

Curriculum Reform or Structural Equity? A Multilevel Analysis of Indonesia's PISA Performance Across Eight Cycles, 2000-2022

Nur Azizah Nasution¹, Valid Hasyimi²

^{1,2}Institut Teknologi Bandung (ITB)

Jl. Ganesha No. 10, Bandung, West Java, Indonesia

29324417@mahasiswa.itb.ac.id, valid_hasyimi@sbm-itb.ac.id

Abstract. This study examines whether curriculum era shifts correlate with Indonesia's PISA score fluctuations. Using multilevel regression on eight PISA cycles (2000–2022) with controls for population coverage and test modes, complemented by semi-structured interviews with six policy informants, curriculum changes showed no significant relationship with scores. Conversely, structural factors like socioeconomic status, urban-rural location, and grade repetition consistently predicted performance. Findings indicate student achievement depends on structural conditions rather than curriculum overhauls. Policy resources should therefore target educational equity, teacher feedback, and socioeconomic support rather than frequent curriculum changes.

Keywords: Curriculum Reform; Structural Equity; Multilevel Analysis; Educational Inequality; Student Achievement.

1. Introduction

The learning crisis in many developing countries is fundamentally not just a matter of low average achievement, but rather an issue of distribution. Over the past two decades, the expansion of educational access in various middle-income countries has often not been accompanied by an equitable distribution of quality; consequently, improvements in participation rates have coincided with stagnation, or even a decline, in learning outcomes. This pattern, where achievement drops precisely as participation rises, has been documented as "learning reversal" (Beatty et al., 2021). Similar phenomena appear across regions: systems that succeed in increasing the number of students in school do not necessarily succeed in enhancing what those students actually learn.

Literature on academic achievement consistently identifies socioeconomic status as one of the strongest correlates of learning outcomes, a finding first brought to prominence at national scale by (Coleman et al., 1966) and confirmed by meta-analytic evidence across countries and decades (Sirin, 2005). Consequently, an education system's average score tends to mask the structural inequalities that lie at the heart of the problem: low achievement is typically concentrated within specific socioeconomic groups and regions, subsequently reproducing itself from one cohort to the next. Viewing achievement merely as a single national figure therefore risks obscuring the true location of the problem and, by extension, obscuring where resources for improvement should be directed.

Learning is also cumulative in nature. Deficits arising at early stages tend to compound and are difficult to rectify at later levels (García et al., 2020). Thus, the achievement of a 15-year-old student represents the accumulation of years of prior learning, and every cohort passing through a low-performing system suffers developmental losses, some of which are irreversible. It is this

cumulative nature that gives education policy choices long-term consequences, transforming the question of which factors truly drive achievement from a mere academic inquiry into an urgent matter of resource allocation. This urgency rests on a long-standing premise in human capital theory: schooling is an investment whose returns accrue over a lifetime, so learning losses in adolescence translate into permanent reductions in productive capacity (Becker, 1964; Schultz, 1961), with cross-national evidence linking cognitive skills rather than years of schooling to long-run growth (Hanushek & Woessmann, 2015).

Indonesia offers a compelling case study for examining this issue. Having participated in PISA without missing a single cycle since 2000, Indonesia possesses a long, internationally comparable series of measurements. Between 2000 and 2022, Indonesia's performance remained stagnant before declining in 2022; average scores in mathematics, reading, and science fell back toward the lows recorded at the time of the country's initial participation (OECD, 2023). The majority of 15-year-old students have yet to reach internationally agreed-upon minimum proficiency levels, while the proportion of high-performing students is virtually zero. Although Indonesia's ranking improved slightly during certain periods, these gains were driven more by the declining performance of other nations than by any rise in Indonesian students' achievement.

The magnitude of the gap becomes even clearer when viewed in terms of learning time. In 2022, the difference between Indonesia's scores and the OECD average ranged from approximately 100 to 117 points per domain; given the common benchmark that 40 PISA points equate to one year of schooling, this indicates a lag of roughly two and a half to three years of education (OECD, 2023). These figures underscore that the problem is not merely one of low average scores, but rather a massive learning deficit entrenched along socioeconomic and geographic lines.

