

The Quality of Budget Institutions and the Efficiency of Public Spending

Santiago Herrera Aguilera

sherrera@worldbank.org

World Bank

Hironobu Isaka

World Bank

Juan Fernando Gomez Lopez

World Bank

Research Article

Keywords: Budget Institutions, PEFA, Public Spending Efficiency, Machine Learning, Causal Forests, Dose-Response

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9751722/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

The Quality of Budget Institutions and the Efficiency of Public Spending

Santiago Herrera Aguilera, sherrera@worldbank.org, corresponding author
Hironobu Isaka, Hisaka@worldbank.org
Juan Fernando Gomez, Jgomezlopez@worldbank.org

Abstract

This study examines how the quality of budget institutions affects public spending efficiency in education, health, and infrastructure. Using Public Expenditure and Financial Accountability (PEFA) indicators to measure fiscal governance and non-parametric efficiency scores as outcomes, the analysis proceeds in two stages. XGBoost models first identify which governance dimensions most strongly predict efficiency: Predictability and Control in Budget Execution (Pillar V) consistently ranks highest across sectors, with effects concentrated in the low-to-mid range of the PEFA distribution. Since income and broader institutional quality confound these associations, the second stage applies Generalized Random Forests (GRF) to estimate Conditional Average Treatment Effects under a continuous dose-response framework. After conditioning on key confounders, a one-unit improvement in key PEFA pillar scores raises spending efficiency by two to nine percentage points depending on sector — with effects smooth, exhibiting diminishing returns at higher governance levels, and systematically larger in lower-income and institutionally weaker countries. Efficiency gains of this magnitude imply 0.6 to 1 additional learning-adjusted years of schooling and up to fourteen percentage-point increases in infrastructure quality indices

JEL Classification: H50, H51, H52, H54, H61, I1, I2, E62

Keywords: Budget Institutions; PEFA; Public Spending Efficiency; Machine Learning; Causal Forests; Dose-Response

1. Introduction

Public finance institutions, namely the rules, organizational structures, and processes governing the allocation and use of public resources, are widely believed to influence the efficiency of public spending. Effective fiscal governance ensures that funds reach their intended uses, minimizes waste, and promotes transparency and accountability. Yet despite its policy relevance, the empirical relationship between fiscal governance quality and spending efficiency remains understudied. Most existing work uses broad governance composites and relies on linear regression methods that impose restrictive functional form assumptions and cannot resolve the multicollinearity among governance dimensions (Herrera and Ouedraogo, 2018; Herrera, Isaka, and Ouedraogo, 2025).

This paper makes two contributions. First, it uses the Public Expenditure and Financial Accountability (PEFA) framework, consisting of thirty-one disaggregated indicators clustered into seven institutional pillars, to examine which specific dimensions of budget institution quality drive efficiency differences across countries. Second, it moves beyond associative evidence to estimate causal effects, using machine learning methods that accommodate the nonlinearities, interaction effects, and multicollinearity inherent in cross-country governance data.

The empirical strategy proceeds in two stages. In the first stage, we use XGBoost with SHAP-based feature importance (augmented by a Modified Index Position correction for correlated predictors) and Accumulated Local Effects (ALE) plots to characterize the association between PEFA pillar scores and efficiency in education, health, and infrastructure. This stage identifies which governance dimensions are most predictive and documents a robust nonlinear, threshold-like pattern: predicted efficiency rises sharply as PEFA scores improve from low to mid-range levels, with diminishing returns thereafter.

The second stage recognizes that this threshold pattern may reflect confounding rather than a causal governance effect. Income level and broader institutional quality are correlated with both PEFA scores and efficiency outcomes; a simple positive association cannot disentangle the independent contribution of fiscal governance. We therefore apply Generalized Random Forests (Athey, Tibshirani, and Wager, 2019) with a continuous treatment specification, which is appropriate because PEFA pillar scores take values from 1 to 7, not a binary treatment, to estimate dose-response CATEs conditional on observed confounders. A key diagnostic result is that after partialling out the effect of confounders, the raw positive correlation between PEFA scores and efficiency largely disappears, confirming that association overstates the governance-efficiency relationship and that isolating the causal component requires the identification strategy developed in Section 6.

The paper is organized as follows. Section 2 reviews the literature. Section 3 describes the data. Section 4 outlines the machine learning methodology. Section 5 presents the predictive results. Section 6 develops the causal identification strategy and reports CATE estimates. Section 7 discusses limitations. Section 8 concludes.

2. Literature Review

Institutional quality has been consistently linked to development outcomes. Acemoglu et al. (2001) established the foundational role of institutions in explaining cross-country income differences, and Pritchett (2021) finds that a composite of four World Bank governance indicators is the most statistically significant predictor of social progress in a global sample. In the fiscal domain, Frankel, Végh, and Vuletin (2013) show that better institutions increase the likelihood that countries adopt countercyclical fiscal policy, treating governance as endogenous and instrumenting with settlers' mortality rates.

Research on governance and spending composition documents that weaker institutions bias expenditure toward less transparent categories (Mauro, 1998), while Grigoli and Mills (2014) find a positive link between governance and infrastructure quality using GMM methods. Herrera and Ouedraogo (2018) estimate technical efficiency of public spending and examine individual WGI dimensions, finding meaningful associations with infrastructure efficiency; Herrera, Isaka, and Ouedraogo (2025) extend this to health, education, and infrastructure using Tobit regressions, assuming linearity throughout.

Only one prior study uses PEFA indicators to explain fiscal outcomes: Keita, Leon, and Lima (2021) find that stronger fiscal governance, particularly on transparency and accounting dimensions, is associated with lower sovereign credit spreads. Herrera and Isaka (2024) document a positive linear correlation between a PCA-based PEFA composite and technical efficiency scores (Supplementary Figures S1–S3). Neither study quantifies the causal impact of changes in budget institutions on efficiency outcomes, accommodates nonlinear functional forms, or addresses multicollinearity among individual pillars.

This paper fills these gaps by combining PEFA's institutional granularity with machine learning methods capable of uncovering nonlinear and heterogeneous relationships, and with causal inference that moves beyond linear association to estimate the efficiency gains from fiscal governance reform.

3. Data

3.1 Fiscal Governance: The PEFA Framework

The primary measure of fiscal governance quality is drawn from the PEFA framework, which assesses the performance of a country's public financial management system through thirty-one indicators organized into seven pillars: (I) Budget Reliability, (II) Transparency of Public Finances, (III) Management of Assets and Liabilities, (IV) Policy-Based Fiscal Strategy and Budgeting, (V) Predictability and Control in Budget Execution, (VI) Accounting and Reporting, and (VII) External Scrutiny and Audit. Each indicator is scored on a letter scale (A through D), transformed into integer values from 7 (best) to 1 (worst). Pillar scores are the mean of constituent indicator scores.

We use both the 2016 PEFA framework (80 countries) and, for robustness, a merged 2011–2016 dataset (111 countries before efficiency score matching). The 2016 framework introduced Pillar III; merging the samples requires dropping it for comparability. Results are robust across both samples (Supplementary Appendix 8). The merged dataset yields 69–95 country observations depending on the efficiency indicator; the 2016-only sample yields 49–64 observations. All main results use the 2016 sample.

[INSERT TABLE 1: PEFA 2016 and 2011 Framework Pillars and Indicators]

A notable feature of PEFA scores is their high correlation with GDP per capita and with the WGI composite. This co-movement motivates both the multicollinearity-corrected attribution approach in Section 5 and the confounder-adjustment strategy in Section 6.

