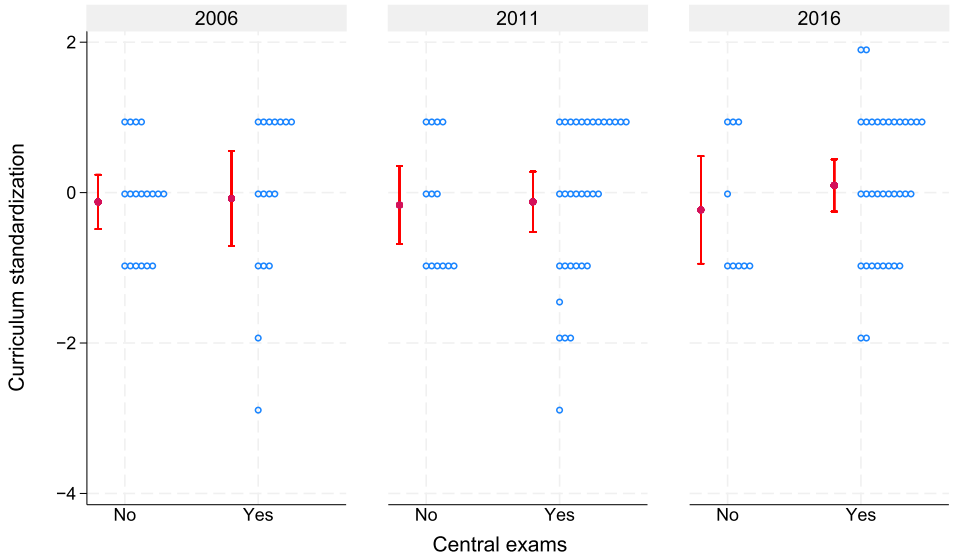

Fig. 4. *Co-variation between central examinations (x-axis; 0 =no, 1 =yes) and the curriculum standardization index (y-axis).* Note: Blue hollow circles are country–year observations (2006/2011/2016; x-positions jittered). Red dots show the pooled mean of the index for systems without/with central exams; whiskers indicate 95 % CIs.

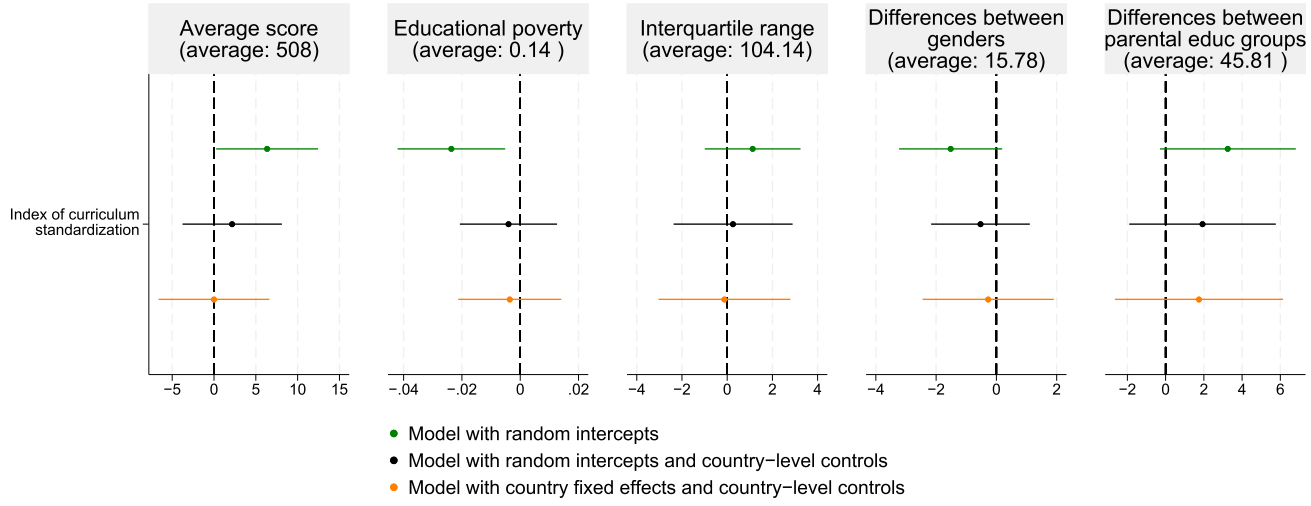

Fig. 5. Results from the meta multilevel models: coefficients (and 95 % confidence intervals) showing the relationship between the index of curriculum standardization and several indicators of achievement and achievement inequality across three model specifications.

country fixed effects, which isolates within-country variation over time by controlling for all time-invariant country characteristics.