During this same period, the national curriculum underwent frequent changes. The 1994 Curriculum, based on the "link and match" principle, was replaced by the Competency-Based Curriculum (2004), followed by the Education Unit-Level Curriculum (2006), which decentralized instructional design; then came the 2013 Curriculum with its scientific approach, and finally the Kurikulum Merdeka (2022), which emphasizes differentiation and project-based learning. On average, a curriculum remained in place for little more than a decade before being replaced. This recurring pattern, often described by the phrase "new minister, new curriculum", has long been a source of complaint among practitioners. However, its impact on student achievement has never been empirically tested at the national level. In 2025, the government decided to halt curriculum changes and instead focus on deepening the implementation of the existing one; this decision implicitly acknowledges that successive overhauls do not necessarily improve learning outcomes, yet this admission lacks sufficient supporting evidence.

The core issue can be framed as a matter of opportunity cost. An education system's capacity for reform is finite: policymakers' time, budgets for teacher training, the development and dissemination of teaching materials, and political capital are all scarce resources that cannot be allocated in two directions simultaneously. Consequently, every reform decision entails an opportunity cost; opting for one policy lever means delaying another. Throughout the 2000–2022 period, these scarce resources were repeatedly expended on a single lever, the overhaul of national curriculum documents, while factors persistently hindering achievement were allowed to persist across eras.

The pattern of successive curriculum policy overhauls, often termed "policy churn" in the literature, is not unique to Indonesia. Many education systems undertake curriculum document revisions as the most visible response to disappointing assessment results, as such overhauls are easy to announce, easy to claim as achievements, and appear comprehensive. At the same time, more

decisive structural factors, such as socioeconomic inequality and disparities in school quality, change much more slowly and are far harder to frame as symbols of short-term success. Consequently, the question of whether curriculum changes are truly associated with shifts in achievement holds relevance beyond any single nation.

The Indonesian case offers two valuable insights for readers in other contexts. First, it provides an empirical test of a premise rarely examined directly: that curriculum overhaul is a primary driver of achievement. Second, it illustrates a replicable analytical approach, leveraging the recurring cycles of an international assessment to disentangle curriculum associations from time trends and measurement artifacts by categorizing cohorts according to the specific curriculum that actually shaped their education. Other education systems participating in PISA or similar assessments, and grappling with frequent policy shifts, can use these findings as a point of comparison. This study does not claim that Indonesia's results automatically apply elsewhere; rather, it presents patterns and methods, leaving it to readers to assess their relevance to their own specific contexts.

Based on the preceding discussion, this study poses three interrelated questions. First, what is the relationship between curriculum changes and Indonesia's PISA scores from 2000 to 2022, once confounding factors are controlled for? Second, which factors consistently contribute to scores across curriculum eras, thereby qualifying as persistent factors that endure despite curriculum changes? Third, if curriculum changes are not strongly linked to score improvements, toward what areas should priority programs be directed? The first two questions are addressed primarily through a quantitative analysis of eight PISA cycles, while the third integrates these quantitative findings with interviews involving policymakers and cross-country comparisons.

This study offers three mutually reinforcing contributions. Methodologically, it utilizes all eight available PISA cycles, combining cohort exposure logic with explicit controls for testing modes and population coverage, thereby allowing repeated cross-sectional data to serve as a basis for trend inference. Empirically, it tests, rather than assumes, the premise that curriculum overhauls are a primary driver of achievement, examining whether this premise holds when persistent and confounding factors are controlled for. Practically, it translates findings into policy reallocation considerations that are validated for implementation feasibility. The study presents these three contributions as claims open to evaluation, rather than as definitive conclusions.

For readers outside Indonesia, an overview of the trajectory of curriculum reform provides necessary context. The 1994 Curriculum championed the "link and match" principle, utilized a trimester system, and was known for its heavy content load. The Competency-Based Curriculum (2004) shifted the orientation toward competencies and learning outcomes using a student-centered approach, emerging in response to broader reforms and regional autonomy. Just two years later, the Education Unit-Level Curriculum (2006) decentralized instructional design to individual schools within the framework of National Education Standards. The 2013 Curriculum recentralized control, introducing a scientific approach, the integration of attitudes, knowledge, and skills, and authentic assessment. The Kurikulum Merdeka (Emancipated Curriculum) of 2022 emphasized flexibility, differentiation, and project-based learning; it was initially optional before being adopted as the national curriculum.