3.2 Public Spending Efficiency

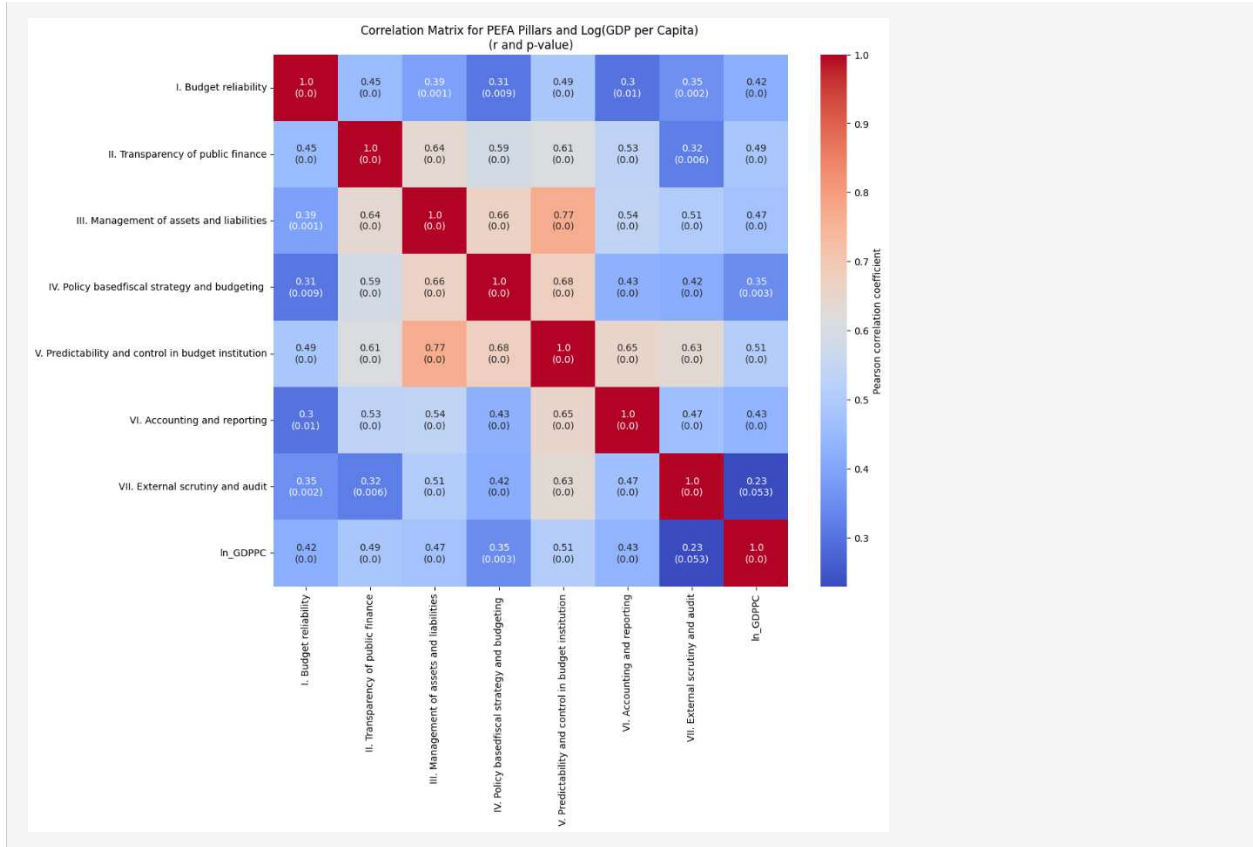
Efficiency scores are drawn from Herrera, Isaka, and Ouedraogo (2025), who apply non-parametric Data Envelopment Analysis to estimate technical efficiency of public spending in education, health, and infrastructure. Public spending per capita (PPP-adjusted) is the input in all three sectors. Output indicators vary by sector: education uses level of education and skills (WEF), youth literacy rate, and learning-adjusted years of schooling; health uses life expectancy, adult mortality rate, and infant survival rate; infrastructure uses WEF quality of overall infrastructure and World Bank Logistics Performance Index. Efficiency scores range from 0 to 1, with higher values indicating better performance relative to the estimated efficiency frontier. Summary statistics and variable definitions are in Supplementary Appendix 1. Average efficiency in our sample is 0.79 in health, 0.66 in education, and 0.47 in infrastructure.

4. Machine Learning Methodology

4.1 Motivation

Traditional regression faces three limitations in this setting. First, the seven PEFA pillars are correlated, with pairwise correlations ranging from 0.23 to 0.77 (Figure 1), originating multicollinearity that destabilizes coefficient estimates and makes it difficult to isolate the independent contribution of specific governance dimensions. Second, standard regression assumes linearity; nonlinear relationships require pre-specified functional forms. Third, the combination of sixteen efficiency measures and seven governance pillars creates a high-dimensional attribution problem.

Figure 1
Correlation Matrix of PEFA Pillars and GDP per Capita



Machine learning methods are data-driven and impose no functional form restrictions. We use XGBoost, a gradient-boosted decision tree ensemble that corrects residual errors sequentially and incorporates L1 and L2 regularization. XGBoost exhibits more stable out-of-sample performance than Random Forest in this small-sample setting (Supplementary Appendix 4). To address overfitting, we use repeated 4-fold cross-validation (3 repeats, 12 evaluations per model). Model performance metrics are in Supplementary Table A4.1.

4.2 Feature Attribution: SHAP with MIP Adjustment

Feature importance is computed using SHAP (Shapley Additive Explanations) values. However, when predictors are correlated, standard SHAP values are unstable: the apparent importance of a given pillar depends on which correlated pillars happen to be included. We quantify instability using the Normalized Movement Rate (NMR) and apply the Modified Index Position (MIP) adjustment (Salih, 2024) to detect instability at $NMR > 3$, reallocating importance to features that consistently gain explanatory power when correlated counterparts are excluded. Technical details are in Supplementary Appendix 4; the NMR diagnostic is in Supplementary Figure S7.

4.3 Marginal Effects: Accumulated Local Effects Plots

To characterize the functional form of the governance-efficiency relationship for the most predictive pillar in each sector, we use Accumulated Local Effects (ALE) plots. This choice is important because the PEFA pillars are strongly correlated, creating multicollinearity that complicates interpretation of marginal effects in the nonlinear XGBoost model. Unlike Partial Dependence Plots, ALE conditions on the local data distribution, thereby reducing reliance on unrealistic feature combinations and avoiding extrapolation to regions of the covariate space with little or no support. Each feature's range is partitioned into 5–8 bins (maintaining 8–9 observations per bin given sample sizes of 51–69 countries), and effects are estimated by comparing predictions at bin boundaries. Effects are expressed as deviations from the sample mean prediction, isolating the marginal association of each pillar while accounting for the observed correlation structure among PEFA pillars.