The first column of Fig. 5 shows the estimated association between the index of curriculum standardization and average reading test scores. In the first model specification, curriculum standardization appears slightly positively correlated with average scores. However, this correlation disappears when time-varying macro-level control variables are included, suggesting that it was driven by other country-level factors shaping differences in students' scores across countries. This null result is even clearer in the third model, with the fixed-effects specification—which relies solely on within-country changes over time—reinforcing that shifts in curriculum standardization within a country are not systematically associated with changes in average reading performance. Taken together, these results indicate that, at least at the primary level and over the period analyzed, curriculum standardization does not appear to enhance or hinder overall reading achievement.

We next examine the association between curriculum standardization and achievement inequality, using multiple indicators as shown in Fig. 5. The second column focuses on educational poverty, defined as the proportion of students scoring below 400 points on the PIRLS reading assessment. Results from the first model specification—the random intercept model—suggest a negative correlation between curriculum standardization and the prevalence of low achievement. However, this association disappears when country-level control variables (black coefficient) or fixed effects (orange coefficient) are included. Both of these more restrictive specifications indicate that no statistically significant or substantially relevant relationship exists between curriculum standardization and educational poverty in primary education.

We then examine interindividual inequalities, measured by the interquartile range of scores (third column). All model specifications indicate a non-significant association between curriculum standardization and the interquartile range of test scores.

The last two columns examine group-based inequalities, defined by gender (fourth column) and parental education (fifth column). We again find no meaningful association between the curriculum standardization index and the score gap between male and female students, nor any

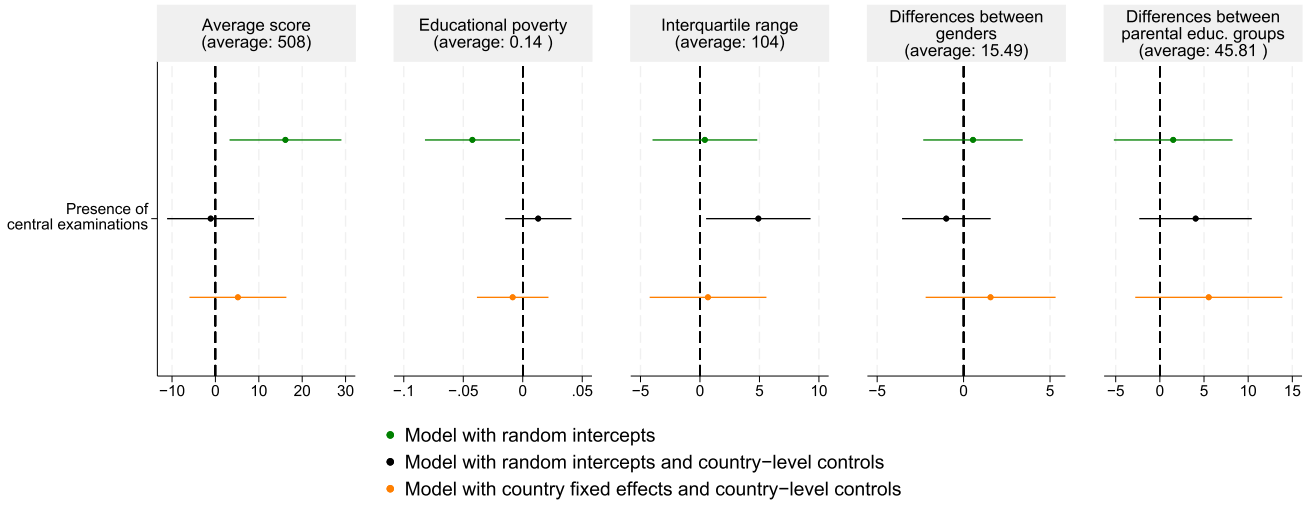

Fig. 6. Results from the meta multilevel models: coefficients (and 95 % confidence intervals) showing the relationship between the presence of central exams and several indicators of achievement and achievement inequality across three model specifications.

