International Budget Partnership

Exploring the Determinants of Budget Credibility

Paolo de Renzio and Chloe Cho

June 2020

Paper

EXECUTIVE SUMMARY

Budgets are one of the key tools that governments use to turn policy intentions into concrete interventions to achieve their objectives. Based on complex planning and forecasting, budgets allow governments to project revenues and spending for the year(s) ahead and, in principle, to allocate resources according to priorities. They are also one of the main avenues through which governments report to citizens on their management of public resources. The extent to which governments stick to what is included in their parliament-approved budgets is what we refer to as the credibility of their budgets. Budget credibility is an important measure of a government's own overall credibility and their capacity to deliver on policy promises.

In another recent paper, our descriptive analysis of budget credibility in 35 countries found that on average these governments under-execute their budgets by almost 10 percent. The extent of underspending was even larger for capital expenditure and for specific government functions, suggesting that countries also suffer from lack of compositional credibility, or shift resources around substantially during budget execution.

In this paper, we examine some of the potential drivers of such deviations, using regression analysis on data from 120 Public Expenditure and Financial Accountability (PEFA) assessments conducted in 94 countries. We look at three different aspects of credibility: aggregate expenditure deviations, aggregate revenue deviations, and deviations in the composition of spending. Drawing on existing literature and other common propositions, we select and examine several explanatory variables related to national-level public financial management (PFM), and include a country's level of income per capita, democracy, and overall government effectiveness as control variables.

KEY FINDINGS

***More credible revenue projections are associated with lower expenditure deviations*.** Poor revenue forecasts and economic shocks are among the most common explanations provided for aggregate budget deviations in PEFA assessments. Our data show a statistically significant positive and high correlation between revenue and expenditure deviations, supporting the hypothesis that credibility on the spending side is affected by credibility on the revenue side.

Better overall quality of PFM systems is also associated with better credibility of expenditure. Weaknesses in PFM systems, from poor planning to challenges in budget execution, can lead to unexpected deviations from approved budgets. In a multivariable regression, composite PEFA scores – in addition to revenue deviations – exhibit a statistically significant relationship with the absolute value of aggregate expenditure deviations. The coefficient has a negative sign, as expected.

Notably, fiscal transparency and timely and comprehensive reporting appear critical for budget credibility. Our model, which includes the budget transparency and oversight scores from the Open Budget Survey, shows that budget transparency is associated with better credibility at the aggregate level. Looking at each of the six PFM pillars covered by the current PEFA framework, we find that "transparency of public finances" and "accounting and reporting" are the most significant factors. This may be because more information and timely reporting allow for more direct real-time control of how public resources are being used.

Poor planning and forecasting are associated with deviations on the revenue side. Capacity to run the necessary macroeconomic models and to limit political meddling is critical to adequate revenue planning. Lack of such capacity can lead to inaccurate or unrealistic projections, thereby increasing revenue deviations. In our multivariable models for revenue credibility, the difference between projected revenue and "expected" revenue – calculated using actual collection from last year and changes in the gross domestic product and inflation – was the only factor that displayed a statistically significant relationship. Some of the technical aspects of revenue forecasting, captured in the PEFA pillar "policy-based fiscal strategy and budgeting", are likely to be an important driver; but there may also be political incentives to over-estimate revenue.

The quality of PFM systems is also associated with deviations in the composition of expenditure. The PEFA pillar "predictability and control in budget execution" – which includes indicators on the effectiveness of payroll and procurement controls, predictability of in-year resource allocation, and internal audit – is important in explaining deviations in the composition of spending by both administrative/functional and economic classification. Fiscal transparency and rules over in-year budget adjustments made by the executive also seem to play a role in deviations in administrative/functional composition, while assets and liabilities management, external audit and procurement seem to play a role in economic composition.

Levels of income and democracy are not significant in most of our models. This does not mean that they do not play a role. However, this finding suggests that low-income countries, which on average underspend more than higher-income ones, might be able to improve their budget credibility through improving their PFM systems, which may be more amenable to reforms than a country's economic performance or political institutions.

RECOMMENDATIONS

What can governments and other actors do to improve budget credibility at different levels? They can:

Take actions to strengthen governments' revenue planning and forecasting function. This can include addressing gaps in technical capacities, key institutional bottlenecks and the political incentives to inflate revenue projections.

Strengthen other aspects of PFM systems that are associated with budget credibility. This includes accounting and reporting systems; managing public investment; budget execution (including procurement systems and virement rules); and external audit.

Improve fiscal transparency practices related to budget credibility. This could include specific measures like more disaggregated reporting on budget execution vs. original allocations, and providing adequate explanations and justifications for deviations, as IBP has suggested in previous work.

Carry out these reforms while not overlooking the importance of complementary government systems. Other systems not related to PFM may have a bearing on governments' capacity and willingness to improve its policy planning and implementation, and thereby its overall policy credibility.

CONTENTS

1. INTRODUCTION

The budget laws that countries' legislators pass every year lay out the policy intentions of a government for the 12 months to come. The resulting key documents include details on how resources will be collected through taxation, how such resources will be spent, and how much borrowing will be necessary to fill any financing gaps.

Research we presented in a previous paper[1], however, shows that governments often deviate substantially from the path indicated in these pieces of legislation. Data from that sample of 35 countries, from various regions and at different levels of development, indicate that on average countries underspend on their originally approved national budgets by about 10 percent. That average worsens to around 14 percent for the low-income countries we studied. Underspending affects areas that are very important for socioeconomic development, such as agriculture and social infrastructure, with a potentially negative impact on people's wellbeing.

These deviations – both in aggregate levels of spending and in the composition of spending – happen for a variety of reasons, not all of which can easily be investigated through cross-country research. In fact, previous research on this topic has been mostly anecdotal or limited to a small sample of countries. In this paper, we attempt to systematically examine some of the possible sources of poor budget credibility across a large number of countries, using data from Public Expenditure and Financial Accountability (PEFA) assessments. PEFA assessments offer a standardized evaluation of the quality of budget systems based on a common methodology that covers many aspects of public financial management (PFM). The first performance pillar includes data on budget credibility, or the difference between what was originally planned in terms of revenues, overall spending and composition of spending over a few years, and the actual outturn for each at the end of each year.

After summarizing prior work and laying out our analytical approach, we present our analysis for three aspects of budget credibility: (1) aggregate expenditure deviations, (2) revenue deviations, and (3) deviations in spending composition. Our findings generally support previous research and confirm many of our expectations. Based on the results, we provide some conclusions and recommendations, especially regarding strengthening PFM systems.

2. BRIEF REVIEW OF EXISTING EVIDENCE

Although practitioners have been aware of the problem for a long time, the literature on budget credibility is very limited. As previously suggested, much of what has been published is largely descriptive and looks at a single

[1] See de Renzio et al. (2019).

country or at a few countries at best.[2] Only a few studies analyze the determinants of budget credibility from a cross-country perspective.

Addison (2013) was the first to take advantage of data from PEFA assessments to identify correlates of levels of budget execution, but his analysis was limited by the number of assessments that had been carried out by that point in time and covered 45 countries. He looks at different possible explanations for aggregate budget deviations and a large number of variables, related to the existence of common pool behavior and pressures on public spending (e.g. based on demographics), to the nature of political institutions (e.g. democratic vs autocratic or presidential vs parliamentary systems), and to public financial management (e.g. PFM capacity). He finds that the accuracy of revenue projections, the quality of budget systems and political institutions that reduce common pool tendencies are all associated with lower levels of underspending.

Mustapha (2019) looks at the extent to which the quality of PFM systems can help explain budget deviations, and finds that better systems are associated with better budget credibility in nonfragile states. In states defined as fragile by the World Bank, even though the effect is weaker and insignificant for overall budget credibility, the relationship between the quality of PFM systems – as well as certain elements, such as budget execution systems – and compositional credibility is positive and statistically significant. Sarr (2015) investigates more specifically the role of fiscal transparency in improving the credibility of health and education spending across 73 countries, finding that it does have a positive impact, alongside other variables like GDP per capita and levels of democracy.

Finally, rather than test specific hypotheses about what might explain budget deviations, Mills (2018) looks at some of the explanations that are provided in PEFA assessments, themselves, for both expenditure and revenue deviations. Poor revenue forecasts and economic shocks were the most cited justifications, followed by challenges in the budgeting process, including the use of supplementary budgets or poor planning for capital spending, and the unreliability of donor funds.

3. OUR ANALYTICAL APPROACH

Topics: Based on our own previous findings, and on insights from others' studies on specific countries, this paper explores the determinants of budget credibility looking at three different types of deviations: (a) aggregate

[2] See, for example, Ablo and Reinikka (1999), Stasavage and Moyo (2000), Williams (2015), Simson and Welham (2014), Atiku and Lakin (2019).

expenditure deviations; (b) revenue deviations; and (c) deviations in the composition of spending, as measured through data provided for the relevant indicators in PEFA assessments.

Variables: Key explanatory variables used for each of the three analyses are introduced in the corresponding section. All models, though, use a consistent set of control variables:

a) a country's *level of income*, based on the finding from our previous paper that richer countries tend to have more credible budgets, and on the more general expectation that in richer countries the conditions (including better qualified human capital both within and outside government) under which governments formulate and execute public policies are more conducive to credibility. For this purpose, we use gross national income (GNI) per capita corrected for purchasing power parity (PPP), and expect it to have a positive impact on budget credibility (i.e. lower deviations);

b) the *strength of its democratic institutions*, hypothesizing that in more democratic countries governments will be held under more scrutiny to deliver on their plans.[3] We measure levels of democracy using Polity2 scores,[4] choosing this over other measures mostly because of the better data availability and for the purpose of capturing different regime types. And, we expect it to have a positive impact on budget credibility;

c) an *indicator of government effectiveness (GEI)*, to account for the existence of broader systems and institutions (i.e. beyond PFM) that may have an impact on budget credibility. The indicator we use for this purpose is part of the World Bank's World Governance Indicators; it tries to capture "perceptions of the quality of public services, the quality of civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies."[5] Assuming that perceptions can serve as a proxy for actual performance and good governance, we expect to find a positive relationship between this variable and budget credibility.[6]

Details about the variables used in this paper can be found in Annex Table 1. Correlation matrices are provided in Annex Table 2.