5. Association Between Fiscal Governance and Efficiency

5.1 What the XGBoost vs. Linear Comparison Reveals

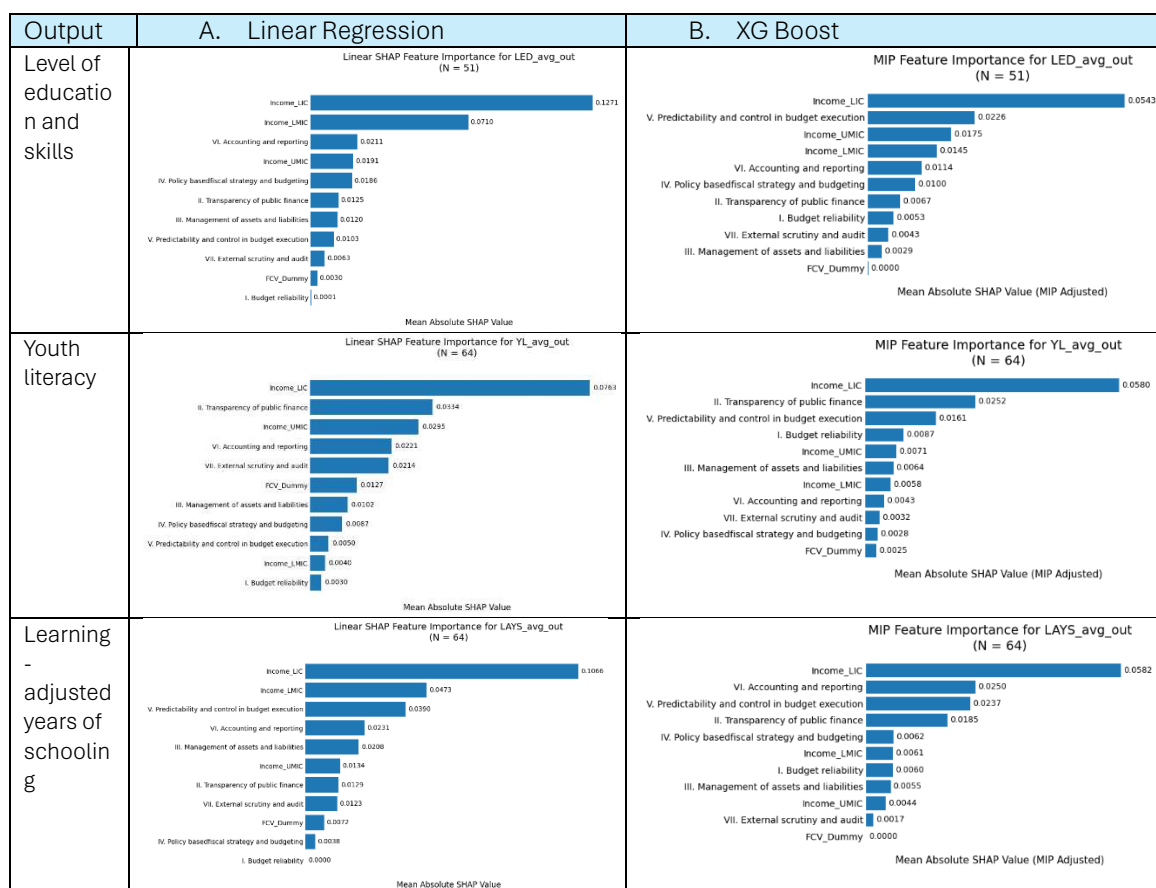
A methodological contribution of this paper is demonstrating when and why the machine learning approach produces materially different conclusions from linear regression. The comparison is not primarily about predictive accuracy, as out-of-sample R^2 values are broadly similar between OLS and XGBoost (Supplementary Table A4.1), but about attribution reliability. In linear specifications, income-group dummies dominate feature importance rankings because they absorb variance that, in a nonlinear setting, is attributable to correlated governance dimensions. The MIP-adjusted XGBoost models shift individual PEFA pillars upward by as many as seven to nine rank positions relative to OLS, while income dummies fall by comparable magnitudes (Supplementary Appendix 2).

The substantive implication is qualitative: linear regression suggests income level is the primary determinant of efficiency, leaving little residual role for fiscal governance. The corrected machine learning approach shows that governance dimensions such as Pillar V and, in certain sectors, Pillars II and III, have independent predictive power that linear specifications mask due to their correlation with income. The consistency of these attribution results across both the 2016-only sample and the merged sample of approximately 120 countries, combined with careful cross-validation-based hyperparameter tuning to prevent overfitting, reinforces confidence that the reranking reflects genuine signal rather than small-sample instability. This distinction matters: if income were the only driver, recommendations to strengthen PEFA performance would carry little empirical weight.

5.2 Education

Across all three education indicators, level of education and skills (LED), youth literacy (YL), and learning-adjusted years of schooling (LAYS), Pillar V (Predictability and Control in Budget Execution) moves into the top three after MIP adjustment, with rank improvements of four to six positions. Income dummies dominate the linear specification but fall sharply in rank after correction (Supplementary Tables A2.1–A2.3).

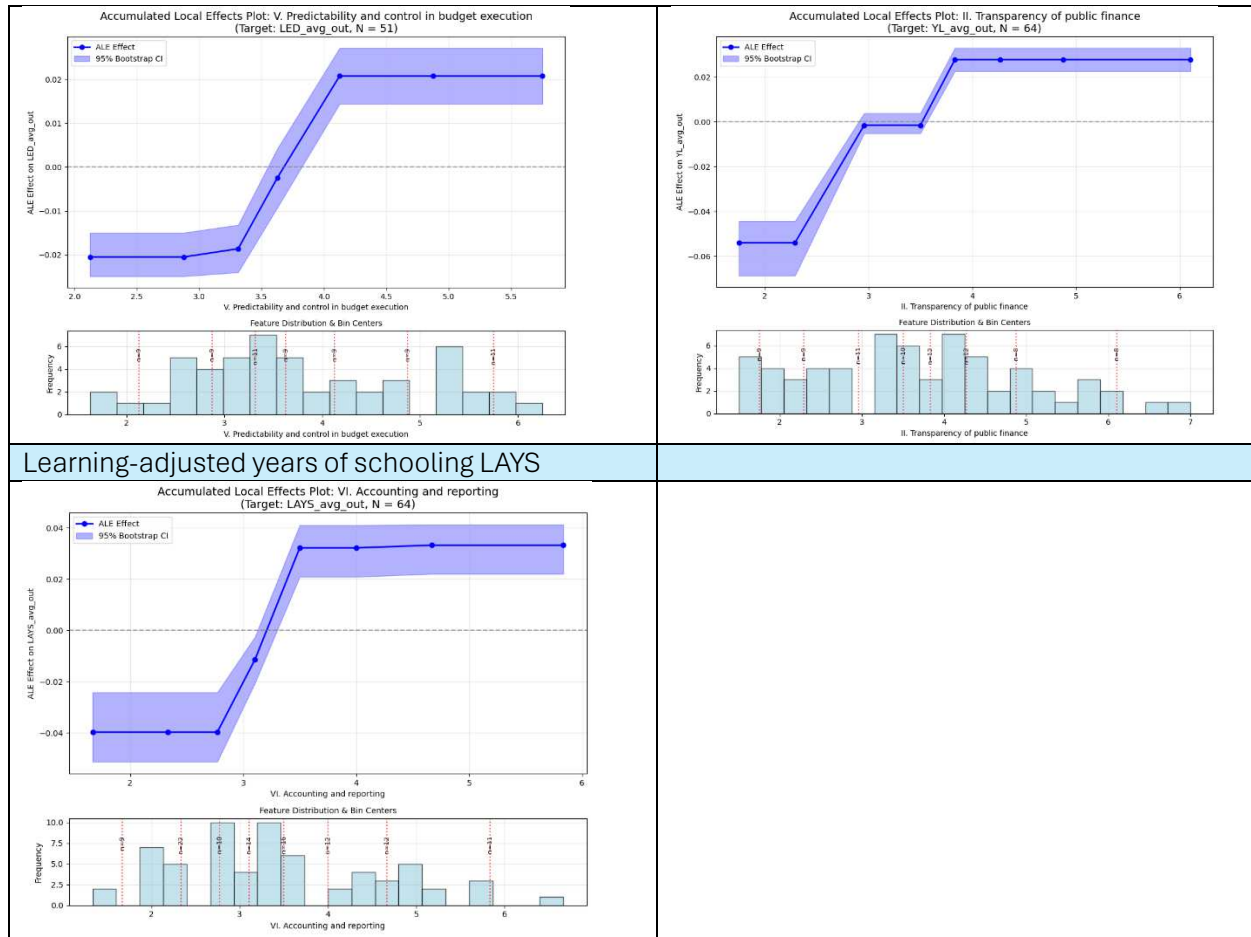
Figure 2. SHAP-Based Feature Importance for Education Spending Indicators: Results from Linear Regression and MIP-Adjusted XGBoost



ALE plots for Pillar V show a consistent threshold pattern across education indicators: predicted efficiency is approximately flat below scores of 3.5, rise sharply between 3.5 and 4.5, then plateaus. The net ALE improvement from the low to high end of the score range is approximately 0.030 units for LED and 0.023 units for LAYS.