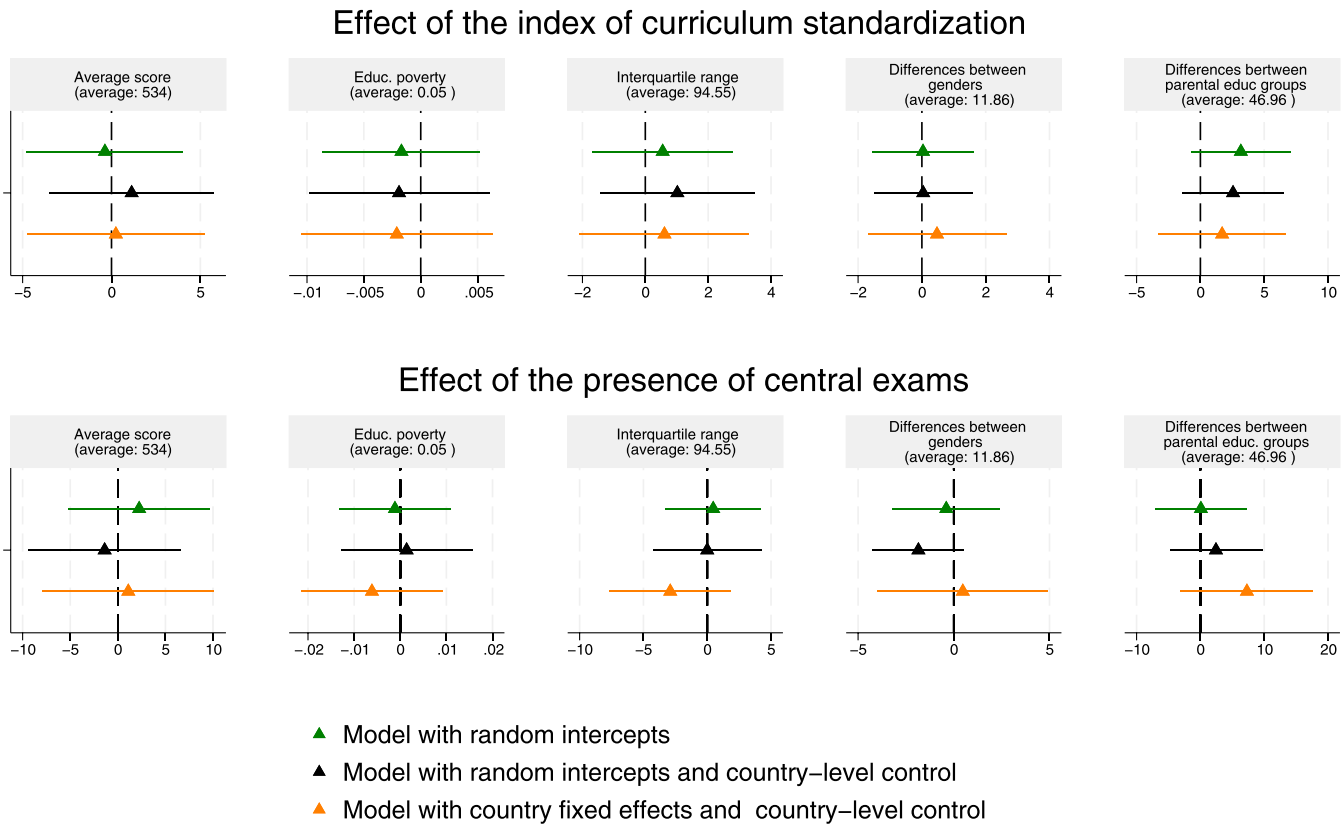

Fig. 7. Results from the meta multilevel models: coefficients (and 95 % confidence intervals) showing the relationship between the index of curriculum standardization and central exams, on one side, and several indicators of score inequalities, on the other, among Western countries only.

significant association with the score gap between students with at least one parent with tertiary education and those without (Model 3).

Fig. 6 presents the results of the meta-multilevel models, with the presence of central examinations in primary schools as a predictor of average performance levels and educational inequality across countries and years. The random-intercept models suggest that central examinations are associated with higher average test scores and a lower proportion of low-achieving students, consistent with previous literature indicating that introducing centralized exams can improve learning outcomes (Bishop, 1997, 1999; Bol & Van de Werfhorst, 2013). However, this initial association does not hold in more restrictive model specifications: once country-level control variables are included, the relationship between central examinations and both average test scores and educational poverty disappears.