[3] It is interesting to note here that Addison (2013) highlights a potential opposite hypothesis, whereby in democracies the existence and interests of more numerous interest groups might generate more demands on the budget and lead to larger budget deviations.

[4] For the data and more information about the methodology, see Polity IV Dataset and Polity IV Users' Manual, available here: https://www.systemicpeace.org/polityproject.html.

[5] For the data and more information about the methodology, see http://info.worldbank.org/governance/wgi/.

[6] While Annex Table 2 shows strong positive correlations between GEI and some other variables – e.g. 0.491 with the composite PEFA score and 0.636 with log-transformed income per capita – diagnostics based on the variance inflation factor (VIF) show that there is no multicollinearity in regression models. As the R-squared values are higher for the models that include GEI, we retain it as a relevant control variable.

Sample: For this analysis, we collected and analyzed data from 120 PEFA reports released since 2012, covering budget years for the period 2007-2017. Most of these reports use the framework that the PEFA initiative released in 2011, but 40 of them are based on the updated 2016 framework, which includes some new indicators including "expenditure composition outturn by economic classification" that we consider as a dependent variable for the composition analysis. This sample includes roughly 360 country/year observations, covering 94 countries with between two and six years of data. The geographical spread of our data includes 39 countries in Sub-Saharan Africa, 16 in the Caribbean, 14 in East Asia and Pacific, 14 more in Europe and Central Asia, 6 in the Middle East and North Africa, and 5 in South Asia. As of 2017, 27 of these countries were classified as low-income, 36 as lower-middle-income, 26 as upper-middle-income, and only five as high-income.

Limitations: While PEFA assessments offer comprehensive data that are extremely useful, the time-series data are limited and inconsistent across countries. As such, we do not undertake panel analysis, and treat each country/year combination as an individual observation. In other words, our analysis is cross-sectional in nature. Accordingly, we rely on the Ordinary Least Squares (OLS) method to assess the relationship between our variables of interest, and we carry out the usual diagnostics for non-linear relationships, model misspecification, heteroskedasticity and multicollinearity. As widely recognized, OLS results should not be interpreted as showing causal relationships. While we explain the theoretical bases for presuming cause and effect relationships between budget credibility and selected explanatory variables, the linkages are complex and reverse causality might affect our results. Our analysis might also suffer from omitted variables, given the difficulties of identifying or measuring all determinants of budget credibility. Readers should keep these limitations in mind when reviewing the regression results.

4. EXPLORING THE DETERMINANTS OF EXPENDITURE DEVIATIONS

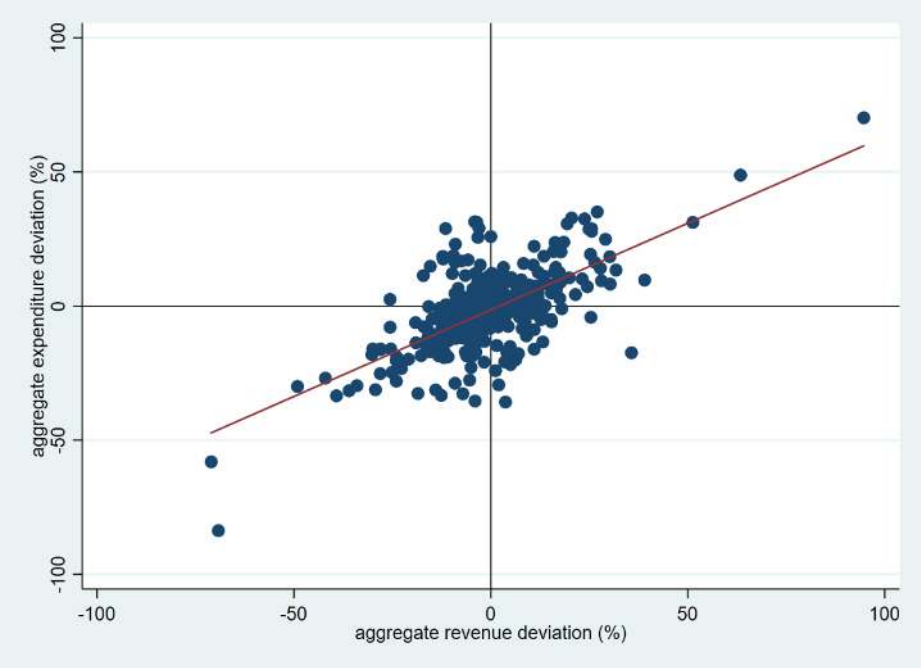
In this section, we look at some of the possible determinants of **aggregate spending deviations**. These are denoted as exp_dev and are measured as the percent difference between the total expenditure in the approved budget and the expenditure actually incurred in any given year.

We first test the simple hypothesis that deviations on the spending side are directly affected by deviations on the revenue side, building on the common justification provided in many PEFA assessments, as shown in Mills (2018). If governments overestimate the revenues that they expect to collect in any given year, they will be forced to cut expenditure – or borrow and increase public debt. If, on the other hand, they collect more than the revenue that was originally forecasted, they will have additional resources to spend, beyond the limits imposed by the originally

approved budget. We therefore expect aggregate revenue deviations, denoted as rev_dev, to have a significant positive impact on exp_dev.

This is confirmed by the data presented in Figure 1. Based on a simple bivariate regression, revenue outturns explain about 44% of the variance in expenditure outturns. Using country averages rather than individual year observations, the correlation is even stronger, with average revenue outturns explaining more than 47% of the variance in average expenditure outturns. The OLS results suggest that overall expenditure outturns are expected to increase by 0.6 percentage points for every percentage point increase in revenue outturns. However, as can be expected, model specification tests indicate omitted variables.

FIGURE 1. AGGREGATE REVENUE AND EXPENDITURE DEVIATIONS



We therefore proceed to include in our model other explanatory variables of interest, the main one being the *quality of PFM systems*. Weaknesses in PFM systems – from poor planning and forecasting to deficient cash management, procurement and reporting – can hamper budget execution, leading to unexpected deviations from approved budgets. We would therefore expect a positive relationship between the quality of PFM systems and the capacity of governments to stick to their plans. Governments with stronger PFM systems, in other words, are expected to have lower budget deviations, as also found in other studies.

We measure the quality of PFM systems using additional data from indicators in the PEFA assessments. Most PEFA indicators have more than one performance dimension and are assessed on an ordinal scale, from A (best) to D (worst). Following past practice, we convert the letter scores into numerical scores (from A = 4 to D = 1) and assume that they do not change over the years covered by each assessment. We then calculate an aggregate composite score for PFM performance, denoted as $PEFA\_comp$. The aggregate score excludes the indicators on budget reliability, which are used as dependent variables, and donor- or revenue-related indicators. We also use the average of PEFA ratings for the individual dimensions, assigning them with equal weight. This approach allows us to be more comprehensive and avoid any bias that may be introduced by PEFA's "weakest link" methodology, used for scoring multi-dimensional indicators.[7] We also calculate sub-scores for specific elements or pillars of the PFM system. Both the aggregate score and sub-scores can range, by design, from 1 (when a country gets D's on all dimensions) to 4 (when it gets all A's).[8]

The other two explanatory variables that we include later in our model are related to *budget transparency* and *budget oversight*. We consider budget transparency based on findings from Sarr (2015) using the principal-agent model to assess the potential impact of transparency on budget credibility. Greater transparency will reduce the informational advantage enjoyed by the agents – in this case government officials with decision-making powers over budget execution – and increase the visibility and public awareness of budget credibility challenges, incentivizing them to execute the budget as planned. We therefore expect a country's score on the Open Budget Index (or OBI, our measure of budget transparency) to have a positive impact on budget credibility.

Budget oversight refers to the role played by oversight actors such as the legislature and supreme audit institution (SAI) in the budget process in keeping executive action under scrutiny. On the one hand, well-functioning and independent oversight can have a similar effect to budget transparency, creating pressures for better budget execution. On the other hand, greater power vested in the legislature can lead to amendments that are not feasible and, in turn, poorer execution.[9] We assess these potential links using the Open Budget Survey (OBS) score on oversight.

[7] Our approach follows de Renzio et al. (2011), Fritz et al. (2014) and Mustapha (2019), although Fritz et al. uses indicators rather than dimensions. The choice between indicators and dimensions would not make a substantial difference – aggregate scores would change by less than 0.1 on average, and by 0.3 at most.

[8] Unlike the past studies noted above, which rely solely on assessments using the 2011 framework, our analysis includes more recent data from those using the 2016 framework. We reassign the scores from older reports to the corresponding indicators and dimensions, based on the tables available in the "Guidance on Tracking PFM Performance for Successive Assessments" provided by the PEFA Secretariat. When multiple indicators and dimensions are comparable to a single one, we take an averaging approach. We use this mapped data to calculate the scores (Annex Table 3 specifies the 2011 indicators used for each element). It is worth observing that some indicators are comparable in "subject only," and as such, the change in PEFA framework may cause some inconsistencies in our measure of the quality of PFM systems.