Figure 3. Accumulated Local Effects of the Most Important Feature across Three Education Indicators

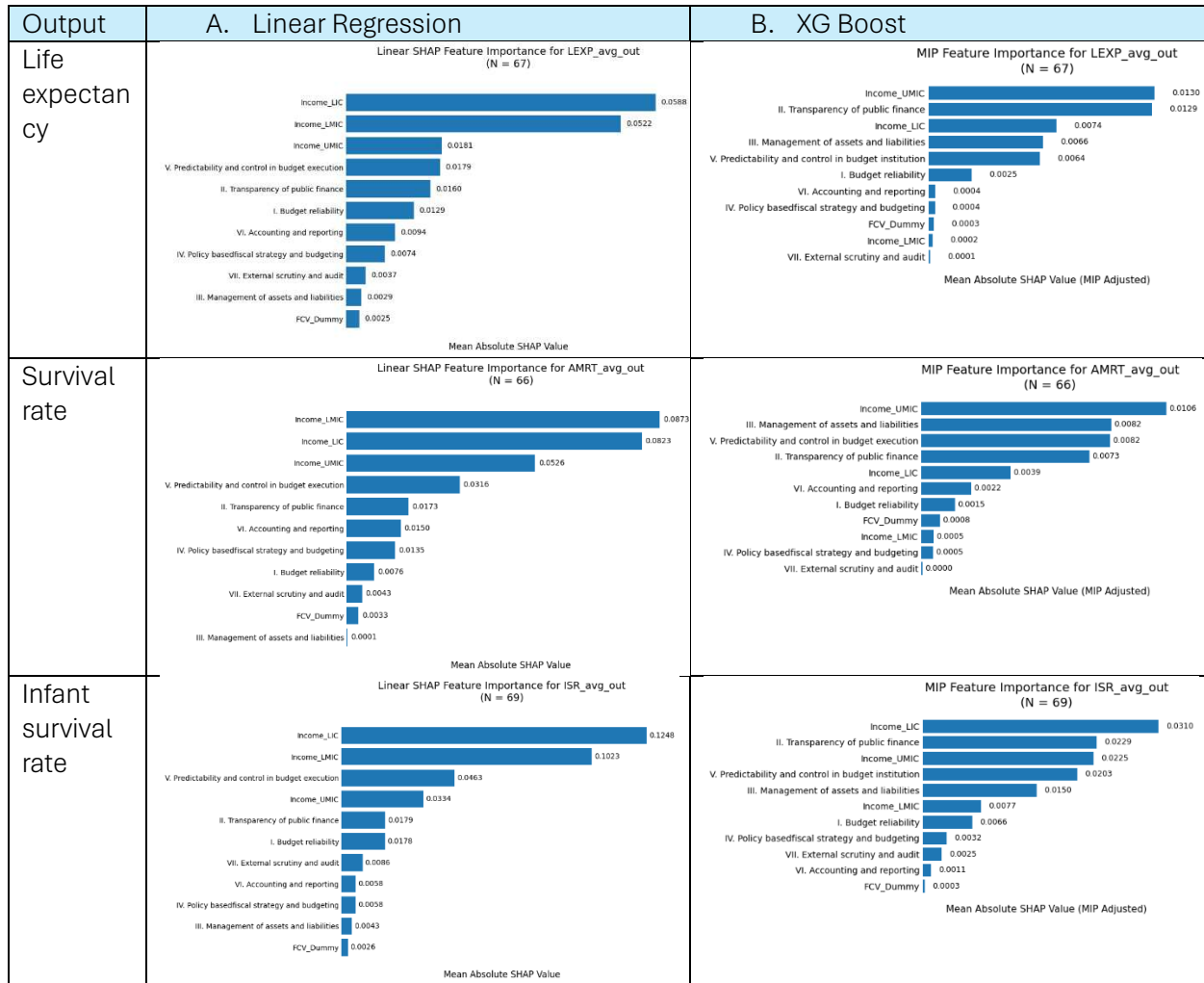
Level of education and skills LED	Youth literacy YL
-----------------------------------	-------------------



5.3 Health

In health, Pillar II (Transparency of Public Finances) is the most consistently important governance dimension after MIP adjustment, ranking among the top three for all three indicators (life expectancy, adult mortality rate, infant survival rate). Pillar III (Management of Assets and Liabilities) rises by up to nine positions for the adult mortality rate, while income dummies fall by up to eight positions. This is consistent with evidence that fiscal transparency improves budget execution rates in health, ensuring resources reach service delivery points (IMF, 2022).

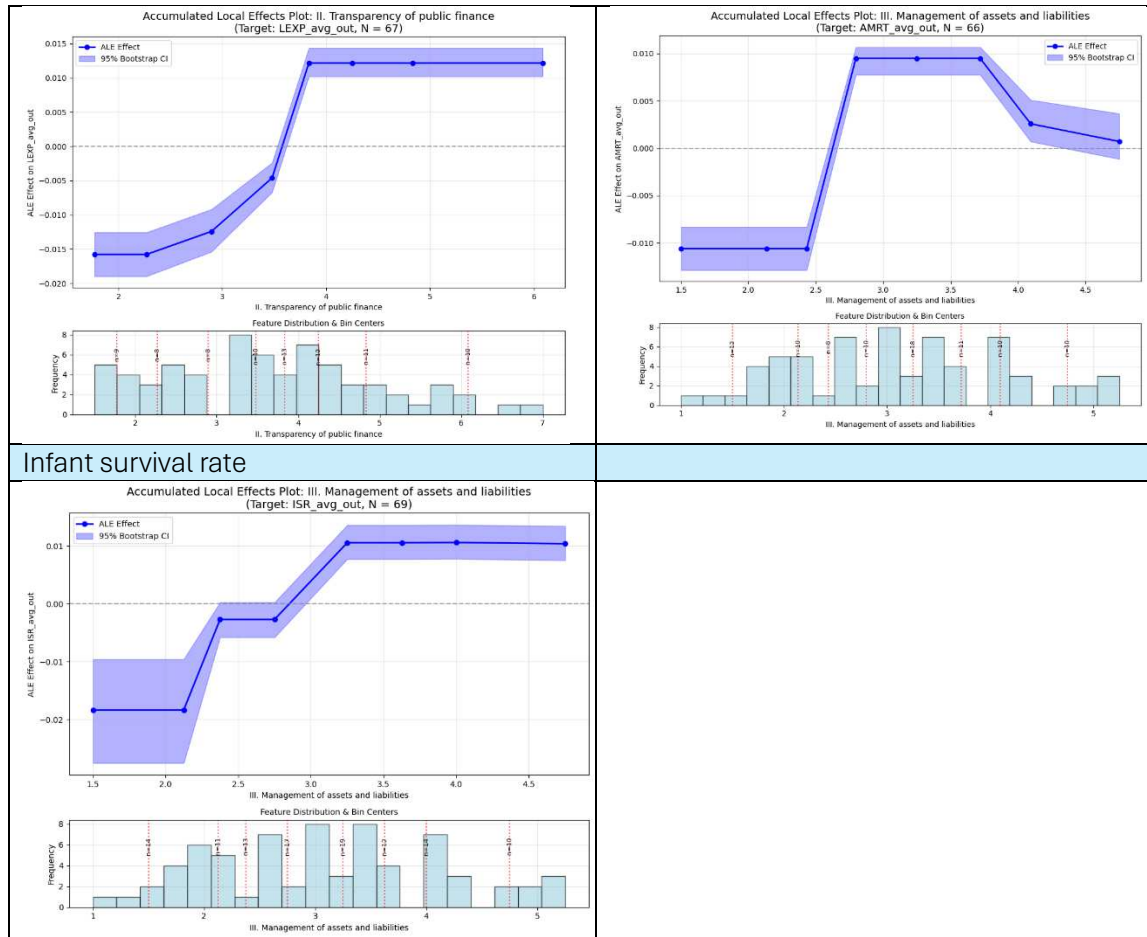
Figure 4. SHAP-Based Feature Importance for Health Spending Indicators: Results from Linear Regression and MIP-Adjusted XGBoost



ALE plots for Pillar II show a sharp transition near score 3.5 for life expectancy (net improvement approximately 0.025 units) and a steeper sustained improvement for the infant survival rate.