Two significant features stand out in this trajectory. First, the pendulum swings from centralization to decentralization and back to centralization, and from content-heavy models to competency-based and then project-based approaches, revealing that debates have centered more on the philosophy of the documents than on strengthening implementation. Second, the intervals between changes were often shorter than the time teachers needed to master the new systems truly; for instance, the transition from the 2004 Competency-Based Curriculum to the 2006 Education Unit-

Level Curriculum spanned only two years. A turning point arrived in 2025, when the government decided to halt curriculum overhauls and instead focus on deepening the implementation of the existing framework. It is precisely this pattern that makes the research question relevant. If achievement levels remain stagnant despite shifting pedagogical principles, the determining variables may well lie outside the curriculum documents themselves. This context also informs the coding of curriculum eras in the methodology section, which maps each cohort according to the curriculum that actually shaped their education, rather than the one coincidentally launched during their examination year.

2. Method

This study employs an explanatory sequential mixed-methods design (Creswell & Plano Clark, 2018). The quantitative phase is conducted first to establish a primary foundation, followed by the qualitative phase to explain and examine the findings. This approach was selected because, while numerical data can reveal patterns, they do not explain the underlying reasons; the qualitative phase addresses questions that coefficients cannot answer, such as why a pattern emerges and what the implementation looks like in reality. It is this transition from the macro-level perspective of the quantitative phase to the actors' perspective in the qualitative phase that links hypothesis testing with the formulation of policy recommendations.

The quantitative phase relies entirely on secondary data from the public OECD PISA database, encompassing student and school responses via structured questionnaires as well as cognitive test results across three domains. The data cover all eight cycles of Indonesia's participation: 2000, 2003, 2006, 2009, 2012, 2015, 2018, and 2022. In addition to student- and school-level microdata, the study compiles cycle-level data from OECD technical reports, specifically the population coverage index and the assessment mode for each cycle. Coverage data serves as a compositional control, given that the coverage of the tested Indonesian student population increased rapidly over the period, rising from approximately 40 percent in the early 2000s to around 85 percent in the final cycle. Information on assessment modes was collected because Indonesia utilized paper-based testing from 2000 to 2015 and computer-based testing in 2018 and 2022.

Given that the PISA database contains thousands of variables, a multi-stage filtering process was required. Candidate variables were first mapped to (Bronfenbrenner, 1979) ecological layers and the input categories of the Education Production Function (Hanushek & Woessmann, 2011) then screened based on four operational criteria: consistent availability across all eight cycles, lack of semantic overlap with other variables, clear policy relevance, and data completeness for the Indonesian sample. For the cycle-specific models, approximately ten student and family variables plus four school variables were used, with the exact number varying by cycle depending on data availability. For the pooled cross-cycle model, only five core variables available across all cycles were utilized: socio-economic status (ESCS), mother's education (MISCED), school type, school size, and gender, supplemented by three specially constructed cycle-level variables: curriculum era, test mode, and population coverage.

A key feature distinguishing this study from other cross-sectional PISA analyses is its explicit handling of the time lag between curriculum implementation and the measurement of outcomes. Since PISA assesses 15-year-old students, a curriculum can only be meaningfully evaluated against a cohort that has studied under it for a sufficient duration; measuring impacts during the launch year would yield inaccurate estimates (Cantoni et al., 2017). Applying a minimum lag rule of approximately three years, curriculum eras were coded based on the dominant curriculum shaping each cohort. Consequently, PISA 2000 scores do not reflect policies from the year 2000 itself, but

rather the 1994 Curriculum that had been in effect for several years prior. The 1994 Curriculum cohort served as the reference category, allowing all era coefficients to be interpreted as deviations from a clear baseline. It should be noted that this mapping relies on a dominant-curriculum proxy; the reality of phased and regionally heterogeneous implementation is simplified here, and this assumption is explicitly acknowledged as a study limitation.