[9] Atiku and Lakin (2019) found that in the case of Nigeria, legislative amendments increased the budget in 6 of 8 years assessed and concluded that they tend to exacerbate underspending.

The expected impact of all of the above explanatory variables, alongside the control variables, on budget deviations does not theoretically depend on whether governments underspend or overspend, as in the case of revenue deviations, but relates to their absolute value, which they are supposed to help reduce. In the multivariate model, therefore, our dependent variable becomes $|exp\_dev|$. In addition to using all observations with available data, we look more specifically at the group that underspent their budgets separately. Our interest in this group stems from the fact that underspending is more likely to hamper public service delivery and that our previous research found that it is more prevalent around the globe.

Summary descriptive statistics for key variables are provided in Table 1 below. Governments in our sample under-execute their budgets by an average of 2% for expenditure and 1% for revenue, but with great variation across countries and over time.[10]

TABLE 1. SUMMARY STATISTICS FOR VARIABLES USED IN AGGREGATE EXPENDITURE CREDIBILITY ANALYSIS

Variable	No. Obs.	Mean	Std. Dev.	Minimum	Median	Maximum
Aggregate expenditure deviation (%)	362	-2.0	14.7	-83.7	-2.1	70.2
Aggregate revenue deviation (%)	362	-0.9	15.2	-71.0	-2.1	94.7
Composite PEFA score (1 to 4)	357	2.4	0.5	1.5	2.5	3.5
Open Budget Index (0 to 100)	197	39.3	18.2	0.0	42.9	91.2
Budget oversight score (0 to 100)	195	45.8	14.6	17.0	45.0	82.6
GNI per capita (1,000 PPP)	360	6.4	5.9	0.6	3.9	41.7
Polity2 score (-10 to 10)	326	4.1	5.1	-9.0	6.0	10.0
Government Effectiveness Index (percentile rank, 0 to 100)	360	33.9	19.8	1.0	31.6	81.7

Notes: No. Obs. = number of observations (with n = country year combination) and Std. Dev. = standard deviation. PEFA scores are applied to all assessed years. For more information including data sources, see Annex Table 1.

Our initial model, which looks at whether the two main explanatory factors – i.e., aggregate revenue deviations (here also expressed in absolute values) and quality of PFM systems as measured by $PEFA\_comp$ – are associated with overall budget credibility, is expressed as:

(1) $|Y_i| = \beta_0 + \beta_1|rev\_dev_i| + \beta_2 PEFA\_comp_i + \beta_x X_i + \varepsilon_i$

[10] Lack of budget credibility is a bigger challenge in low-income countries, which underspend their budgets by 5% and under-collect their revenue by 4% on average. These average figures differ substantially from those in our previous study, partly because of the larger sample and partly because PEFA reports using the 2011 framework did not assess donor-funded projects, which were a great driver of underspending.

where i signifies each country/year observation, Y is the dependent variable (exp_dev in this case), X includes our control variables, and ε represents the error term. We then add OBI and oversight scores as additional independent variables to estimate:

(2) $|Y_i| = \beta_0 + \beta_1|rev\_dev_i| + \beta_2 PEFA\_comp_i + \beta_3 OBI_i + \beta_4 oversight_i + \beta_x X_i + \varepsilon_i$

When we estimated equations (1) and (2) using OLS, however, we found that both models still suffered from model misspecification (see Annex Table 4 for OLS results). After examining ACPR plots and density curves, we log transformed the dependent variable, revenue credibility, and income per capita.[11] Equations (1) and (2) were thus modified to:

(3) $\ln|Y_i| = \beta_0 + \beta_1 \ln|rev\_dev_i| + \beta_2 PEFA\_comp_i + \beta_x X_i + \varepsilon_i$

(4) $\ln|Y_i| = \beta_0 + \beta_1 \ln|rev\_dev_i| + \beta_2 PEFA\_comp_i + \beta_3 OBI_i + \beta_4 oversight_i + \beta_x X_i + \varepsilon_i$

Table 2 provides the OLS results for equations (3) and (4). Both models display statistically significant, positive coefficients for revenue deviation, as could be expected from the previous bivariate analysis. OLS estimates suggest that for every 1% increase in revenue deviations, expenditure deviations are expected to increase by around 0.3%, holding all other variables constant. The quality of PFM systems is also significantly correlated with aggregate expenditure deviations, with better quality systems associated with lower deviations. A similar negative association can be seen for budget transparency and for government effectiveness, which could suggest that certain aspects of institutional performance – including the public availability of budget information and others less directly related to PFM – are also relevant for improving budget credibility. Other variables do not show any statistically significant coefficients. Interestingly, neither higher incomes nor stronger democratic institutions are associated with better budget execution when other variables are taken into account. For oversight, this could be linked to the potentially contradictory role of legislatures in incentivizing credibility or aggravating underspending.

Looking at the group that underspent their aggregate budget (i.e., $exp\_dev < 0$), the coefficient for revenue deviation is still positive and significant at 5%, but smaller – possibly reflecting governments' capacity to borrow to fill the financing gap left by lower than expected revenues. The coefficient for PFM quality remains negative although it loses statistical significance, while the one for OBI remains statistically significant and negative (p-value of 0.014). The coefficient for $oversight$ also becomes significant at 10% (p-value of 0.080). There are more

[11] We use log transformation partly to reduce the skew in these variables and thus reduce the variability, but primarily to address the issue of incorrect functional form and heteroskedasticity, as we rely on OLS. This is the most conventional approach and, at least in our analysis, seems to serve its purpose without creating any apparent problems (e.g., difficulty in interpretation, changed relationship between dependent and independent variables).

uncertainties around its direction, but the coefficient shows a positive sign, possibly suggesting the role of legislative amendments in exacerbating underspending. While we remain wary of test results for model specification, these estimates seem to suggest that factors related to fiscal accountability can be especially important in explaining and addressing underspending.

TABLE 2. NATURAL LOG OF AGGREGATE EXPENDITURE DEVIATIONS, OLS RESULTS

	(3) ln|exp_dev|	(4) ln|exp_dev|	
		all	underspend
ln(|rev_dev|)	0.332***	0.320***	0.187**
	(0.058)	(0.073)	(0.094)
PEFA_comp	-0.373**	-0.403*	-0.377
	(0.165)	(0.243)	(0.333)
OBI		-0.014**	-0.019**
		(0.006)	(0.007)
oversight		0.010	0.015*
		(0.007)	(0.008)
ln(gni_pc)	0.050	0.006	-0.020
	(0.092)	(0.133)	(0.154)
polity2	0.009	0.023	-0.011
	(0.013)	(0.021)	(0.028)
GEI	-0.009*	-0.014*	-0.011
	(0.005)	(0.007)	(0.008)
constant	2.295***	2.672***	2.955***
	(0.391)	(0.591)	(0.734)
no. obs.	*318*	*184*	*111*
R-squared	*0.170*	*0.252*	*0.317*
RESET p-value	*0.532*	*0.549*	*0.049*
White's p-value	*0.523*	*0.680*	*0.832*
mean VIF	*1.46*	*1.51*	*1.78*

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Standard errors in parentheses.

It is also interesting that $PEFA\_comp$ becomes less significant when OBI and $oversight$ are added as regressors. The p-value increases from 0.025 to 0.099, and the confidence interval changes from [-0.698, -0.047] to [-0.882, 0.077], as we move from equations (3) to (4), although the size of the coefficient increases. This might be because $PEFA\_comp$, to some extent, already captures a country's performance on budget transparency and oversight through some of its indicators. To address this issue and try and get to a more nuanced assessment of which aspects of PFM quality might matter more for budget credibility, we calculated sub-scores for each one of the pillars of the PEFA framework, again using individual indicator dimensions and excluding those related to tax administration. In the 2016 framework, they are called: (a) transparency of public finances ($PEFA\_tran$); (b) management of assets and liabilities ($PEFA\_mgmt$); (c) policy-based fiscal strategy and budgeting ($PEFA\_budg$);

(d) predictability and control in budget execution ($PEFA\_exec$); (e) accounting and reporting ($PEFA\_acct$); and (f) external scrutiny and audit ($PEFA\_audit$).[12]

Using variations of model (3), we checked to see if results changed when using the sub-scores for each PEFA pillar instead of $PEFA\_comp$. Estimates for each variable for the PEFA sub-scores are presented in Table 3; full results for all variables are available in Annex Table 5. As expected from the results for Open Budget Index or OBI, we find a negative relationship between $PEFA\_tran$ and aggregate expenditure deviations that is significant at 1% and has the largest coefficient of all pillars. Among the remaining pillars, dimensions in "management of assets and liabilities", "accounting and reporting" and "external scrutiny and audit" also show significant negative coefficients. This suggests that improving these aspects of PFM systems – for example by strengthening the selection, costing and monitoring of public investment projects, improving the coverage and timeliness of budget reports, and ensuring adequate follow up to audit findings – could be an important step in addressing lack of credibility on the expenditure side.

TABLE 3. NATURAL LOG OF AGGREGATE EXPENDITURE DEVIATIONS, OLS ESTIMATES FOR PEFA SUB-SCORES

	Coefficient	Std. Error	P-value	95% Conf. Interval	
				lower limit	upper limit
PEFA_tran* *fiscal transparency*	-0.315	0.109	0.004	-0.529	-0.102
PEFA_mgmt* *asset and liability management*	-0.245	0.112	0.030	-0.465	-0.024
PEFA_budg *policy-based budgeting*	-0.091	0.147	0.536	-0.381	0.198
PEFA_exec *control and predictability in budget execution*	-0.193	0.147	0.189	-0.481	0.096
PEFA_acct* *accounting and reporting*	-0.254	0.108	0.019	-0.466	-0.042
PEFA_audit* *external audit*	-0.184	0.106	0.083	-0.392	0.024

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively.