Figure 5. Accumulated Local Effects of the Most Important Feature across Three Health Efficiency Indicators

Life expectancy	Adult Survival rate
-----------------	---------------------



5.4 Infrastructure

For infrastructure, Pillars V and III lead in MIP-adjusted importance, with gains of seven to nine positions for quality of overall infrastructure and the Logistics Performance Index, respectively. ALE plots show a pronounced threshold near score 3.5 for quality of overall infrastructure (net improvement approximately 0.04 units) and an inverted-U pattern for the LPI under Pillar III, where the effect declines at higher values, suggesting diminishing returns from asset management complexity beyond an optimal range.

Figure 6. SHAP-Based Feature Importance for Infrastructure Spending Indicators: Results from Linear Regression and MIP-Adjusted XGBoost

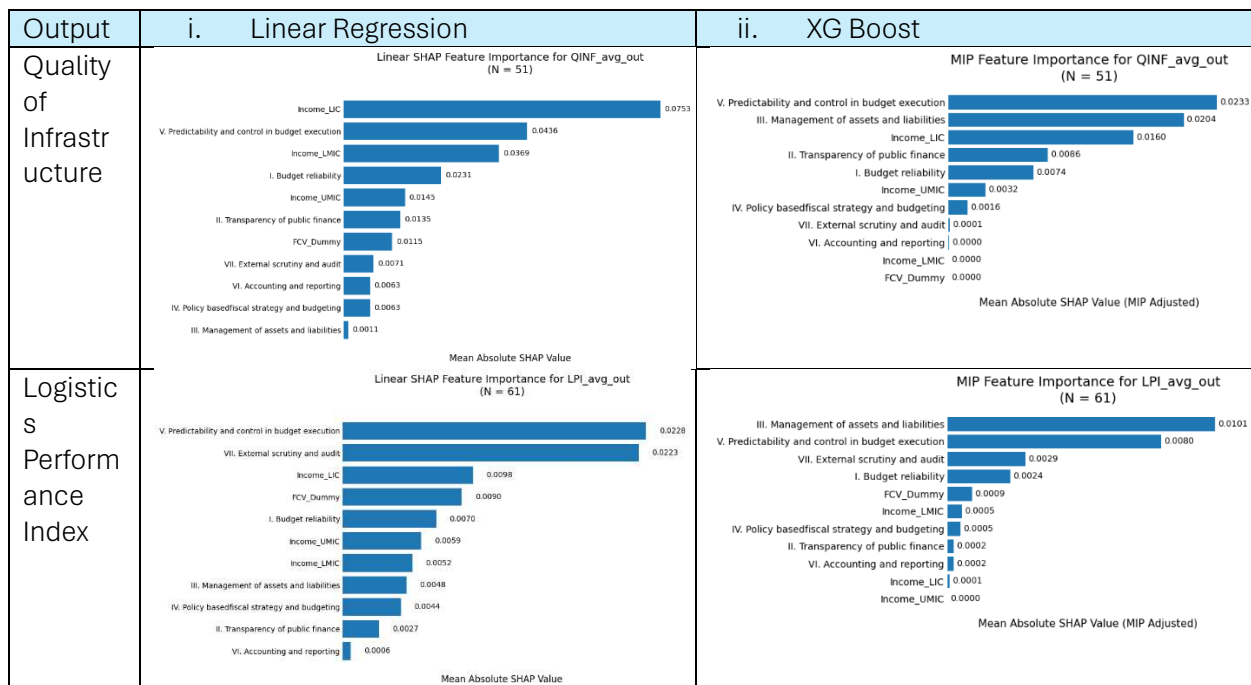
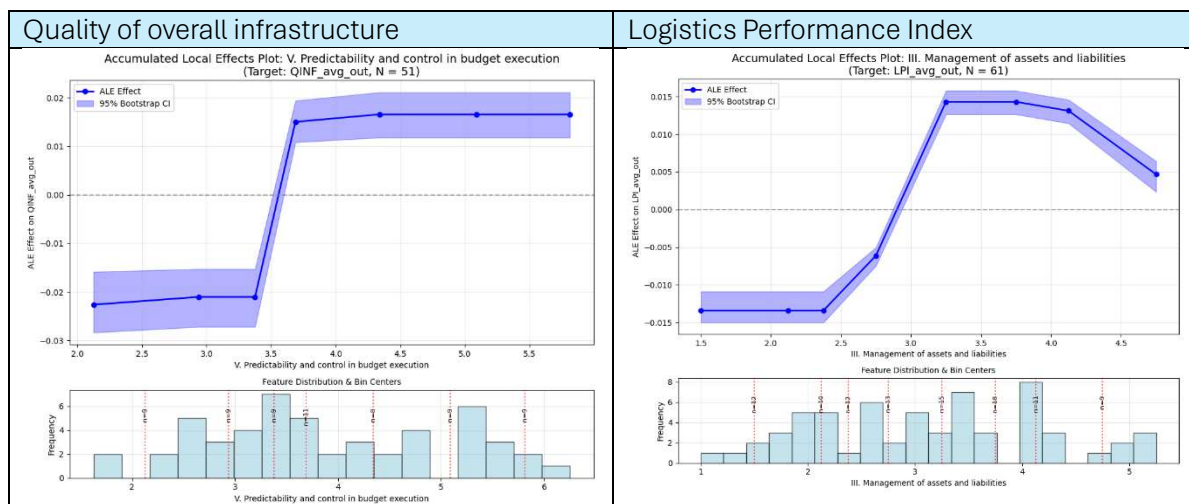


Figure 7. Accumulated Local Effects of the Most Important Feature across Two Infrastructure Indicators



5.5

Summary and Transition to Causal Analysis

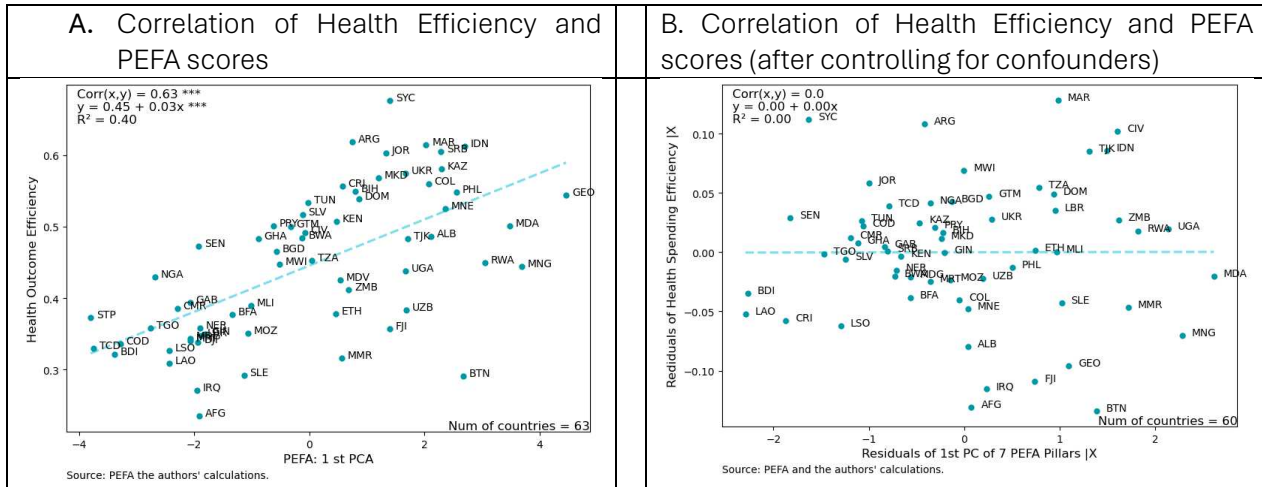
Across all three sectors, Pillar V (Predictability and Control in Budget Execution) is the most consistently predictive governance dimension after MIP adjustment. Pillar II (Transparency) is prominent in health and education; Pillar III (Management of Assets and Liabilities) is prominent in infrastructure and health, consistent with the centrality of asset maintenance in both sectors. The

nonlinear ALE patterns, with efficiency gains concentrated in the low-to-mid PEFA score range, are robust across sectors and output indicators.