Interestingly, only the interquartile range of scores remains significantly associated with the presence of central examinations when country-level controls are included, indicating a potential increase in score variability. However, this association becomes non-significant

once country fixed effects are introduced, suggesting that it may still be driven by unobserved, country-specific factors. Overall, these findings provide little evidence that central examinations reduce learning inequalities; if anything, they could be linked to greater disparities in student performance.

Regarding group-based inequalities, none of the model specifications show a significant effect of central examinations on achievement gaps between male and female students or between students with at least one university-educated parent and the others. Similarly, both the random-intercept model and the specification with country-level controls reveal no significant association with performance differences by parental education. Overall, the evidence suggests that central examinations are unlikely to be an effective tool for reducing learning inequalities.

4.4. Robustness checks

As a robustness check, we repeated the analyses on a restricted sample including only Western countries. This approach helps mitigate concerns about unobserved heterogeneity arising from large cross-national differences in educational systems, political institutions, and cultural norms. By focusing on a more institutionally and culturally homogeneous subset of countries, we reduce the risk of conflating the effects of curriculum standardization with broader structural or contextual disparities, thereby strengthening the internal validity of our findings.

The results, shown in Fig. 7, provide even clearer evidence against a negative relationship between standardization and literacy inequality. Across all model specifications, we find no indication that either the curriculum standardization index or the presence of central examinations significantly affects reading achievement levels or inequality measures.

As a further robustness check, we examined the relationship between the two country-level predictors and another measure of achievement inequality: between-school inequality, net of individual-level characteristics. The rationale is that greater standardization could lead to more consistent teaching practices and learning experiences across schools. Between-school inequality was operationalized using the interclass correlation coefficient, which measures the proportion of total score variance in a country in a given year that is attributable to differences between schools. The results of this supplementary analysis are reported in Figure A3 of Appendix A. These results confirm the absence of a significant relationship between either the level of curriculum standardization or the presence of central exams and learning inequality.

Finally, we analyzed the association between achievement inequality and the individual items within the curriculum standardization index that exhibited the most change over time. Figure A4 in Appendix A shows the standard deviation of each item and its values for every country and year. Three items displayed greater variation both between and within countries compared to the others: whether reading instruction constitutes a separate curriculum area (item 2); whether teaching methods and procedures are prescribed (item 4); and whether materials are prescribed (item 5). We therefore reproduced the main analyses using these items as predictors, with results shown in Figure A5 of Appendix A. These analyses confirm that higher standardization is not associated with improvements in average scores or reductions in score inequality. In fact, the only statistically significant findings indicate slightly larger inequalities between genders and between parental education groups when materials are chosen nationally. However, this result is based on a single item and should not be overinterpreted. Overall, these supplementary analyses are highly consistent with those obtained using the overall index, further reinforcing the robustness of our findings.

5. Discussion and conclusions

The standardization of educational systems has been widely discussed as a potential lever for improving both the effectiveness and equity of schooling. Rooted in theories of institutional regulation and accountability, standardization is often viewed as a way to reduce disparities in educational experiences by ensuring that all students are exposed to similar curricula and evaluated using common benchmarks (Allmendinger, 1989; Van de Werfhorst & Mijs, 2010). While much of the empirical literature has focused on secondary education—particularly in relation to tracking, national examinations, and labor market outcomes—less is known about how standardization operates in the earlier stages of schooling.

This study set out to address this gap by examining the extent to which standardization of the intended curriculum and the use of central examinations are associated with differences in reading achievement and educational inequality among fourth-grade students across a broad set of countries. Using three waves of PIRLS data (2006, 2011, and 2016), our research contributes to the literature both substantively and methodologically. Substantively, we shift the focus to primary education, a level often underexamined despite its critical role in shaping foundational skills and long-term educational trajectories.

Methodologically, we contribute to the literature by designing and implementing a novel two-step multilevel meta-regression strategy for repeated-cross-sectional international survey data. This approach offers several important advantages. First, it is specifically designed to handle the complex multilevel structure of international educational data, in which students are nested in schools and observations are repeated across multiple country-years. By separating estimation into two stages, it accounts for the PIRLS sampling design, the use of plausible values, and hierarchical dependencies in the data. Second, it provides a flexible framework to estimate diverse outcome measures—including average performance, interindividual inequality, and group-based disparities—while incorporating appropriate standard errors and weights at each stage. Third, and most critically, the use of repeated cross-sectional data and country-year variation allows us to exploit within-country changes in curriculum standardization and the implementation of central exams. This enables a more credible identification of the effects of standardization than conventional cross-sectional designs, which are limited in their ability to isolate institutional effects from stable country characteristics. By modeling country-year outcomes as a function of year-specific measures of standardization, we assess whether temporal reforms or shifts in primary school educational standardization are systematically associated with changes in both the effectiveness and equity of student learning.