[12] For more details about the indicators covered under each pillar, see Annex Table 3.

5. EXPLORING THE DETERMINANTS OF REVENUE DEVIATIONS

Based on the evidence of a strong positive relationship between revenue and expenditure credibility, it is imperative to also explore what causes collected revenues to deviate from their original forecasts and projections. For this purpose, we set up some additional analyses using rev_dev as our dependent variable.

In addition to the usual controls for income levels, democracy and government effectiveness, which can be assumed to maintain their relevance in explaining revenue deviations, we assess the relationship between revenue deviations and the following explanatory variables:

a) the quality of the *revenue planning* process, based on the hypothesis that more accurate revenue projections depend on a government's capacity to run the necessary sophisticated macroeconomic models and to limit political meddling with the revenue forecasting process.[13] There are no indicators, however, that cover this aspect of the budgeting process directly. As a consequence, we follow the approach adopted in Addison (2013) by calculating the difference between "expected revenue" and forecasted revenue, denoted as rev_plan, as a proxy for inaccurate macroeconomic forecasting or for overoptimistic budgeting. We define expected revenue as the actual collected revenue from the previous year adjusted by the annual growth in gross domestic product (GDP) and by inflation. This corresponds to what revenue projections should look like without any substantial policy change. Large, positive differences between the expected and forecasted value could indicate either lack of technical capacity in revenue forecasting or deliberate over-estimation. For this reason, we expect to find a negative relationship between rev_plan and rev_dev;[14]

b) unexpected *macroeconomic shocks*, to assess their potential impact on revenue collections, and therefore on deviations from the original forecasts. When countries face an unexpected drop (or rise) in their growth rates during a fiscal year, this is likely to result in an over (or under) estimation of their revenues. We measure this effect through the difference in GDP growth rates between the current and the previous year, and expect to find a positive relationship between gdp_shock and rev_dev;

[13] For a recent review of revenue forecasting in developing countries, and a discussion of governments' tendency to adopt optimistic projections, see Cangiano and Pathak (2019). A recent IMF blog post also reviews some evidence on the politics of revenue forecasting (see J. Grinyer, 'The Politics of Revenue Forecasting', 7 February 2019).

[14] Note that we are not able to calculate this variable for the first year covered in each PEFA assessment, for the reports that do not provide data on the amount of revenue projected and collected (or, in other words, only provide the percent deviation figures), and for the countries with no inflation data. This affects our sample size.

In addition, in the second model, we include measures of *aid dependency* and *natural resource dependency* for the following reasons:

a) Many governments in our sample receive substantial external financial assistance, and this can have a contradictory impact. On the one hand, foreign aid could provide needed support to building domestic institutions and improve governance. In this case, we could expect it to have a positive impact on revenue credibility. On the other hand, governments who rely heavily on external assistance to finance spending may have less of an incentive to prioritize efforts at mobilizing domestic revenues.[15] Foreign aid is also known to be a volatile source of revenue, which might imply that the relationship between aid dependency and revenue credibility is a negative one. We try to assess these conflicting forces by including the 5-year average of net official development assistance (ODA) as a share of GNI, denoted as aid_5yr, as an independent variable in our model.

b) Natural resource rents also tend to be volatile and are easily affected by factors that are outside of government control, such as changes in international demand or price. Previous studies also found a negative impact of resource dependency on economic performance and governance, which can in turn affect budget execution and revenue collection.[16] Governments that rely heavily on resource rents for their revenue could then be expected to suffer from a lack of revenue credibility. There are different measures for resource dependency that we considered – some of them are continuous, using total natural resource rents or oil rents as a share of GDP, while others are dummy variables that use specific criteria for defining resource-dependent countries.[17]

Table 4 provides descriptive statistics for the four additional variables considered for revenue deviations (for resource dependency, we present summary statistics for the two continuous variables). It shows that countries in our sample tend to over-estimate their revenues relative to the expected amount and, on average, did not experience any significant macroeconomic shock. It also suggests heavy right skews in the data for aid and resource dependency.

[15] For a review of arguments looking at the relationship between aid dependency and revenue collection in Africa, see Moss et al. (2006).

[16] The literature on the "resource curse" is vast. For two overview papers from one of the major experts in the field, see Ross (1999) and Ross (2015).

[17] See Mustapha (2019), which relies on the definition used by IMF, or Bleaney and Halland (2016).

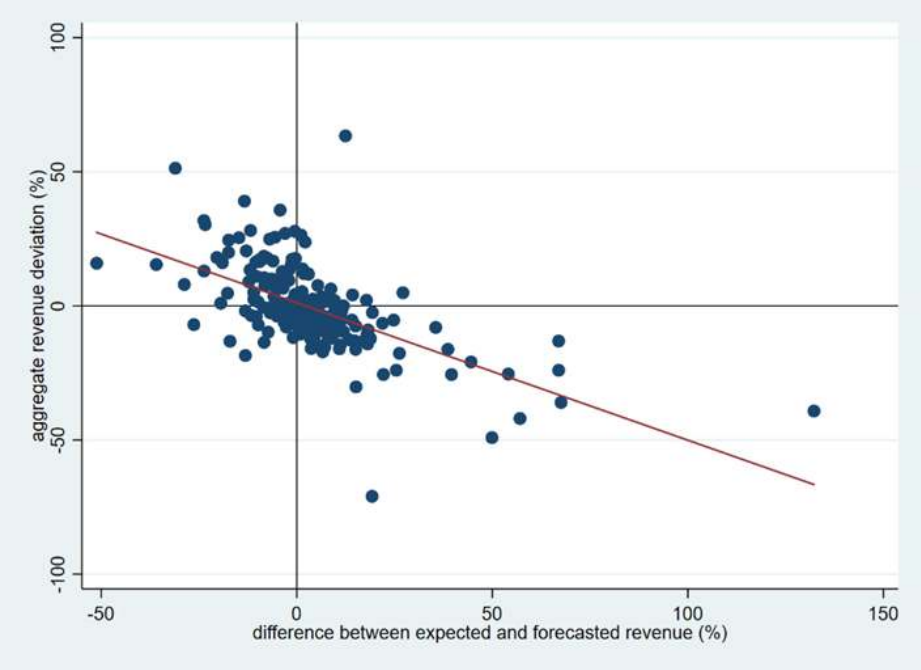
TABLE 4. SUMMARY STATISTICS FOR ADDITIONAL VARIABLES USED IN AGGREGATE REVENUE CREDIBILITY ANALYSIS

Variable	No. Obs.	Mean	Std. Dev.	Minimum	Median	Maximum
Revenue planning (%)	215	3.7	18.0	-51.2	2.1	132.2
Annual change in GDP growth (p.p.)	361	0.1	4.9	-25.2	0.0	29.7
Net ODA as a share of GNI (average over 5 years, %)	360	8.2	10.8	-0.1	4.8	67.8
Total natural resource rents as a share of GDP (%)	360	9.8	11.7	0.0	5.5	59.3
Oil rents as a share of GDP (%)	353	2.8	8.6	0.0	0.0	55.2

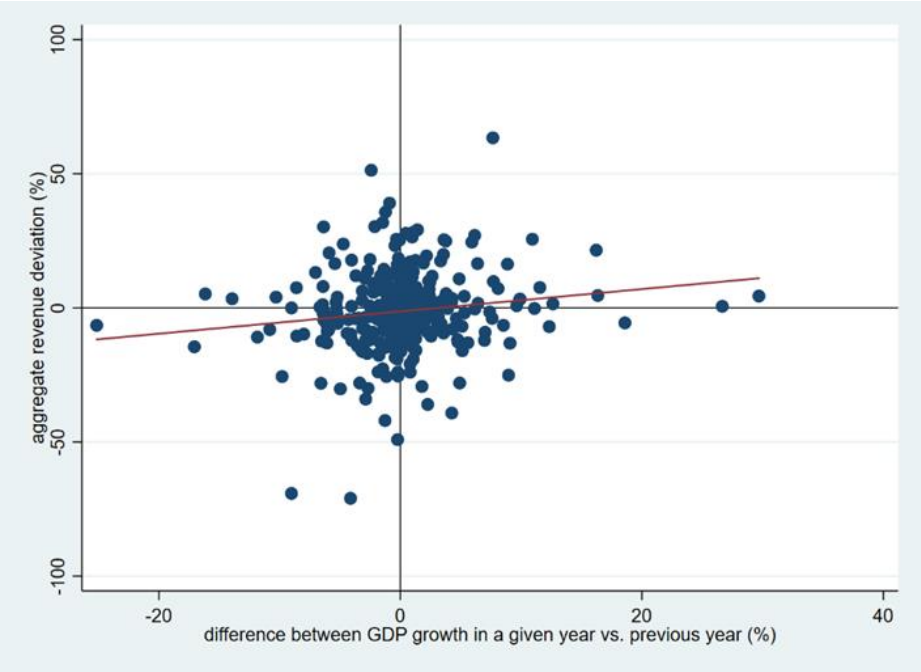
As with the aggregate expenditure credibility analysis, many regressors are expected to have a non-linear relationship with the raw value of rev_dev. The exceptions are rev_plan and gdp_shock, which were also provided as key explanations of lack of credibility in a large number of PEFA reports, according to Mills (2018). Figures 2A and 2B illustrate the bivariate relationships between these two variables and aggregate revenue deviation. As expected, we find a strongly linear and negative association for rev_plan (coefficient of -0.512, p-value of 0.000 and R-square of 0.369). For gdp_shock, we find a positive association, but with more uncertainties around its impact (coefficient of 0.416, p-value of 0.007 and R-square of 0.020).