These findings establish an important empirical baseline, but they are associative: the XGBoost models predict efficiency using PEFA scores without resolving whether higher governance quality causes higher efficiency or whether the correlation reflects common dependence on income level and broader institutional development. The co-movement of PEFA scores with GDP per capita, the Human Capital Index, and WGI dimensions means that simple association overstates the causal contribution of fiscal governance. Figure 8 illustrates this directly: the panel on the left shows the raw positive correlation between PEFA scores and health efficiency; the panel on the right shows the same relationship after both variables are residualized on confounders, where the correlation largely disappears.

This result does not contradict the Section 5 findings; prediction of efficiency levels benefits from using both PEFA scores and income-level information together. But it does signal that the causal impact of improving fiscal governance, while holding income, human capital, and other institutional factors constant, requires a different identification approach. Section 6 develops that approach using Generalized Random Forests with a continuous treatment specification.

Figure 8. Correlation of Outcome and Treatment Variable



6. The Causal Impact of Fiscal Governance on Spending Efficiency

6.1 Identification Framework

We adopt the potential outcomes framework for a continuous treatment (Hirano and Imbens, 2004; Imbens, 2000). Because PEFA pillar scores take values on a discrete 1–7 scale, the treatment is multi-valued, not binary. The binary notation $Y(1) - Y(0)$ is not applicable here; it does not generalize directly to the multi-valued case. The relevant estimand is the dose-response function:

$$\mu(t, x) = E[Y_i | T_i = t, X_i = x]$$

where Y_i is spending efficiency, $T_i \in [1, 7]$ is the PEFA pillar score (treatment), and X_i is a vector of confounding covariates. The object of interest is the Conditional Average Treatment Effect (CATE) for a one-unit increment:

$$\tau(x) = E[Y_i(t+1) - Y_i(t) | X_i = x]$$

The sample average of $\tau(x)$ over observed country characteristics is the CATE reported in Tables 4–6. Identification rests on generalized unconfoundedness (Hirano and Imbens, 2004):

$$Y_i(t) \perp\!\!\!\perp T_i | X_i \quad \forall t$$

This assumption requires that, conditional on the observed country characteristics in X_i , the assignment of PEFA scores is as good as random. We cannot verify this assumption for unobserved factors, and reverse causation — countries that improve efficiency becoming more likely to undertake governance reforms — is a genuine concern. Section 7 discusses these limitations and their implications for interpreting results.

6.2 Confounder Selection

We select confounders using the double-selection LASSO approach of Belloni, Chernozhukov, and Hansen (2014): LASSO is applied separately to the outcome (efficiency) and the treatment (PEFA pillar score) and the union of selected variables is retained. Candidate variables include log GDP per capita, the Human Capital Index (HCI), and six World Bank Governance Indicators (Voice and Accountability, Political Stability, Government Effectiveness, Regulatory Quality, Rule of Law, Control of Corruption). Selection is conducted independently for each sector.

[INSERT TABLE 3: Selected Confounding Variables by Sector]

Government Effectiveness, Regulatory Quality, and HCI are selected across all three sectors. Education and health additionally include Political Stability and Voice and Accountability. This selection is data-driven and transparent, though it cannot fully rule out bias from factors not in the candidate set.

6.3 Generalized Propensity Score and Overlap

For a continuous treatment, identification requires the generalized propensity score (GPS): the conditional density of the observed treatment given covariates, $r(t | x) = f_{T|X}(t | x)$. We estimate the GPS via beta regression after rescaling pillar scores to the unit interval following Cribari-Neto and Zeileis (2010). The beta regression yields good fit (log-likelihood 30.50; LR $\chi^2(6) = 53.3$, $p < 0.001$) with estimated precision parameter $\phi = 3.20$. Full estimation details are in Supplementary Appendix 6.

The GPS distribution is bell-shaped with no extreme mass at the tails (mean 1.51, s.d. 1.09), indicating adequate overlap. We verify the balancing property following Hirano and Imbens (2004): splitting

countries into tertiles of the treatment and testing independence of each confounder from the pillar score conditional on the GPS. Across eighteen regressions, no treatment coefficient is statistically significant at the 5% level (Supplementary Appendix 6, Table 6.3), providing empirical support for unconfoundedness within the range of observed covariates.

6.4 Estimation: Generalized Random Forests

We estimate $\tau(x)$ using Generalized Random Forests (GRF; Athey, Tibshirani, and Wager, 2019) with the continuous treatment extension. GRF uses recursive partitioning with an honesty property — splitting tree construction and effect estimation across separate subsamples — to produce asymptotically valid confidence intervals for heterogeneous treatment effects without distributional assumptions. We separately estimate CATEs for each combination of PEFA pillar and efficiency outcome indicator. Results for the three most impactful pillars in each sector are reported in Tables 4–6.

6.5 Results

Health.

Table 4 shows that Pillar V (Predictability and Control in Budget Execution) yields the largest average CATE across health indicators: a one-unit improvement raises efficiency by 5.7 percentage points on average, with the largest single-indicator effect of 9 percentage points for the infant survival rate. Pillar III generates the second-largest average effect. An efficiency gain of this magnitude translates into approximately 1.5–3.5 additional years of life expectancy at birth.

[INSERT TABLE 4: CATE Estimates — Health Sector]

Education.

Table 5 shows broad-based efficiency gains across education indicators: all three output measures respond similarly to Pillar V improvements, with an average CATE of eight percentage points. This homogeneity suggests that stronger budget execution supports both access and quality of learning outcomes. An efficiency gain of 8.2 percentage points translates into approximately 0.4–1 additional learning-adjusted years of schooling.

[INSERT TABLE 5: CATE Estimates — Education Sector]

Infrastructure.

Table 6 shows the largest single-indicator CATE in the full analysis: a one-unit improvement in Pillar V raises efficiency for quality of overall infrastructure by 7.8 percentage points. The Logistics Performance Index is less responsive (2–3 percentage points on average), possibly because LPI captures logistics ecosystem quality extending beyond the fiscal governance domain. An efficiency

improvement of 7.8 percentage points for infrastructure quality implies an increase in the WEF index of up to 14%.

[INSERT TABLE 6: CATE Estimates — Infrastructure Sector]

These results have profound implications for public finances, given the fiscal savings they could generate over time and because, more importantly, they matter for developmental outcomes: such efficiency gains would accelerate progress toward the Sustainable Development Goals (SDGs).¹ For example, in health, measured by years of life expectancy at birth (Table 4), an efficiency improvement of up to 4.2% translates into an increase of between 1.5-3.5 years. In education, measured by learning-adjusted years of schooling, efficiency gains of up to 8.2% (Table 5) represent an increase of between 0.4-1 years. In infrastructure, where efficiency scores are lowest, the potential impact is even more significant; when measured as the Quality of Overall Infrastructure (Table 6), the output index will increase by up to 14%.²

6.6 Do Threshold Effects Survive Confounder Adjustment?

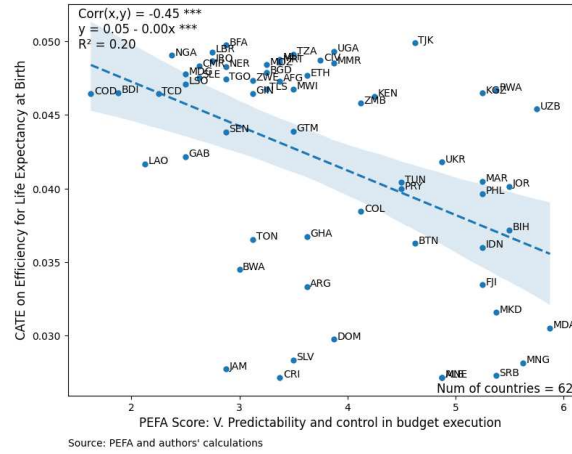
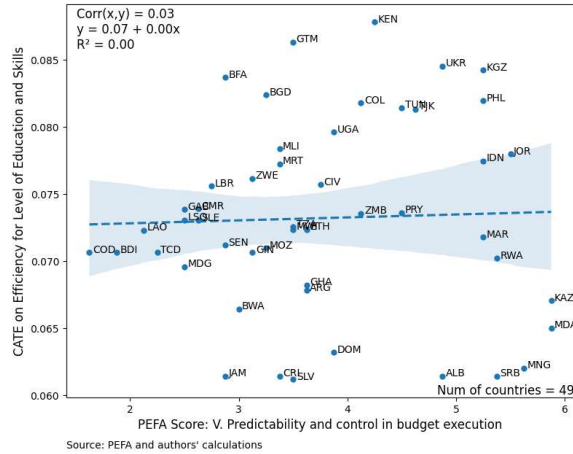
Figure 9 plots estimated CATEs against PEFA pillar scores. Two patterns stand out. First, the sharp threshold documented in the ALE plots of Section 5 is not present in the CATE estimates: effects evolve smoothly across the score range. This confirms that the threshold pattern in the associative analysis was partly attributable to confounding variables — particularly income level and government effectiveness — that change across PEFA score clusters, rather than to a genuine discontinuity in returns from fiscal governance reform. Second, there is a slight downward trend in CATE magnitudes at higher governance levels, consistent with diminishing returns: countries with already-strong fiscal institutions gain less from a marginal one-unit improvement.