Since PISA data are hierarchical, with students nested within schools, observations are not independent; ordinary linear regression would ignore this structure and yield incorrect standard errors (Rabe-Hesketh & Skrondal, 2006). Two model forms were constructed. The per-cycle model (two-level) places students at the first level and schools at the second, with its output used to identify persistent factors. The pooled cross-cycle model nests students within schools and schools within cycles; curriculum-era dummy variables and controls for confounding factors are included at the cycle level to address the first research question directly. Given that there are only eight cycles, cycle-level variables are treated as fixed effects with cycle-level cluster-robust standard errors, preventing the model from becoming overly complex for the limited number of cycles.

PISA does not report a single score per student but rather a set of plausible values representing the distribution of potential proficiency. All plausible values are treated equally: analyses are performed on each set, and the results are pooled according to Rubin's rules, ensuring the final standard errors reflect both the sample design and the uncertainty of proficiency measurement. Final student weights and replication weights are applied to ensure standard errors align with the complex sample design. All data processing was conducted using R software.

Two confounding factors require special handling. The shift in testing mode from paper to computer can generate pseudo-trends easily mistaken for curriculum effects, necessitating explicit control (Robitzsch et al., 2020); furthermore, computer-based assessments require digital navigation skills that can influence scores (Dyrvold & Bergvall, 2023). Changes in population coverage are also controlled for, as expanding coverage from approximately 40 to 85 percent can create pseudo-improvements or score declines driven by shifts in sample composition rather than changes in learning quality. Prior to interpretation, the model was checked for assumptions of normality, homoscedasticity, multicollinearity, linearity, and autocorrelation at the appropriate level; models violating the homoscedasticity assumption were re-estimated using robust standard errors.

The robustness of the era coefficient was tested using several alternative specifications: with and without transition cycles, with and without controls for testing mode and population coverage, and by excluding specific cycles. A factor is deemed persistent if its direction and significance remain consistent across all or most cycles and hold up across model variations. This rule was established a priori to ensure that conclusions were not driven by a single cycle or a solitary specification (Angrist & Pischke, 2009).

The quantitative findings were examined through semi-structured, in-depth interviews with six informants, selected purposively and expanded via snowball sampling, comprising policymakers from standard-setting and curriculum bodies, national assessment administrators, independent policy research institutes, intergovernmental organizations, and former high-level policymakers. To maintain a double-anonymized process, informants were referred to by codes (INF-1 through INF-6) and general role descriptions, without disclosing names or specific institutions. Prior to the interviews, each informant was provided with a summary of the quantitative findings and asked to assess the plausibility of the observed patterns and identify potential inaccuracies. Informant adequacy was evaluated based on the concept of "information power" rather than mere sample size, given the highly specific nature of the interviews and the informants' deep, directly relevant expertise (Malterud et al., 2016).

Transcripts were processed using a six-stage thematic analysis that combined deductive coding, based on the framework of the three research questions, with inductive coding of concepts emerging from the data (Braun & Clarke, 2006). As the researcher was also the designer of the model under review, three measures were taken to mitigate confirmation bias: interview protocols were designed to elicit counter-arguments, coding adhered to a codebook with operational definitions, and evidence potentially contradicting the findings was explicitly presented. Comparisons with selected reference countries served as secondary evidence, analyzed alongside the quantitative results to assess the realism of various policy models while accounting for cultural and regional contexts.

Certain limitations are openly acknowledged. The results are associative in nature, as PISA is a cross-sectional study that does not track individuals over time, precluding the determination of causal directionality. The mapping of curriculum eras relied on proxies based on the dominant curriculum. The transition in testing modes and the pandemic-related disruptions during the final cycle were addressed through control variables but could not be eliminated. The voices of the schools and teachers implementing the curriculum on a daily basis have not been represented, as all the resource persons came from the level of national policy and assessment.

3. Result

Cross-school nested structure

An unconditional model (a model without predictors) was estimated first to confirm the relevance of the data's nested structure (Peugh, 2010). Intraclass correlation coefficient (ICC) values ranged from 0.28 to 0.51, even exceeding 0.50 in 2012 and 2018. This indicates that nearly half of the score variation stems from between-school differences, far above the OECD average, where such variation typically accounts for about one-third. In practical terms, two students with similar family backgrounds could achieve vastly different results primarily because they attend different schools. The fact that ICC values remained stable despite multiple curriculum changes offers an early indication that the issue lies not in the instructional content itself, but rather in the differing conditions of the learning environments, pointing to a relatively stratified system (Willms, 2010).