FIGURE 2. REVENUE PLANNING, CHANGE IN GDP GROWTH AND AGGREGATE REVENUE DEVIATIONS

2A. REVENUE PLANNING AND DEVIATIONS



2B. CHANGE IN GDP GROWTH AND REVENUE DEVIATIONS



Using just these two variables as regressors, we find that one percentage point increase in the difference between expected and forecasted revenue would result in a 0.5 percentage point decrease in revenue deviations, holding the change in GDP growth constant.[18] In simpler terms, the more ambitious a government's revenue projections, the more likely that it will see a revenue shortfall. Once again, however, diagnostic tests indicate model misspecification. Thus, taking an approach similar to the one adopted with analysis of aggregate spending deviations, we then proceed to log transform some variables and estimate the following equations:

(5) $\ln|Y_i| = \beta_0 + \beta_1 \ln|rev\_plan_i| + \beta_2|gdp\_shock_i| + \beta_x X_i + \varepsilon_i$

(6) $\ln|Y_i| = \beta_0 + \beta_1 \ln|rev\_plan_i| + \beta_2|gdp\_shock_i| + \beta_3 \ln(aid\_5yr_i) + \beta_4 resource_i + \beta_x X_i + \varepsilon_i$

where Y now denotes rev_dev, X includes the usual control variables and $resource$ is a dummy variable for dependency on natural resources.[19] In addition to rev_plan, macroeconomic shocks are included using their absolute value. This is because we expect that governments will be able to plan their revenues better when the unexpected change in GDP growth is smaller, regardless of its direction.

[18] p-value of 0.000, confidence interval of -0.424 and -0.610.

[19] As OLS results using the different measures of resource dependency were essentially identical, we selected the measure that produced the largest R-squared value.

Table 5 provides the OLS results for these models, where rev_plan stands out as the only statistically significant factor at 5%. This does not mean that other regressors have no meaningful association with revenue deviations – confounding factors and spurious relationships are likely to exist in our model. But the OLS results suggest that poor planning may be the most significant factor. Again, we also look specifically at the cases in our sample when revenues were under-collected (i.e., $rev\_dev < 0$), and find similar results, with the only change being that resource dependency is weakly associated with larger revenue shortfalls, suggesting that the presence of large resource revenues might exacerbate deviations from projected revenues.

TABLE 5. NATURAL LOG OF AGGREGATE REVENUE DEVIATIONS, OLS RESULTS

	(5) ln|rev_dev|	(6) ln|rev_dev|	
		all	shortfall
ln(|rev_plan|)	0.367***	0.351***	0.336***
	(0.065)	(0.066)	(0.072)
|gdp_shock|	-0.025	-0.027	-0.009
	(0.021)	(0.021)	(0.024)
ln(aid_5yr)		-0.031	0.121
		(0.091)	(0.095)
resource		0.239	0.344*
		(0.176)	(0.197)
ln(gni_pc)	0.028	-0.001	0.130
	(0.107)	(0.158)	(0.168)
polity2	-0.010	-0.010	0.024
	(0.017)	(0.017)	(0.023)
GEI	-0.005	-0.004	-0.011
	(0.005)	(0.005)	(0.007)
constant	1.424***	1.421***	1.030***
	(0.229)	(0.356)	(0.386)
no. obs.	*196*	*191*	*109*
R-squared	*0.163*	*0.171*	*0.336*
RESET p-value	*0.108*	*0.259*	*0.246*
White's p-value	*0.144*	*0.091*	*0.123*
mean VIF	*1.30*	*1.74*	*1.80*

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Standard errors in parentheses.

What might explain poor planning on the revenue side? Is it the inability to carry out accurate forecasting or political motives? While more – and more detailed – research would be necessary to answer these questions,[20] countries that received higher PEFA scores on dimensions in "policy-based fiscal strategy and budgeting" tend to forecast their revenue closer to the expected amount.

[20] See Herrera and Lakin (2019) for evidence of governments artificially inflating revenue projections from case study evidence in IBP's budget credibility project.

TABLE 6. SUMMARY STATISTICS FOR REVENUE PLANNING BY PEFA PILLAR SCORE FOR POLICY-BASED BUDGETING

Group (by increasing PEFA score)	**No. Obs.**	**Mean**	**Std. Dev.**	**Minimum**	**Median**	**Maximum**
D	2	14.2	3.2	11.9	14.2	16.5
D+ to C	10	14.1	10.0	4.4	11.5	36.5
C to C+	33	14.6	19.1	1.7	8.2	92.1
C+ to B	32	9.9	8.5	0.7	7.8	42.0
B to B+	23	7.9	5.4	1.7	6.6	21.9
A	5	3.2	1.6	1.1	2.8	5.3

Notes: n = country. Shows multi-year average of absolute values of rev_plan. When we look at raw values, the mean value for each group would be -3.1% (with a standard deviation of 18.9%), 8.4% (15.5%), 6.2% (22.8%), 2.8% (10.4%), 2.7% (9.0%) and 0.3% (3.8%). This suggests that over-estimation is a bigger problem in our sample, especially in countries that do not meet performance standards in this PEFA pillar.

In fact, when we replace rev_plan with $PEFA\_budg$ in equations (5) and (6), the results are similar – except that resource dependency gains stronger statistical significance in model (6). The results using $PEFA\_budg$ are provided in Annex Table 6 and indicate that some of the technical aspects of revenue forecasting captured by PEFA assessments – like the preparation of macroeconomic and fiscal forecasts and the use of sensitivity analysis – are associated with improved revenue credibility.

6. EXPLORING THE DETERMINANTS OF DEVIATIONS IN COMPOSITION OF SPENDING

In the final part of our analysis we turn to "compositional" credibility, which looks at changes in the composition of spending during budget execution. For this purpose, we use "variance in expenditure composition" in PEFA's indicator PI-2, which measures the extent to which budget reallocations during execution result in variance in expenditure composition, as our dependent variable.

Annexes of PEFA reports generally provide data on the approved and executed budgets of the 20 largest administrative or programmatic heads (for PI-2.1); annexes of reports using the 2016 framework also include data by economic classification (for PI-2.2). Using these disaggregated data retrieved from 103 reports and replicating

PEFA's current methodology, we calculate composition variance, $PI2i\_var$ and $PI2ii\_var$.[21] Table 7 summarizes data on expenditure composition for our sample. Average variance was 17.4% for PI-2.1 and 18.5% for PI-2.2, both of which would receive a D in PEFA. It is also worth noting that low-income countries tend to suffer more from lack of compositional credibility; within this income group, the average variance was 23% for PI-2.1 and 25% for PI-2.2.

Looking at the percent deviation in each category of the budget, we can also see that on average, capital projects and economic sectors (e.g., Ministry of Agriculture, housing and community amenities) are underspent the most, with large variations.

TABLE 7. SUMMARY STATISTICS FOR EXPENDITURE COMPOSITIONAL CREDIBILITY

Variable	No. Obs.	Mean	Std. Dev.	Minimum	Median	Maximum
PI2i_var: Variance in expenditure composition by administrative or programmatic head (%)	306	17.4	12.3	1.7	14.2	74.9
Deviation in expenditure on general government functions (%)	*306*	*0.6*	*19.5*	*-87.5*	*-1.0*	*102.4*
Deviation in expenditure on economic or infrastructure development (%)	*306*	*-5.5*	*31.9*	*-89.5*	*-7.4*	*216.7*
Deviation in expenditure on social sectors, e.g., health, education (%)	*306*	*-4.8*	*14.6*	*-89.7*	*-3.0*	*51.2*
PI2ii_var: Variance in expenditure composition by economic classification (%)	104	18.5	14.3	2.8	14.7	87.7
Deviation in recurrent expenditure (%)	*104*	*-3.6*	*15.2*	*-83.7*	*-1.8*	*46.7*
Deviation in capital expenditure (%)	*101**	*-14.9*	*31.3*	*-97.4*	*-15.4*	*88.0*

*The number of observations for the deviation in capital spending is smaller because the 2018 PEFA report for Afghanistan shows zeroes for "consumption of fixed capital" in all 3 years that were assessed, which are treated as missing values.

Factors we considered to explain deviations in expenditure composition are similar to those we considered for overall spending credibility, but the analysis differs in two ways:

a) Here we include aggregate expenditure deviation (in absolute terms) as a regressor, because we expect that better overall credibility can help minimize compositional deviations. Deviations at the aggregate level would require redistribution of resources, and the impact is likely to be more unequal with greater under or overspending if some budget categories tend to be more protected than others. Conversely, deviations in certain categories can also influence overall credibility. For instance, capital budgets could be underspent due

[21] Compositional variance is calculated as the total sum of "absolute deviation" as a share of the total sum of "adjusted budget" across budget heads. "Adjusted budget" is calculated by applying the percent deviation in aggregate expenditure equally across the approved budget for each budget head – for example, if education was allocated $100 and the overall budget was underspent by 5%, the adjusted budget for education would be $95 (95% of $100). "Absolute deviation" is calculated as the difference between the adjusted budget and actual spending – building on the previous example, if actual spending for education was $90, its absolute deviation would be $5 (the difference between adjusted budget of $95 and actual spending of $90). For the calculation sheet and more information about PEFA's methodology, see: http://pefa.org/resources/calculation-sheets-pefa-performance-indicators-pi-1-pi-2-and-pi-23-november-2018.

to procurement delays and this could be reflected at the aggregate level. We control for some PFM challenges that can lead to such situations.

b) We look at specific components of PFM systems using PEFA data, through both sub-scores and selected indicators or dimensions. More specifically, we look at indicators on the rules governing in-year budget adjustments by the executive – testing the hypothesis that when the executive has more flexibility to shift allocations between different budget heads the compositional credibility for functional or administrative classification will suffer[22] – and on the quality of procurement systems, which are more likely to directly affect capital budget execution, and therefore compositional credibility for economic classification.