Figure 9. CATE and PEFA scores in Different Sectors

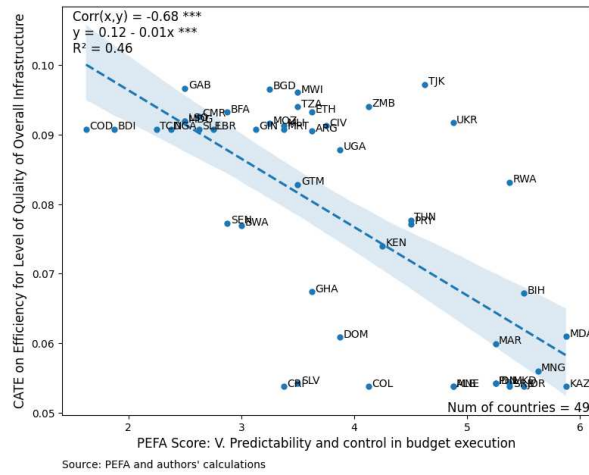
A. Education – Level of Education and Skills B. Health – Life Expectancy

¹ Since we are using output efficiency measures, the CATE shows how output indicators can increase while holding public spending constant. Appendix 12 describes the changes in output indicators implicit in the efficiency gains predicted by CATE estimates.

² Appendix 12 in the Supplementary Material file shows the descriptive statistics of all output indicators used in Tables 4-6. It shows the change in the output indicator implicit in the efficiency gain taking the three top pillars of Tables 4-6 in absolute terms and in percentage terms.



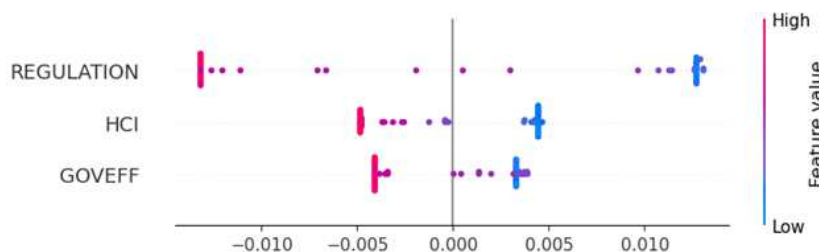
C. Infrastructure – Quality of Overall Infrastructure



6.7 Heterogeneity by Country Context

Figure 10 presents SHAP decompositions of CATE variance for infrastructure, the sector with the most pronounced heterogeneity. Regulatory Quality is the dominant driver: countries with weaker regulatory environments have above-average CATEs, while those with strong regulatory quality have below-average CATEs. Human Capital Index and Government Effectiveness show analogous patterns, validating diminishing-returns interpretation. Results for health and education (Supplementary Figures S5–S6) show similar patterns with somewhat weaker heterogeneity. Country-level waterfall charts illustrating the largest and smallest CATEs by sector are in Supplementary Appendix 9.

Figure 10. SHAP Values in Infrastructure – Quality of Overall Infrastructure



6.8 Robustness

Results are robust to the merged 2011–2016 PEFA sample: the same pillars retain the highest CATE estimates, and magnitudes change by less than 1.5 percentage points across frameworks (Supplementary Appendix 8). A Kolmogorov-Smirnov test on the 25 variables included in the analysis fails to reject equal distributions between samples at the 5% level for 21 of 25 variables, suggesting the additional countries are drawn from the same underlying distribution.

7. Limitations

Sample size and GRF precision. The 2016 PEFA framework yields 51–69 country observations depending on the efficiency indicator. We mitigate this through repeated cross-validation and regularization in the XGBoost stage, and through GRF's honest estimation in the causal stage; the robustness of core results to the merged sample, which nearly doubles the observation count, increases confidence that the small-sample constraint is not binding. GRF's honest splitting further reduces the effective sample in each leaf, limiting precision for fine-grained heterogeneity estimates; country-level CATEs should therefore be treated as indicative rather than precise.

Unconfoundedness and reverse causation. Causal identification relies on the assumption that all relevant confounders are observed and included in X_i . GRF estimates heterogeneous treatment effects under the conditional independence assumption rather than by algorithmically eliminating omitted variable bias; we verify plausibility through balance diagnostics and find broadly satisfactory results, and the honesty condition prevents overfitting. Reverse causation is a separate and more fundamental concern: countries that achieve efficiency improvements may be more likely to pursue fiscal governance reforms. The timing of PEFA assessments, often driven by donor coordination cycles rather than domestic policy milestones, partially mitigates this but cannot eliminate it; where upward bias from reverse causation is present, CATE estimates should be interpreted as upper bounds on the true causal effect. Future research using CPIA fiscal governance components, available as a panel, would permit a larger sample and IV-based identification strategies that could more cleanly resolve both concerns.

8. Conclusions and Policy Implications

This paper examines the relationship between the quality of budget institutions and public spending efficiency using a two-stage empirical strategy combining machine learning for robust association analysis with Generalized Random Forests for causal inference.

The associative analysis reveals that Predictability and Control in Budget Execution (Pillar V) is the governance dimension most consistently linked to efficiency across sectors, with Transparency (Pillar II) prominent in health and education and Management of Assets and Liabilities (Pillar III) prominent in infrastructure. The relationship is nonlinear: predicted efficiency gains are largest in the low-to-mid range of governance quality, with diminishing returns at higher levels.

The causal analysis finds that a one-unit improvement in key PEFA pillar scores raises spending efficiency by two to nine percentage points, depending on sector and output indicator. In health, this implies 1.5–3.5 additional years of life expectancy; in education, 0.4–1 additional learning-adjusted years of schooling; in infrastructure, up to a 14% increase in quality indices. Importantly, the threshold pattern of the associative stage does not survive confounder adjustment, causal effects evolve smoothly, confirming that part of the threshold effect in raw data reflects confounding by income and governance co-movement rather than a genuine discontinuity in reform returns. Impacts are systematically larger in lower-income and institutionally weaker countries.

Policy priorities should vary by country context. In low-income countries with PEFA scores at or below 3, the highest-return interventions are in basic budget execution discipline: reliable cash releases, commitment controls, and timely publication of execution data. In middle-income countries near the midrange (scores 3–4.5), gains come from institutionalizing predictability, expanding performance information in service delivery, and establishing public investment appraisal frameworks. In stronger institutional environments (scores above 4.5), marginal returns shrink; emphasis shifts to lifecycle asset management, strengthened audit follow-up, and open contract data.

These results carry important caveats. The causal interpretation relies on observed confounders fully capturing selection into governance quality, and reverse causation cannot be ruled out. The estimates should be used to prioritize reform directions and set conservative output targets, rather than as precise predictions for specific country contexts outside the sample. The combined findings nonetheless underscore that fiscal governance institutions are meaningful levers for improving the value of public spending, though they operate alongside, and interact with, a range of non-PFM factors, including staff capacity, geographic conditions, and socioeconomic context, that jointly determine spending outcomes; strengthening budget institutions is thus best understood as one high-return component of a broader reform agenda for accelerating progress toward the Sustainable Development Goals at lower fiscal cost.