Cross-cycle factor in the per-cycle model

Sequential analysis across the eight cycles reveals consistent patterns. Socio-economic status shows a positive and significant association across nearly all cycles, strengthening to approximately 13 points per index unit in 2018, the first cycle in which Indonesia fully transitioned to computer-based testing; the heightened impact of family economic background at this juncture aligns with the hypothesis that computer-based exams require digital familiarity that is unevenly distributed. A history of grade retention, observable only from 2012 onwards, emerged immediately as one of the strongest negative predictors of achievement, with an effect size ranging from roughly 20 to 33 points and remaining consistently negative across the four cycles for which data were available. School location revealed an urban-rural gap that widened in tandem with city size, reaching approximately 59 points in 2018. In contrast, school size consistently showed a positive association, serving as an indicator of more established areas.

Several additional patterns are noteworthy. Student attachment to school was prominent in the earliest cycles, the classroom disciplinary climate strengthened during the middle of the period, and boys consistently lagged in reading across almost every cycle. Conversely, the direction of the school-type variable was unstable across cycles, flipping from positive to negative depending on sample composition, rendering it unreliable as a robust determinant. The coefficient for maternal education often appeared negative once socio-economic status and household resources were

included; this is best interpreted as a complementary factor whose influence was largely captured by other family variables, rather than as evidence that maternal education itself hinders achievement. It must be emphasized that the negative association linked to grade retention history should be viewed as an indicator of early academic lag rather than proof that repeating a grade diminishes ability, given that cross-sectional data cannot establish the direction of causality (Hattie, 2009).

Curriculum era in the combined model (H1 testing)

In the cross-cycle pooled model, the curriculum era coefficients do not follow a sequence based on curriculum recency. If the premise that newer curricula yield higher achievement were valid, one would expect to see a consistent trend where more recent eras are associated with higher scores. No such trend emerged. The 2013 Curriculum era, which has been in effect the longest and has absorbed the most policy resources, was actually the only era showing no significant association in any domain (mathematics, reading, or science). Conversely, the variables for population coverage and socioeconomic status showed strong, significant effects across all three domains. In contrast, the time trend showed a significant negative association with mathematics and science, mirroring the long-term stagnation that forms the study's backdrop. This overall pattern is consistent with H1: once confounding factors are controlled for, the curriculum era does not exhibit a strong or consistent relationship with achievement.

Sensitivity analysis

Sensitivity analysis yields nuanced insights. Two scenarios show stable results: removing the test mode control produces virtually identical coefficients, since test mode and curriculum era overlap, while excluding the 2015 cycle results in only a negligible shift in coefficients. Two other scenarios reveal particularly informative sensitivity. When the control for population coverage is removed, the coefficient for the 2013 Curriculum era surges from non-significant to significant in both mathematics and science. This surge demonstrates that, without controlling for coverage, the curriculum era would appear to have raised scores. In contrast, much of that increase actually stemmed from the expansion of the test's coverage. These findings underscore that population coverage is not merely a technical control but a safeguard preventing the expansion of coverage from being misconstrued as curriculum success. Consequently, the conclusion drawn is not that a specific era proved superior, but rather that curriculum changes, in general, do not exhibit a strong, consistent relationship with student achievement.

Persistence factor (H2 test)

Cross-cycle analysis tallies the number of times a factor was found to be significant across all cycles where data was available. Socioeconomic status and school size stand out, having proven significant in seven out of eight cycles across all domains. A history of grade retention was consistently significant, showing significance in all four cycles for which data existed. School location and student-teacher ratios were notable within specific domains. Based on these patterns, H2a and H2b are supported; H2c is supported, with the caveat that the proxy for teacher quality was limited; and H2d is partially supported because the direction of the school-type effect was unstable, even though specific categories consistently suppressed achievement in the combined model. These results reinforce the study's central focus on an asymmetric pattern: curriculum inputs, which are more removed from students, showed only weak and inconsistent associations, whereas structural inputs that directly impact students demonstrated stronger, enduring relationships across different eras.