Our initial model looks at which pillars of PFM systems, as defined by PEFA, help explain the variance in expenditure composition by administrative or programmatic unit. It is expressed as:

(7) $Y_i = \beta_0 + \beta_1 PEFA_i + \beta_2 |exp\_dev_i| + \beta_x X_i + \varepsilon_i$

where Y is now the compositional credibility, measured using $PI2i\_var$, and $PEFA$ may be any sub-score used in the aggregate expenditure analysis.

OLS results looking at variance by administrative classification and using each PEFA sub-score are provided in Table 8. To account for heteroskedasticity in the error term in the variations of model (7), we present robust standard errors.[23]

In all cases, as expected, $|exp\_dev|$ shows a statistically significant and positive relationship with $PI2i\_var$. Among different aspects of PFM systems, "transparency of public finances" and "predictability and control in budget execution" are the ones that show strong negative associations with compositional variance by administrative or functional heads. This again highlights the importance of fiscal transparency in keeping governments accountable for delivering on the promised allocation of resources, and hints that weak budget execution systems can exacerbate deviations in expenditure composition, even more so than having an impact on overall budget credibility.

[22] See Herrera and Lakin (2019) for some case study evidence in this respect.

[23] The presence of heteroskedasticity is not very surprising. High overall credibility, for example, may be a "necessary but not sufficient condition" for high compositional credibility. Also, as previously discussed, there are likely interaction effects that our model does not consider. As our model does not seem to suffer from violation of any other assumptions (e.g., model specification seems to be correct), we deal with this issue by relaxing the assumption that the error terms are drawn from identical distributions.

The coefficients for our control variables have the right sign in almost all cases, but only the indicator of government effectiveness reaches the minimum level of statistical significance in some cases. The confidence interval, however – with a positive upper limit – suggests uncertainties in the estimators.

TABLE 8. EXPENDITURE COMPOSITIONAL VARIANCE BY ADMINISTRATIVE CLASSIFICATION, OLS RESULTS USING PEFA SUB-SCORES

	(7) PI2i_var					
	PEFA_tran	PEFA_mgmt	PEFA_budg	PEFA_exec	PEFA_acct	PEFA_audit
PEFA score	-3.939***	-0.442	-1.157	-4.853***	-0.404	-1.505
	(0.820)	(1.123)	(0.985)	(1.404)	(1.094)	(1.045)
|exp_dev|	0.568***	0.624***	0.623***	0.598***	0.623***	0.617***
	(0.065)	(0.065)	(0.066)	(0.061)	(0.066)	(0.063)
ln(gni_pc)	-0.463	-0.999	-1.108	-0.844	-0.971	-0.938
	(1.154)	(1.181)	(1.151)	(1.108)	(1.221)	(1.170)
polity2	0.023	-0.039	-0.031	-0.069	-0.040	-0.062
	(0.112)	(0.125)	(0.123)	(0.115)	(0.126)	(0.131)
GEI	-0.037	-0.066*	-0.056	-0.004	-0.068*	-0.062
	(0.038)	(0.037)	(0.041)	(0.041)	(0.038)	(0.039)
constant	23.246***	15.691***	17.512***	24.537***	15.714***	16.988***
	(2.310)	(2.522)	(2.923)	(2.718)	(2.635)	(2.398)
no. obs.	270	270	270	270	270	270
R-squared	0.416	0.382	0.384	0.411	0.382	0.385
RESET p-value	0.578	0.815	0.848	0.658	0.836	0.771
White's p-value	0.004	0.000	0.001	0.005	0.001	0.000
mean VIF	1.42	1.40	1.41	1.47	1.38	1.36

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Robust standard errors in parentheses.

When looking at the variance by economic classification, or $PI2ii\_var$, we find that the coefficient for PEFA's transparency sub-score becomes statistically insignificant, while the budget execution sub-score continues to have a significant, negative relationship even with economic compositional credibility. This again highlights the importance of efficient budget execution systems for different types of spending.

Other interesting changes include the coefficients for both $PEFA\_mgmt$ (sub-score for "management of assets and liabilities", which includes the indicator on "public investment management") and $PEFA\_audit$ which both become significant at 1%, pointing to other areas of the budget system which may help limit compositional deviations.[24] In addition to $|exp\_dev|$, the coefficient for GEI, our control variable on government effectiveness, is

[24] Small RESET p-value of 0.02 indicates specification issues with this variation of model (6) – in fact, most show somewhat uncertain results for the test, which may mean that our model is probably missing some key explanatory factors for economic compositional credibility.

negative and significant, again highlighting the likely importance of the quality of government systems beyond PFM. Complete OLS regression results for $PI2ii\_var$ are presented in Annex Table 7.

Finally, to assess our hypotheses about more specific aspects of the budget system, we estimate:

(8) $Y_i = \beta_0 + \beta_1|exp\_dev_i| + \beta_2 PI18iv_i + \beta_3 PI24\_comp_i + \beta_x X_i + \varepsilon_i$

where Y can be either $PI2i\_var$ or $PI2ii\_var$, $PI18iv$ is the PEFA score on the more specific dimension "rules for budget adjustments by the executive" and $PI24\_comp$ is the average PEFA score across the four dimensions that make up the indicator related to the quality of procurement systems.[25]

Table 9 displays the OLS results, which are aligned with our expectations. When rules that set limits on executive virements are present, clear, and adhered to, administrative or functional compositional variance is smaller – moving up a grade on the PEFA dimension is expected to reduce $PI2i\_var$ by 2.9 percentage points.[26] A well-functioning procurement system, signaled by accurate and complete data available for monitoring, competitive procedures, public access to key information, and complaints management, is a significant factor associated with compositional credibility by economic classification – moving up a grade on this indicator would reduce $PI2ii\_var$ by 2.5 percentage points. This reflects the importance of various aspects of the budget systems for different types of operations that could lead to compositional deviations. Interestingly, but somewhat unclearly, other aspects of government effectiveness also have a strong negative association with deviations in composition by economic categories.

[25] For assessments using the 2011 framework, we use PI-27iv – which is directly comparable to PI-18.4 – and PI-19 "Competition, value for money and controls in procurement", which is relevant to PI-24.

[26] We also considered Question 115 from the Open Budget Survey as an alternative variable here. It asks if the executive seeks approval "prior to shifting funds between administrative units that receive explicit funding in the Enacted Budget." When used instead of PI-18.4, OLS results remain virtually the same. A dummy variable for Q115 (0=does not obtain legislative approval in practice, 1=obtains legislative approval in practice) has a statistically significant coefficient in model (7) using PI2i_variance as the dependent variable but is not statistically significant for the model using PI2ii_variance.

TABLE 9. EXPENDITURE COMPOSITIONAL VARIANCE BY ADMINISTRATIVE CLASSIFICATION AND BY ECONOMIC CLASSIFICATION, OLS RESULTS USING SPECIFIC PEFA INDICATORS

	(8) PI2i_var	(8) PI2ii_var
|exp_dev|	0.583***	0.670***
	(0.066)	(0.109)
PI18iv	-2.891***	-2.637
	(0.855)	(1.602)
PI24	-0.628	-2.493**
	(0.694)	(1.096)
ln(gni_pc)	-1.035	2.193
	(1.084)	(1.359)
polity2	-0.002	-0.365*
	(0.116)	(0.214)
GEI	-0.009	-0.185***
	(0.040)	(0.055)
constant	22.634***	28.394***
	(3.066)	(6.763)
no. obs.	*261*	*95*
R-squared	*0.419*	*0.711*
RESET p-value	*0.750*	*0.103*
White's p-value	*0.053*	*0.181*
mean VIF	*1.42*	*1.64*

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Robust standard errors in parentheses.

7. CONCLUSIONS AND RECOMMENDATIONS

This paper explores possible factors behind the lack of budget credibility across countries and builds on the descriptive analysis carried out in another recent publication (de Renzio et al. 2019). Our earlier work found large discrepancies between approved and executed budgets across countries including substantive underspending by governments in areas critical to sustainable development, especially in lower-income countries.

This analysis, which uses data from PEFA assessments, investigates the possible determinants of three aspects of budget credibility: (a) aggregate expenditure deviations; (b) revenue deviations; and (c) deviations in the composition of spending, all of which have corresponding PEFA indicators. We use a standard set of control variables and adapt the choice of explanatory variables to each aspect of credibility, drawing on previous literature, descriptive data patterns and some of the issues raised in the PEFA assessments, themselves, when discussing the reasons for lack of budget credibility.

The results of our analysis *per se* are not very surprising, as they broadly confirm several previous findings and other propositions and intuitions based on "common sense" and practical experience. However, that they do so based on a thorough analysis of a large sample of countries and years, rather than on single cases or anecdotal evidence, is significant.

SUMMARY OF FINDINGS

1) Aggregate expenditure deviations are most strongly associated with revenue credibility, but also with the quality of PFM systems, including the presence of more transparent budgets and better accounting and reporting systems which allow for more direct real-time control of how public resources are being used.

2) Revenue deviations, in turn, are mostly associated with governments' capacities in revenue planning. These seem related to technical aspects of revenue forecasting, but could also have a more political explanation, given the incentive that many governments face to adopt overly optimistic revenue projections.