TABLES

Table 1. Pillars and Indicators of the PEFA 2016 and 2011 frameworks

PEFA2016		PEFA2011	
I. Budget reliability		A. Credibility of the budget	
PI-01	Aggregate expenditure outturn	PI-01	Aggregate expenditure out-turn compared to original approved budget
PI-02	Expenditure composition outturn	PI-02	Composition of expenditure out-turn compared to original approved budget
PI-03	Revenue outturn	PI-03	Aggregate revenue out-turn compared to original approved budget
		PI-04	Stock and monitoring of expenditure payment arrears
II. Transparency of public finances		B. Comprehensiveness and Transparency	
PI-04	Budget classification	PI-05	Classification of the budget
PI-05	Budget documentation	PI-06	Comprehensiveness of information included in budget documentation
PI-06	Central government operations outside financial reports	PI-07	Extent of unreported government operations
PI-07	Transfers to subnational governments	PI-08	Transparency of inter-governmental fiscal relations
PI-08	Performance information for service delivery	PI-09	Oversight of aggregate fiscal risk from other public sector entities
PI-09	Public access to fiscal information	PI-10	Public access to key fiscal information
III. Management of assets and liabilities			
PI-10	Fiscal risk reporting		
PI-11	Public investment management		
PI-12	Public asset management		
PI-13	Debt management		
IV. Policy-based fiscal strategy and budgeting		C(i) Policy-Based Budgeting	
PI-14	Macroeconomic and fiscal forecasting	PI-11	Orderliness and participation in the annual budget process
PI-15	Fiscal strategy	PI-12	Multi-year perspective in fiscal planning, expenditure policy and budgeting
PI-16	Medium-term perspective in expenditure budgeting		
PI-17	Budget preparation process		
PI-18	Legislative scrutiny of budgets		
V. Predictability and control in budget execution		C(ii) Predictability and Control in Budget Execution	
PI-19	Revenue administration	PI-13	Transparency of taxpayer obligations and liabilities
PI-20	Accounting for revenue	PI-14	Effectiveness of measures for taxpayer registration and tax assessment

PI-21 Predictability of in-year resource allocation

PI-22 Expenditure arrears

PI-23 Payroll controls

PI-24 Procurement

PI-25 Internal controls on non-salary expenditure

PI-26 Internal audit

PI-15 Effectiveness in collection of tax payments

PI-16 Predictability in the availability of funds for commitment of expenditures

PI-17 Recording and management of cash balances, debt and guarantees

PI-18 Effectiveness of payroll controls

PI-19 Competition, value for money and controls in procurement

PI-20 Effectiveness of internal controls for non-salary expenditure

PI-21 Effectiveness of internal audit

VI. Accounting and reporting

PI-27 Financial data integrity

PI-28 In-year budget reports

PI-29 Annual financial reports

C(iii) Accounting, Recording and Reporting

PI-22 Timeliness and regularity of accounts reconciliation

PI-23 Availability of information on resources received by service delivery units

PI-24 Quality and timeliness of in-year budget reports

PI-25 Quality and timeliness of annual financial statements

VII. External scrutiny and audit

PI-30 External audit

PI-31 Legislative scrutiny of audit reports

C(iv) External Scrutiny and Audit

PI-26 Scope, nature, and follow-up of external audit

PI-27 Legislative scrutiny of the annual budget law

PI-28 Legislative scrutiny of external audit reports

Table 2. MIP-adjusted re-ranked feature importance

index	SHAP	MIP-Adjusted	SHAP Rank	MIP Rank	Rank Shift
Income_UMIC	0.012	0.011	1	1	0
III. Management of assets and liabilities	0.009	0.008	3	2	1
V. Predictability and control in budget institution	0.009	0.008	2	3	-1
II. Transparency of public finance	0.007	0.007	4	4	0
Income_LIC	0.003	0.004	5	5	0
VI. Accounting and reporting	0.002	0.002	6	6	0
I. Budget reliability	0.001	0.001	7	7	0
FCV_Dummy	0.001	0.001	8	8	0
Income_LMIC	0	0.001	11	9	2
IV. Policy based fiscal strategy and budgeting	0	0.001	9	10	-1

Table 3. Selected Confounding Variables in Each Sector

Sector	Variables
Education	- Government Effectiveness (GOVEFF) - Human Capital Index (HCI) - Regulatory Quality (REGULATION) - Political Stability (STABILITY) - Voice and Accountability (VOICE) - GDP per Capita (log) (ln_GDPPC)
Health	- Government Effectiveness (GOVEFF) - Human Capital Index (HCI) - Regulatory Quality (REGULATION) - Political Stability (STABILITY) - Voice and Accountability (VOICE) - GDP per Capita (log) (ln_GDPPC)
Infrastructure	- Government Effectiveness (GOVEFF) - Human Capital Index (HCI) - Regulatory Quality (REGULATION)

Table 4. Conditional Average Treatment Effects (Sample Average) in Health Sector³

Pillar	Life Expectancy	Infant Survival Rate	Survival Rate from Age 15-60
I. Budget reliability	0.027 (0.003)	0.047 (0.01)	0.019 (0.002)
II. Transparency of public finance	0.036 (0.003)	0.064 (0.011)	0.026 (0.006)
III. Management of assets and liabilities	0.038 (0.006)	0.078 (0.014)	0.028 (0.01)
IV. Policy based fiscal strategy and budgeting	0.025 (0.007)	0.049 (0.015)	0.014 (0.008)
V. Predictability and control in budget execution	0.042 (0.007)	0.092 (0.018)	0.036 (0.009)
VI. Accounting and reporting	0.021 (0.013)	0.057 (0.018)	0.020 (0.012)
VII. External scrutiny and audit	0.018 (0.008)	0.045 (0.013)	0.014 (0.006)

Standard deviations in parentheses.

Table 5. Conditional Average Treatment Effects (Sample Average) in Education Sector⁴

³ The average efficiency scores are 0.81, 0.73, and 0.84 for life expectancy, infant survival rate, and survival rate from age 15 to 60, respectively.

⁴ The average efficiency scores are 0.61, 0.85, and 0.52 for level of education and skills, youth literacy, and learning-adjusted years of schooling, respectively.

Pillar	Level of education and skills	Youth literacy rate	Learning-Adjusted Years of Schooling
I. Budget reliability	0.033 (0.006)	0.041 (0.015)	0.032 (0.008)
II. Transparency of public finance	0.050 (0.007)	0.067 (0.007)	0.050 (0.006)
III. Management of assets and liabilities	0.064 (0.006)	0.068 (0.007)	0.065 (0.012)
IV. Policy based fiscal strategy and budgeting	0.050 (0.006)	0.041 (0.01)	0.041 (0.005)
V. Predictability and control in budget execution	0.073 (0.007)	0.076 (0.011)	0.080 (0.008)
VI. Accounting and reporting	0.049 (0.006)	0.064 (0.012)	0.065 (0.005)
VII. External scrutiny and audit	0.033 (0.008)	0.046 (0.02)	0.040 (0.01)

Standard deviations in parentheses.

Table 6. Conditional Average Treatment Effects (Sample Average) on the Infrastructure Sector⁵

Pillar	Quality of overall infrastructure	Logistic Performance Index
I. Budget reliability	0.044 (0.009)	0.012 (0.005)
II. Transparency of public finance	0.046 (0.016)	0.012 (0.004)
III. Management of assets and liabilities	0.068 (0.018)	0.018 (0.013)
IV. Policy based fiscal strategy and budgeting	0.039 (0.017)	0.009 (0.004)
V. Predictability and control in budget execution	0.078 (0.017)	0.021 (0.008)
VI. Accounting and reporting	0.049 (0.015)	0.008 (0.008)
VII. External scrutiny and audit	0.032 (0.009)	-0.003 (0.008)

Standard deviations in parentheses.

⁵ The average efficiency scores are 0.55 and 0.38 for the quality of overall infrastructure and logistics performance index, respectively.