Expert review and synthesis of three sources of evidence

Interviews were conducted not to refute quantitative patterns, but to explain why the curriculum-related association appeared weak on the ground. Two explanations recurred. First, the curriculum operates indirectly through teachers rather than directly on students; one interviewee noted that the curriculum was designed to alter teacher behavior, meaning its impact could not properly be assessed for several years (INF-4), while policymakers themselves emphasized that adapting to a new curriculum takes time (INF-1). Second, part of the score decline reflected changes in measurement methods rather than a drop in quality; PISA implementation managers explained that the shift in testing mode altered the construct being measured and that the expanded population coverage included more students from groups previously not captured (INF-3). Both explanations align with two variables directly controlled in the model, variables that also proved most decisive in the sensitivity analysis.

Regarding persistent factors, interviewees added contextual nuance that coefficients alone could not capture. A history of grade retention was interpreted as an indicator of a child having fallen behind early on, rather than the cause of that lag, and holding students back without additional support was deemed ineffective (INF-4; INF-5). Socio-economic status was identified as a factor beyond school control, operating through an ecosystem far broader than the classroom (INF-4; INF-6). The urban-rural gap was recognized as a consistent finding across all national assessment instruments (INF-1). At the same time, the negative coefficient associated with school type was interpreted as reflecting the low-cost private school segment, which serves many children from disadvantaged families (INF-4). A convergence matrix of three evidence sources identifies the equalization of quality across regions as the proposal with the strongest support; it rests on two mutually reinforcing persistent factors, enjoys balanced backing from both within and outside the government, and aligns with comparative lessons.

4. Discussion

The weakness of the curriculum as a lever

The pattern of results aligns with the mechanisms established in the theoretical framework. When policy signals from the macrosystem must traverse loosely coupled layers before reaching the classroom, and when the interval between changes is shorter than the time required for teacher adaptation, the gap between the intended and implemented curricula rarely closes, resulting in stagnant outcomes (Van den Akker, 2003; Weick, 1976). Substantive reform typically requires several years of sustained implementation before effects on practice become visible, and reforms abandoned before that point tend to alter documents without altering classrooms (Fullan, 2007). This interpretation is reinforced by findings showing that the most recent curriculum era exhibits the weakest association and that the coefficients are sensitive to controls for scope. It must be emphasized that these findings do not imply the curriculum is unimportant; rather, they suggest that overhauling curriculum documents is not the most effective lever for boosting achievement in the next round of resource allocation. This distinction is crucial to ensure the findings are not misconstrued as a license to disregard the curriculum, but rather as a rationale for redirecting reform efforts toward more productive areas.

Persistence of structural factors

Socio-economic status, school location, and a history of grade retention persist across eras because they either directly affect the student or constitute a stable endowment that is not reset whenever curriculum documents are revised. The consistency of socio-economic status aligns with cross-national findings identifying socio-economic background as one of the strongest predictors of achievement (Sirin, 2005), while significant inter-school variation signals a stratified system (Willms,

2010). The pattern also echoes the classic finding that family background outweighs measured school resources in explaining achievement (Coleman et al., 1966), though with one important qualification: the high between-school variance recorded here indicates that in a stratified system schools differentiate outcomes far more than in the context Coleman studied. School location and socio-economic status are best understood as two sides of the same structural inequality: children in underdeveloped areas typically come from families with limited economic means and attend schools with fewer resources, creating a mutually reinforcing effect. This explains why the urban-rural gap has not narrowed despite repeated curriculum changes: the issue lies not in what is being taught, but in the inequality of the settings where that teaching takes place.

International relevance and comparative lessons

The experiences of several benchmark countries reinforce, rather than displace, conclusions drawn from domestic data. Vietnam, with a comparable income level, achieved far superior results through stable policies, robust investment in early childhood education, and low grade-retention rates, rather than through frequent curriculum changes. Malaysia demonstrates that a consistent long-term blueprint must still be accompanied by explicit attention to equity, as socioeconomic disparities can persist even within a stable policy framework. Japan maintains a long-cycle approach to curriculum revision, upholding quality through a culture of improving teaching practices rather than frequent document overhauls. The common thread aligns with this study's findings: higher achievement is more closely linked to policy stability, early learning support, and the quality assurance of teaching practices.