3) When it comes to deviations in expenditure composition, again there are certain aspects of the PFM system that are associated with better credibility, or governments sticking to their resource allocation plans. These include levels of budget transparency and the quality of budget execution systems. More specifically, clearer and better adhered rules for in-year amendments by the executive are associated with better compositional credibility by functional or administrative classification, and better quality of procurement systems with better credibility by economic classification.

4) In most cases, a country's level of income and the strength of its democratic institutions do not show up as significantly correlated with better budget credibility. It is other factors that seem to drive better budget credibility, and these could be more amenable to reform and improvement than a country's economic performance or its political institutions. On the other hand, aspects related to broader government effectiveness – above and beyond what happens with budget systems – show up several times as being associated with better budget credibility.

RECOMMENDED ACTIONS

Potential action points that governments and other actors can take up to improve budget credibility at different levels include:

1) Strengthening the revenue planning and forecasting function, in order to limit aggregate revenue deviations. This would require additional, country-specific analysis to identify specific gaps in technical capacities, key institutional bottlenecks and the existence of political incentives to inflate revenue projections, as well as a process to design reform interventions to adequately address each of the above obstacles.

2) Strengthening components of PFM systems that are associated with improved budget credibility. This includes several specific areas, from accounting and reporting systems to managing public investment, and from various aspects of budget planning and execution (including virement rules and procurement systems) to external audit. Again, specific priority interventions for each country may need to be identified through additional analysis and dialogue.

3) Further improving fiscal transparency practices, in general, seems to be beneficial for different aspects of budget credibility, and could therefore be an additional focus of PFM reforms across countries. This could include specific, credibility-related transparency reforms such as more disaggregated reporting on budget execution vs. original allocations across various classifications and providing adequate explanations and justifications for deviations.

4) Not overlooking the importance of more general and complementary government systems beyond PFM, that may have bearing on governments' capacity and willingness to address policy formulation and implementation challenges, therefore improving overall policy credibility.

We have tried our best to address statistical shortcomings and to be transparent about our assumptions and approaches in this paper, but we acknowledge the inevitable limitations of this work. Cross-country regressions like the ones used here rely on the availability of large amounts of data on different variables, many of which are not regularly produced for a large enough number of countries, or relate to concepts that do not lend themselves easily to operationalization and quantification. Consequently, it is often necessary to rely on inadequate proxy variables. For example, there is no specific and available indicator of governments' capacities for revenue planning, or for assessing the influence of political factors on inflated revenue projections. Similarly, our indicator of government effectiveness is based on a composite measure which has been questioned in terms of its reliability. It is also possible that budget credibility has a positive impact on the perception of government effectiveness, as well as some other explanatory variables we looked at, rather than vice versa. The OLS results show statistical correlation rather than causality, and may not fully explain the intricate relationships between budget credibility and its explanatory factors.

Potential issues of omitted variables, measurement or identification errors and reverse causality can be identified as the key limitations of our analysis. Our results, therefore, should be interpreted as preliminary and exploratory. Still, our findings offer a better understanding of budget credibility using recent data and can help guide further research and action. IBP intends to continue deepening and improving on the analysis begun in this paper. We will also incorporate some of the findings and lessons into the next phase of our work on budget credibility, which will have a much stronger country focus. In the end, it is at the country level where factors affecting budget credibility can be properly identified and addressed through tailored approaches and appropriate reform initiatives.

REFERENCES

Ablo, Emmanuel and Ritva Reinikka. (1999). *Do Budgets Really Matter? Evidence from Public Spending on Education and Health in Uganda*. Policy Research Working Paper No. 1926. Washington, DC: World Bank.

Addison, Douglas. (2013). *The Quality of Budget Execution and Its Correlates*. Policy Research Working Paper No. 6657. Washington, DC: World Bank.

Atiku, Samuel. and Jason Lakin (2019). *That's Incredible! The Contours of Budget Credibility in Nigeria*. Washington, DC: International Budget Partnership.

Bleaney, Michael Francis and Havard Halland (2016). *Do resource-rich countries suffer from a lack of fiscal discipline?* Policy Research Working Paper No. 7552. Washington, DC: World Bank.

Cangiano, Marco and Rahul Pathak (2019). 'Revenue Forecasting in Low-Income and Developing Countries: Biases and Potential Remedies.' In D. Williams, T. Calabrese (eds) *The Palgrave Handbook of Government Budget Forecasting*. Springer International Publishing.

de Renzio, Paolo, Chloe Cho and Jason Lakin (2019). *Budget Credibility Across Countries: How Deviations are Affecting Spending on Social Priorities*. Washington, DC: International Budget Partnership.

de Renzio, Paolo, Matt Andrews and Zac Mills (2011). *Does donor support to public financial management reforms in developing countries work? An analytical study of quantitative cross-country evidence*. ODI Working Paper 329. London: Overseas Development Institute.

Fritz, Verena, Stephanie Sweet and Marjin Verhoeven. (2014). *Strengthening Public Financial Management: Exploring Drivers and Effects.* Policy Research Working Paper No. 7084. Washington, DC: World Bank.

Herrera, Guillermo and Jason Lakin (2019). *Ultra-Vired: How Budget Rules and Practices Can Reduce Budget Credibility.* Washington, DC: International Budget Partnership.

Mills, Zac. (2018). *Budget Credibility: What Can We Learn from PEFA Reports?* Washington, DC: International Budget Partnership.

Moss, Todd, Gunilla Pettersson and Nicolas van de Walle. (2006). *An Aid-Institutions Paradox? A Review Essay on Aid Dependency and State Building in Sub-Saharan Africa.* Working Paper No. 74. Washington, DC: Center for Global Development.

Mustapha, Shakira. (2019). 'Budget Credibility, Fiscal Outcomes, and PFM Performance in Fragile and Nonfragile Countries.' Chapter 4 in *PEFA, Public Financial Management, and Good Governance.* Working Paper No. 135051. Washington, DC: World Bank.

Ross, Michael L. (2015). 'What Have We Learned about the Resource Curse?' *Annual Review of Political Science*, 18(1): 239-259.

Ross, Michael L. (1999). 'The Political Economy of the Resource Curse.' *World Politics*, 51(2): 297-322.

Sarr, Babacar. (2015). *Credibility and Reliability of Government Budgets: Does Fiscal Transparency Matter?* Washington, DC: International Budget Partnership.

Simson, Rebecca. and Bryn Welham. (2014). *Incredible Budgets: Budget credibility in theory and practice*. Working Paper 400. London: Overseas Development Institute.

Stasavage, David and Dambisa Moyo. (2000). 'Are Cash Budgets a Cure for Excess Fiscal Deficits (and at What Cost)?' *World Development*, 28(12): 2105-2122.

Williams, Martin. (2015). *One Size Does Not Fit All: Budget Institutions and Performance in Ghana's Spending Ministries*. Unpublished manuscript.

ANNEX TABLE 1. DEFINITIONS, EXPECTATIONS AND SOURCES FOR KEY VARIABLES

Variable	Definition	Expectation	Source
exp_dev	% difference between budgeted and actual expenditure at the aggregate level	dependent variable, but compositional variance expected to increase as *exp_dev* increases in absolute term	calculations based on PEFA data
rev_dev	% difference between budgeted and actual revenue at the aggregate level	dependent variable, but expected to spill over to expenditure and have a positive impact on *exp_dev*	
PI2i_var	% variance in expenditure composition by program, administrative or functional classification	dependent variable	
PI2ii_var	% variance in expenditure composition by economic classification (available only in 2016 framework)		
rev_plan	% difference between the "expected" revenue (previous year's actual revenue adjusted by growth in GDP and inflation) and budgeted revenue	expected to have a negative impact on *rev_dev* - the more "ambitious" the forecast, the more likely to under-collect	calculations based on PEFA data and World Bank Open Data
gdp_shock	percentage point difference between previous and this year's % change in GDP	used to calculate the expected revenue, but also expected to also have a direct positive impact on *rev_dev* if the unforeseen growth leads to an increase in the revenue collected	World Bank Open Data
gni_pc	GNI per capita based on purchasing power parity	expected to have a positive impact on budget credibility (i.e., reduce absolute deviations and compositional variance), given positive implications regarding better talent and system	
aid_5yr	5-year average of net ODA received as % of GNI	expected to have a negative impact on the absolute value of *rev_dev*, given negative implications regarding volatility; expected to have a positive impact if it improves PFM capacity	
resource / oil	total natural resource or oil rents as % GDP (for the analysis: a dummy for natural resource rents > 10%)	expected to have a negative impact on the absolute value of *rev_dev*, given negative implications regarding volatility	
GEI	percentile rank for perceptions of government effectiveness, 0=lowest to 100=highest	expected to have a positive impact on budget credibility, as effective civil and public services can increase the quality of formulation and implementation of the budget	Worldwide Governance Indicators
polity2	polity2 scores, -10=hereditary monarchy to 10=consolidated democracy	expected to have a positive impact on credibility, as accountability can motivate better budget execution	Polity IV database
OBI	Open Budget Index, 0=least to 100=most transparent; missing values interpolated	expected to have a positive impact on credibility (especially *exp_dev*), as transparency can reduce informational advantage and induce policymakers to execute the budget as planned	Open Budget Survey
oversight	role and effectiveness of oversight institutions, 0=weak to 100=strong; missing values interpolated	expected to have a positive impact on credibility (especially *exp_dev*) if it leads to better execution; expected to have a negative impact if it leads to unrealistic legislative amendments	

ANNEX TABLE 2. CORRELATION MATRIX FOR KEY VARIABLES, BY TOPIC

2(a) aggregate expenditure deviation

	exp_dev	ln|exp_dev|	rev_dev	ln|rev_dev|	PEFA_comp	OBI	oversight	ln(gni_pc)	polity2	GEI
exp_dev	1.000									
ln|exp_dev|	-0.087*	1.000								
rev_dev	0.666***	-0.060	1.000							
ln|rev_dev|	-0.017	0.384***	-0.004	1.000						
PEFA_comp	0.158***	-0.264***	0.178***	-0.205***	1.000					
OBI	0.162**	-0.276***	0.061	-0.186***	0.324***	1.000				
oversight	-0.024	-0.074	-0.104	-0.084	0.281***	0.431***	1.000			
ln(gni_pc)	0.064	-0.117**	0.102*	-0.098*	0.319***	0.346***	0.209***	1.000		
polity2	-0.079	-0.040	-0.102*	-0.095*	0.142**	0.295***	0.329***	0.156***	1.000	
GEI	0.074	-0.219***	0.106**	-0.154***	0.492***	0.340***	0.378***	0.636***	0.233***	1.000

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively.