Our results offer several insights. First, the analysis reveals substantial cross-national variation and modest but meaningful temporal changes, especially in curriculum standardization and, to a lesser extent, in the use of central examinations in primary education. Some countries—particularly in East Asia and parts of the Middle East—consistently exhibit high levels of both input and output standardization, whereas others, including many Anglo-Saxon and Eastern European countries, maintain lower or mixed levels. Curriculum standardization is only slightly more common among countries with centralized education governance, although notable exceptions exist. Central examinations at the primary level, by contrast, are more widespread and show limited temporal fluctuations: Several countries introduce them during the observation period, while only one (South Africa) suspends them. Although curriculum standardization and central examinations are often conceptualized as complementary dimensions of educational governance, our findings reveal only a moderate empirical correlation between the two. This indicates that countries may adopt prescriptive national curricula without implementing centralized assessments, or vice versa. This pattern reflects a degree of institutional decoupling between the formal definition of instructional content and the mechanisms used to evaluate student learning.

Despite meaningful cross-country and over-time variation, we find no consistent association between either form of standardization—curricular or assessment-based—and reading achievement levels or inequality at the primary level. The apparent positive effects of standardization, namely higher average student performance and reduced educational poverty observed in the random-intercept models, do not hold once more refined specifications are applied that account for other differences between countries, such as general levels of inequality and public expenditure on education. This result does not align with parts of the secondary-education literature, where standardization is often linked to higher average performance (e.g., Bol & Van de Werfhorst, 2013; Gamoran, 1996; Montt, 2011). By contrast, evidence on whether—and which—forms of standardization reduce achievement disparities is mixed even at the secondary level (Checchi et al., 2014). In line with prior critiques, claims about the benefits of standardization often rely on non-causal designs, which makes their apparent effectiveness easy to overstate without rigorous identification strategies (Jürges & Schneider, 2004).

The absence of a consistent and statistically significant effect between curriculum standardization or the presence of central exams, on one side, and student outcomes in PIRLS, on the other, requires careful interpretation. In the introduction, we anticipated that both forms of standardization might bring benefits but also involve trade-offs. A centrally defined curriculum may reduce teachers' ability to adapt content to local contexts or student needs, potentially disadvantaging those who do not fit the "standard" student profile (Remillard, 2005; Zhao et al., 2023). Similarly, central examinations can focus teaching too narrowly on tested content, limit pedagogical creativity, and place pressure on students and teachers alike. Moreover, if not carefully designed, they may reflect cultural or linguistic biases that unintentionally disadvantage minority or lower-income students (Hopmann et al., 2007; Looney, 2009).

Several factors—both empirical and theoretical—could help explain these null results. First, there may simply be limited cross-national variation in primary-level standardization. Many education systems already exhibit a high degree of standardization in primary education at the beginning of the studied period, especially in terms of national curricula. This narrow range of variation across countries may limit our ability to detect effects. When nearly all systems prescribe a core reading curriculum and administer national assessments at the primary level, the absence of a "low-standardization" reference group may obscure associations with outcomes.

A second, related explanation concerns possible threshold or nonlinear effects: It is plausible that standardization influences outcomes only once it exceeds a certain level of implementation or salience. Minor changes in curriculum coherence or superficial national assessments may be insufficient to shift teaching practices or student learning. If countries differ only marginally in their degree of standardization, or if reforms are recent or partial, the effect sizes may remain undetectable in aggregate models.

Third, central exams in primary education may simply have low salience. In contrast to secondary education, primary-level central exams are often formative rather than high-stakes. They may exist as part of a national evaluation system but typically do not carry consequences for students, teachers, or schools, except in some early-tracking countries (Blossfeld et al., 2016). As a result, they may lack the motivational and accountability mechanisms that make them more impactful in later stages of schooling (e.g., selection into tracks or school types), thereby diluting their measurable effects on achievement or equity.