Cross-country comparisons require methodological caution. Differences in the timing of testing-mode transitions and population coverage can render score comparisons misleading if overlooked; for instance, a country's superior performance in a given cycle might be partly attributable to the fact that it had not yet undergone the transition-related disruptions experienced by other nations. Consequently, lessons from abroad cannot be cherry-picked without considering the structural readiness of each respective system. For international readers, the most transferable lesson is not to replicate a specific curriculum, but rather to examine whether document overhauls are consuming policy energy in their own systems. At the same time, more decisive structural factors remain unaddressed.

Policy Implications

The study's key implication is that limited reform resources could yield greater returns if policy priorities shifted from revising curriculum documents to addressing the structural foundations that have proven most decisive. This shift can be understood through the systemic pattern known as "shifting the burden": when a system faces severe, deeply rooted problems, decision-makers tend to opt for quick fixes that alleviate immediate pressure rather than fundamental solutions that take longer to bear fruit, thereby increasing the system's reliance on superficial remedies that fail to cure the root cause. Repeated curriculum overhauls resemble such quick fixes, whereas ensuring equitable quality and strengthening learning foundations represent fundamental solutions that have been deferred.

Based on triangulated evidence, the priority program can be structured as a multi-layered architecture. At the foundational level, curriculum stability is maintained so that resources previously consumed by overhauls can be redirected toward strengthening teaching practices. Above this, ensuring quality equity across regions serves as the overarching program, as it simultaneously addresses the two most persistent factors: location and socioeconomic status. This umbrella initiative rests on three pillars: an early detection and learning support system to replace the practice of grade retention without intervention; the strengthening of formative teacher feedback

as a more productive lever for quality than merely reducing class sizes; and the refinement of cross-sectoral socioeconomic support. An additional agenda item, regulating low-cost private schools, is designated for further study, as the supporting evidence is not yet sufficiently robust.

The program formulation outlined above is a synthesis derived from the three sources of evidence rather than a proposal articulated verbatim by the informants; the informants provided the strategic direction, feasibility assessments, and implementation guidelines. Four principles ensure the proposal does not become a collection of isolated suggestions: addressing root causes rather than symptoms, utilizing existing budgetary channels instead of establishing new institutions, focusing energy on a single umbrella program rather than dispersing it across multiple initiatives, and institutionalizing the agenda within long-term planning to ensure its continuity beyond leadership transitions (Cairney, 2016). These implications are framed as viable policy considerations, rather than mandatory prescriptions, and are presented to allow for an assessment of their applicability to other education systems facing similar challenges.

5. Conclusion

This study examines whether shifts in curriculum eras correlate with changes in Indonesia's PISA scores between 2000 and 2022 and identifies factors that remain consistent across these eras. After controlling for testing modes, population coverage, and time trends, curriculum eras showed no strong or consistent relationship with achievement; notably, the most recent curriculum era showed no significant association in any domain and proved sensitive to the removal of coverage controls. Conversely, family socioeconomic status, school location, and a history of grade retention were consistently associated with scores across all eight cycles. These patterns suggest that achievement is driven more by structural conditions external to curriculum documents than by curriculum changes themselves, pointing policy priorities toward regional quality equity, supported by early detection, enhanced teacher feedback, and socioeconomic assistance, anchored in curriculum stability.

These findings carry several specific policy implications. First, the Ministry of Education should institute a moratorium on wholesale curriculum revision for a minimum of one full PISA cycle (approximately six years), redirecting the freed budgetary and administrative capacity toward continuous professional development for teachers in disadvantaged regions. Second, the persistent negative effect of grade retention suggests that an early warning and intervention system—combining diagnostic literacy and numeracy screening at grades 1–3 with targeted remedial support—would be more productive than allowing students to repeat grades without additional assistance. Third, the widening urban-rural gap calls for an affirmative resource allocation formula that channels additional funding, experienced teachers, and digital infrastructure to rural and remote schools, benchmarked against measurable equity targets rather than uniform per-capita distribution. Fourth, the sensitivity of curriculum-era coefficients to population coverage controls underscores the need for transparent reporting of coverage and mode changes in every future national and international assessment, so that score movements are not mistakenly attributed to policy interventions. Finally, the evidence on socioeconomic status as a persistent predictor warrants cross-ministerial coordination linking conditional cash transfer programs, school feeding schemes, and community health services to address the out-of-school determinants of learning that no curriculum document, however well designed, can overcome on its own.