Mismatch between the intended and implemented curriculum could be a fourth reason behind the null effects. PIRLS measures the intended curriculum (i.e., what policymakers prescribe), but this may not reflect what is actually taught and learned in classrooms (Schmidt & Prawat, 2006). Implementation gaps—driven by local school autonomy, teacher discretion, or resource disparities—can weaken the relationship between formal curriculum standardization and student competencies. This "curriculum misalignment" is particularly relevant in large or decentralized systems, where national policies may not translate into uniform classroom practices (Klieme et al., 2009).

Overall, it could be that early education outcomes are driven by other mechanisms. In primary education, teacher quality, early childhood education, the family literacy environment, and instructional time may play a more direct role in shaping reading competencies and inequality than structural institutional features like standardization. Some literature also argues that institutional features of educational systems may be more consequential for mathematics than for reading, especially in the early years, since reading competencies are learned widely outside school and are more closely tied to the home environment.

It is also possible that the potential beneficial effects of standardization emerge more strongly later in the educational trajectory, when curriculum differentiation, tracking, and exam stakes become more salient. Lagged effects and the short observation window could also help explain the limited effects of curriculum standardization and central exams in primary education. Institutional changes in curriculum or assessment may take time to influence learning environments and outcomes. If standardized curricula were introduced or reformed only shortly before the PIRLS waves, there may not have been enough time for measurable effects to appear. Moreover, reading literacy develops cumulatively, and early standardization reforms may have effects that become observable only in later grades or transitions.

These findings challenge some prevailing assumptions about the universal benefits of standardization and call for a more nuanced view. While standardization may promote equity and effectiveness under certain conditions, its presence in policy documents or curriculum guidelines alone may not be sufficient. Instead, attention should be paid to how policies are enacted in practice and to how different institutional arrangements interact with local implementation capacities and the role of a given educational stage in the broader educational career.

From a policy perspective, our results suggest that reforms aimed at increasing standardization in primary education should be approached with caution and grounded in evidence about actual mechanisms of influence. Uniform curricula and central exams are unlikely to improve student outcomes unless they are meaningfully integrated into pedagogical practice and supported by adequate resources and teacher training. Moreover, our findings underline the importance of considering other unobserved variables that are likely to have a stronger influence on student outcomes. Factors such as teacher quality and turnover, school-level resources, class size, and regional funding disparities may play a more decisive role in shaping educational inequality than national curriculum frameworks in primary education. For instance, unequal access to well-trained educators and supportive learning environments may undermine the intended equity goals of curriculum reforms.

While our study relies on a relatively rich cross-national dataset and exploits meaningful within-country variation over time, some methodological limitations should be noted. First, the number of country-year units is modest, which may reduce statistical power to detect small effects. Second, although the indicators of curriculum standardization and central examinations are based on internationally standardized sources and display substantial temporal variation, they rely on a limited number of items and may not fully capture qualitative differences across contexts. These limitations imply that small or highly context-specific effects could remain undetected. Despite these constraints, our design provides strong initial evidence by leveraging the best available temporal variation in international data on curricula and central exams. The fact that the point estimates are generally close to zero rather than merely imprecise suggests that the lack of significant effects is not solely attributable to data constraints but likely reflects the limited role of curriculum standardization and central exams in shaping reading outcomes at the primary level. As future PIRLS cycles and additional policy reforms generate more within-country variation, this analytic

framework can be replicated and extended to yield more precise estimates of how standardization influences achievement inequality.

Moreover, we acknowledge that our analysis focuses on the intended curriculum and its official evaluation mechanisms but cannot fully capture how curricula are implemented in classrooms. Future research should aim to link standardization not only to policy documents but also to the enacted and attained curriculum, including teacher practices, school leadership, and student engagement. Additionally, while our results pertain to primary education, further work is needed to compare the role of standardization across educational levels and to explore how its effects accumulate over the life course. More granular analyses focusing on sub-national units or using mixed methods could also illuminate the institutional, cultural, and contextual conditions under which standardization matters most.

In sum, this study offers one of the first cross-national, longitudinal analyses of the effects of curriculum standardization and central exams in primary education. While the results do not support the expected equity-enhancing or effectiveness-promoting effects of standardization, they raise important questions about when, how, and under what conditions standardization can serve as an effective policy tool.

This research was supported by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101086663 — EDUPOL, ERC Consolidator Grant, call reference: ERC-2022-COG), awarded to Principal Investigator Moris Triventi (Department of Social and Political Sciences, University of Milan).

During the preparation of this work the authors used generative AI (ChatGPT) to polish English grammar and clarity, as well as to help identify a few selected potential data inconsistencies by cross-referencing external sources (as documented in the supplementary materials). After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

CRediT authorship contribution statement

Nathalie Vigna: Writing – original draft, Formal analysis, Data curation. **Moris Triventi:** Writing – original draft, Conceptualization, Supervision, Project administration. **Giuliana Parente:** Writing – original draft, Conceptualization.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.stueduc.2026.101570.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review, 72*(2), 167–180. https://doi.org/10.1177/000312240707200202

Allmendinger, J. (1989). Educational systems and labor market outcomes. *European Sociological Review, 5*(3), 231–250. https://doi.org/10.1093/oxfordjournals.esr.a036524

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Ayalon, H., & Livneh, I. (2013). Educational standardization and gender differences in mathematics achievement: A comparative study. *Social Science Research, 42*(2), 432–445. https://doi.org/10.1016/j.ssresearch.2012.10.001

Baker, D. P. (2012). Ten: institutional change in education: Evidence from cross-national comparisons. In H.-D. Meyer, & B. Rowan (Eds.), *The new institutionalism in education* (pp. 163–185). State University of New York Press. https://doi.org/10.1515/9780791481080-012.

Barro, R. J., & Lee, J. W. (2015). *Education matters: Global schooling gains from the 19th to the 21st Century*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199379231.001.0001

Bishop, J. H. (1997). The effect of national standards and curriculum-based examinations on achievement. *American Economic Review, 87*(2), 260–264. ⟨http://www.jstor.org/stable/2950928⟩.

Bishop, J. H. (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Review, 6*(2), 348–398.

Blossfeld, H. P., Buchholz, S., Skopek, J., & Triventi, M. (Eds.). (2016). *Models of secondary education and social inequality: An international comparison*. Edward Elgar Publishing.

Bol, T., & Van de Werfhorst, H.G. (2013). *The measurement of tracking, vocational orientation, and standardization of educational systems: A comparative approach*. GINI Discussion Paper 81. ⟨https://www1feb-uva.nl/aias/81-3-3-1.pdf⟩.

Brint, S. (2017). *Schools and societies*. Stanford University Press.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Checchi, D., van de Werfhorst, H., Braga, M., & Meschi, E. (2014). The policy response: Education. In W. Salverda, B. Nolan, I. Marx, A. McKnight, I. G. Tóth, & H. van de Werfhorst (Eds.), *Changing inequalities and societal impacts in rich countries: Analytical and comparative perspectives* (pp. 294–327). Oxford University Press.

Chiu, M. M., & Chow, B. W. Y. (2015). Classmate characteristics and student achievement in 33 countries: Classmates' past achievement, family socioeconomic status, educational resources, and attitudes toward reading. *Journal of Educational Psychology, 107*(1), 152–169. https://doi.org/10.1037/a0036897

Dubet, F. (2004). *Le déclin de l'institution*. Seuil.

Gamoran, A. (1996). Curriculum standardization and equality of opportunity in Scottish secondary education: 1984–90. *Sociology of Education, 69*(1), 1–21. https://doi.org/10.2307/2112720

Han, S., & Buchmann, C. (2016). Aligning science achievement and STEM expectations for college Success: A comparative study of curricular standardization. *RSF: The Russell Sage Foundation Journal of the Social Sciences, 2*(1), 192–211. https://doi.org/10.7758/rsf.2016.2.1.09

Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116*(510), C63–C76.

Hopmann, S., Brinek, G., & Retzl, M. (Eds.). (2007), *PISA zufolge PISA / PISA according to PISA: Hält PISA, was es verspricht? / Does PISA keep what it promises?, 6*. LIT Verlag.

Horn, D. (2009). Age of selection counts: A cross-country comparison of educational institutions. *Educational Research and Evaluation, 15*(4), 343–366. https://doi.org/10.1080/13803610903087011

Jürges, H., & Schneider, K. (2004). International differences in student achievement: An economic perspective. *German Economic Review, 5*(3), 357–380. https://doi.org/10.1111/j.1465-6485.2004.00113.x

Jürges, H., Schneider, K., & Büchel, F. (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association, 3*(5), 1134–1155. https://doi.org/10.1162/1542476054729400

Klieme, E., Pauli, C., & Reusser, K. (2009). The pythagoras study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In Tomáš. Janík, & Tina Seidel (Eds.), *The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (pp. 137–160). Münster: Waxmann.