Several limitations constrain the interpretation of these findings. The results are associative, as PISA is a cross-sectional study that cannot establish causality, and the mapping of curriculum eras relies on proxies based on the dominant curriculum. The qualitative review relied on informants at

the policy and national measurement levels, leaving the voices of schools and teachers unrepresented, and the assessment of low-cost private schools relied on limited sources.

Three avenues for future research are proposed. First, longitudinal or quasi-experimental designs tracking a single cohort from the early grades would strengthen the causal basis and measure curriculum impact at the stage where exposure remains "pure." Second, studies dissecting the low-cost private school segment while controlling for socioeconomic status would test whether and why their quality lags. Third, rigorous impact evaluations of refined affirmative action formulas in specific regions, leveraging national assessments disaggregated to the local level, would transform recommendations from evidence-based proposals into policies with proven impact. Overall, these findings convey a message that is simple yet requires patience to put into practice: improving education is less about finding the perfect curriculum document and more about the persistence required to build the most fundamental foundations and nurture them consistently.

6. References

- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. <https://doi.org/10.2307/j.ctvc4j72>
- Beatty, A., Berkhout, E., Bima, L., Pradhan, M., & Suryadarma, D. (2021). Schooling progress, learning reversal: Indonesia's learning profiles between 2000 and 2014. *International Journal of Educational Development*, 85, 102436. <https://doi.org/10.1016/j.ijedudev.2021.102436>
- Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. National Bureau of Economic Research.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press.
- Cairney, P. (2016). *The Politics of Evidence-Based Policy Making*. Palgrave Macmillan.
- Cantoni, D., Chen, Y., Yang, D. Y., Yuchtman, N., & Zhang, Y. J. (2017). Curriculum and ideology. *Journal of Political Economy*, 125(2), 338–392. <https://doi.org/10.1086/690951>
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. U.S. Government Printing Office.
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and Conducting Mixed Methods Research* (3rd ed.). Sage Publications.
- Dyrvold, A., & Bergvall, I. (2023). Computer-based assessment in mathematics: Issues about validity. *LUMAT: International Journal on Math, Science and Technology Education*, 11(3), 49–76. <https://doi.org/10.31129/LUMAT.11.3.1877>
- Fullan, M. (2007). *The New Meaning of Educational Change* (4th ed.). Teachers College Press.
- García, J. L., Heckman, J. J., Leaf, D. E., & Prados, M. J. (2020). Quantifying the life-cycle benefits of an influential early-childhood program. *Journal of Political Economy*, 128(7), 2502–2541. <https://doi.org/10.1086/705718>
- Hanushek, E. A., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 89–200). North-Holland. <https://doi.org/10.1016/B978-0-444-53429-3.00002-8>
- Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*.

Routledge.

- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, 26(13), 1753–1760. <https://doi.org/10.1177/1049732315617444>
- OECD. (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1), 85–112. <https://doi.org/10.1016/j.jsp.2009.09.002>
- Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4), 805–827. <https://doi.org/10.1111/j.1467-985X.2006.00426.x>
- Robitzsch, A., Lüdtke, O., Goldhammer, F., Kroehne, U., & Köller, O. (2020). Reanalysis of the German PISA data: A comparison of different approaches for trend estimation with a particular emphasis on mode effects. *Frontiers in Psychology*, 11, 884. <https://doi.org/10.3389/fpsyg.2020.00884>
- Schultz, T. W. (1961). Investment in human capital. *The American Economic Review*, 51(1), 1–17.
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417–453. <https://doi.org/10.3102/00346543075003417>
- Van den Akker, J. (2003). Curriculum perspectives: An introduction. In J. Van den Akker, W. Kuiper, & U. Hameyer (Eds.), *Curriculum Landscapes and Trends* (pp. 1–10). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-017-1205-7_1
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21(1), 1–19. <https://doi.org/10.2307/2391875>
- Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record*, 112(4), 1008–1037. <https://doi.org/10.1177/016146811011200408>