Kulic, N., Skopek, J., Triventi, M., & Blossfeld, H. P. (2019). Social background and children's cognitive skills: The role of early childhood education and care in a cross-national perspective. *Annual Review of Sociology, 45*(1), 557–579. https://doi.org/10.1146/annurev-soc-073018-022401

Lee, J. W., & Lee, H. (2016). Human capital in the long run. *Journal of Development Economics, 122*, 147–169. https://doi.org/10.1016/j.jdeveco.2016.05.006

Looney, J. W. (2009). *Assessment and innovation in education*. OECD Education Working Papers No. 24. OECD Publishing. https://doi.org/10.1787/222814543073.

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review, 86*(3), 532–565. https://doi.org/10.1177/00031224211004187

Martin, M. O., Trong, K. L., & Mullis, I. V. S. (Eds.). (2007). *PIRLS 2006 encyclopedia: A guide to reading education in the forty PIRLS 2006 countries*. International Study Center.

Montt, G. (2011). Cross-national differences in educational achievement inequality. *Sociology of Education, 84*(1), 49–68. https://doi.org/10.1177/0038040710392717

Mullis, I. V. S. (2012). Ed. *PIRLS 2011 Encyclopedia: Education policy and curriculum in reading*. TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Goh, S., & Prendergast, C. (Eds.). (2017). *PIRLS 2016 Encyclopedia: Education Policy and Curriculum in Reading*. TIMSS & PIRLS International Study Center. ⟨https://timssandpirls.bc.edu/pirls2016/encyclopedia/⟩.

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. OECD Publishing. https://doi.org/10.1787/9789264266490-en

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research, 75*(2), 211–246. https://doi.org/10.3102/00346543075002211

Schmidt, W. H., & Prawat, R. S. (2006). Curriculum coherence and national control of education: Issue or non-issue? *Journal of Curriculum Studies, 38*(6), 641–658. https://doi.org/10.1080/00220270600682804

Skopek, J., Triventi, M., & Buchholz, S. (2019). How do educational systems affect social inequality of educational opportunities? The role of tracking in a comparative perspective. In R. Becker (Ed.), *Research handbook on the sociology of education* (pp. 214–232). Edward Elgar Publishing Ltd. https://doi.org/10.4337/9781788110426.00022.

Schmidt-Catran, A. W., & Fairbrother, M. (2016). The random effects in multilevel models: Getting them wrong and getting them right. *European Sociological Review, 32*(1), 23–38. https://doi.org/10.1093/esr/jcv090

StataCorp, LLC. (2025). *Stata 19 meta-analysis reference manual*. Stata Press.

Stevenson, D. L., & Baker, D. P. (1991). State control of the curriculum and classroom instruction. *Sociology of Education, 64*(1), 1–10. https://doi.org/10.2307/2112887

Tanzi, V., & Schuknecht, L. (2000). *Public spending in the 20th century: A global perspective.* Cambridge University Press. ISBN 0-521-66291-5.

Thompson, S. G., Turner, R. M., & Warn, D. E. (2001). Multilevel models for meta-analysis, and their application to absolute risk differences. *STatistical Methods in Medical Research, 10*(6), 375–392. https://doi.org/10.1177/096228020101000602

UNESCO Institute for Statistics . (2025). UIS bulk data download [Data set]. ⟨https://databrowser.uis.unesco.org/resources/bulk⟩.

United Nations Development Programme. (2024). *Human development report 2023–24: Breaking the gridlock: Reimagining cooperation in a polarized world.* United Nations Development Programme.

United Nations Department of Economic and Social Affairs, Population Division. (2024). *International migrant stock 2024* [Data set]. ⟨https://www.un.org/development/desa/pd/content/international-migrant-stock⟩.

Van de Werfhorst, H. G., & Mijs, J. J. B. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology, 36*(1), 407–428. https://doi.org/10.1146/annurev.soc.012809.102538

World Bank. (2024). *Poverty and Inequality Platform (PIP) [Data set].* ⟨https://pip.worldbank.org/⟩.

Zhao, Y., Li, T., & Liu, W. (2023). The benefits and drawbacks of standardized curriculum in education. *Research and Advances in Education, 2*, 41–47. https://doi.org/10.56397/RAE.2023.10.05