2(b) aggregate revenue deviation

	rev_dev	ln|rev_dev|	rev_plan	ln|rev_plan|	gdp_shock	ln(aid_5yr)	resource	ln(gni_pc)	polity2	GEI
rev_dev	1.000									
ln|rev_dev|	-0.004	1.000								
rev_plan	-0.608***	0.160**	1.000							
ln|rev_plan|	-0.188***	0.369***	0.310***	1.000						
gdp_shock	0.143***	-0.069	-0.094	-0.081	1.000					
ln(aid_5yr)	-0.044	0.086	0.150**	0.091	0.018	1.000				
resource	-0.080	0.172***	0.246***	0.113	-0.060	0.181***	1.000			
ln(gni_pc)	0.102*	-0.098*	-0.273***	-0.142**	0.013	-0.723***	-0.317***	1.000		
polity2	-0.102*	-0.095*	0.028	-0.004	0.071	-0.018	-0.212***	0.156***	1.000	
GEI	0.106**	-0.154***	-0.206***	-0.172**	0.030	-0.396***	-0.310***	0.636***	0.233***	1.000

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively.

2(c) expenditure compositional variance

	PI2i_var	PI2ii_var	|exp_dev|	PEFA_tran	PEFA_exec	PI18iv	PI24_comp	ln(gni_pc)	polity2	GEI
PI2i_var	1.000									
PI2ii_var	0.729***	1.000								
|exp_dev|	0.594***	0.765***	1.000							
PEFA_tran	-0.416***	-0.536***	-0.293***	1.000						
PEFA_exec	-0.365***	-0.577***	-0.236***	0.651***	1.000					
PI18iv	-0.371***	-0.639***	-0.291***	0.412***	0.548***	1.000				
PI24_comp	-0.132**	-0.218**	-0.105**	0.198***	0.657***	0.313***	1.000			
ln(gni_pc)	-0.276***	-0.289***	-0.147***	0.299***	0.272***	0.245***	0.000	1.000		
polity2	-0.117*	-0.231**	-0.059	0.270***	0.064	0.090	0.063	0.156***	1.000	
GEI	-0.358***	-0.554***	-0.242***	0.399***	0.470***	0.393***	0.114**	0.636***	0.233***	1.000

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively

ANNEX TABLE 3. MEASURES CONSIDERED FOR QUALITY OF PFM SYSTEMS IN THE ANALYSIS

Variable	Indicators in 2016 framework	Indicators in 2011 framework
PEFA_comp	PI-4 through PI-31, excluding PI-19 and PI-20	PI-4 through PI-28, excluding PI-13, PI-14, and PI-15 (as well as D-1, D-2, and D-3)
PEFA_tran	Those under the pillar "transparency of public finances": budget classification, documentation, central government operations outside financial reports, public access to fiscal information, etc.	PI-5 through PI-8, PI-10, and PI-23: budget classification, comprehensiveness, transparency of inter-governmental relations, availability of information on resources for delivery units, etc.
PEFA_mgmt	Those under the pillar "management of assets and liabilities": fiscal risk reporting, public investment and asset management, etc.	PI-9 and PI17: oversight of aggregate fiscal risk from other public sector entities, recording and management of cash balances, etc.
PEFA_budg	Those under the pillar "policy-based fiscal strategy and budgeting": macroeconomic and fiscal forecasting, fiscal strategy, budget preparation process, etc.	PI-11, PI-12, and PI-27: orderliness and participation in the annual budget process, multi-year perspective in fiscal planning, legislative scrutiny of the budgets, etc.
PEFA_exec	Those under the pillar "predictability and control in budget execution" (excluding those related to revenue administration): predictability of in-year resource allocation, expenditure arrears, procurement, internal audit, etc.	PI-4, PI-16, and PI-18 through PI-21: stock and monitoring of expenditure payment arrears, effectiveness of payroll controls, competition, value for money and controls in procurement, effectiveness of internal audit, etc.
PEFA_acct	Those under the pillar "accounting and reporting": financial data integrity, in-year budget reports, annual financial reports	PI-22, PI-24, and PI-25: timeliness and regularity of accounts reconciliation, quality and timeliness of in-year reports and annual statements
PEFA_audit	Those under the pillar "external scrutiny and audit": external audit, legislative scrutiny of audit	PI-26 and PI-28: scope, nature and follow-up of external audit and legislative scrutiny

Notes: When one-to-one comparison was not possible between dimensions in the 2011 and 2016 frameworks, scores were recalculated as averages of all relevant dimensions. The numerical score for "timeliness of information on transfers" under the transparency pillar, for example, would equal the score for PI-7.2 in 2016 or the average score for PI-8ii and PI-8iii in 2011.

ANNEX TABLE 4. AGGREGATE EXPENDITURE DEVIATIONS, OLS RESULTS BEFORE TRANSFORMATION

(A1) $|exp\_dev_i| = \beta_0 + \beta_1|rev\_dev_i| + \beta_2 PEFA\_comp_i + \beta_x X_i + \varepsilon_i$

(A2) $|exp\_dev_i| = \beta_0 + \beta_1|rev\_dev_i| + \beta_2 PEFA\_comp_i + \beta_3 OBI_i + \beta_4 oversight_i + \beta_x X_i + \varepsilon_i$

	(A1)	(A2)
|rev_dev|	0.565***	0.682***
	(0.044)	(0.052)
PEFA_comp	-3.956***	-3.490**
	(1.110)	(1.457)
OBI		-0.063*
		(0.033)
oversight		0.057
		(0.042)
gni_pc	0.189**	-0.115
	(0.093)	(0.147)
polity2	0.103	0.273**
	(0.089)	(0.125)
GEI	-0.048	-0.042
	(0.033)	(0.041)
constant	14.691***	13.275***
	(2.532)	(3.546)
no. obs.	*320*	*184*
R-squared	*0.425*	*0.613*
RESET p-value	*0.000*	*0.000*
White's p-value	*0.000*	*0.100*
mean VIF	*1.44*	*1.54*

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Standard errors in parentheses.

ANNEX TABLE 5. NATURAL LOG OF AGGREGATE EXPENDITURE DEVIATIONS, OLS RESULTS USING PEFA SUB-SCORES

$$\text{(A3)}\quad \ln|exp\_dev_i| = \beta_0 + \beta_1 PEFA_i + \beta_2 \ln|rev\_dev_i| + \beta_x X_i + \varepsilon_i$$

where $PEFA_i$ can be any of the six sub-scores

	(A3)					
	PEFA_tran	PEFA_mgmt	PEFA_budg	PEFA_exec	PEFA_acct	PEFA_audit
PEFA score	-0.315***	-0.245**	-0.091	-0.193	-0.254**	-0.184*
	(0.109)	(0.112)	(0.147)	(0.147)	(0.108)	(0.106)
ln(|rev_dev|)	0.323***	0.336***	0.344***	0.342***	0.339***	0.340***
	(0.058)	(0.058)	(0.059)	(0.058)	(0.058)	(0.058)
ln(gni_pc)	0.066	0.051	0.035	0.040	0.060	0.047
	(0.092)	(0.092)	(0.093)	(0.093)	(0.092)	(0.092)
polity2	0.016	0.008	0.009	0.007	0.008	0.004
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
GEI	-0.010**	-0.010**	-0.013***	-0.011**	-0.010**	-0.012**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
constant	2.178***	1.999***	1.743***	1.912***	2.061***	1.892***
	(0.294)	(0.289)	(0.400)	(0.351)	(0.297)	(0.285)
no. obs.	*318*	*318*	*318*	*318*	*318*	*318*
R-squared	*0.178*	*0.169*	*0.157*	*0.161*	*0.171*	*0.164*
RESET p-value	*0.310*	*0.195*	*0.064*	*0.194*	*0.462*	*0.187*
White's p-value	*0.942*	*0.618*	*0.909*	*0.232*	*0.263*	*0.904*
mean VIF	*1.39*	*1.38*	*1.40*	*1.45*	*1.36*	*1.34*

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Standard errors in parentheses.

ANNEX TABLE 6. NATURAL LOG OF AGGREGATE REVENUE DEVIATIONS, OLS RESULTS USING PEFA_BUDG

(A4) $$\ln|rev\_dev_i| = \beta_0 + \beta_1 PEFA\_budg + \beta_2|gdp\_shock_i| + \beta_x X_i + \varepsilon_i$$

(A5) $$\ln|rev\_dev_i| = \beta_0 + \beta_1 PEFA\_budg + \beta_2|gdp\_shock_i| + \beta_3 \ln(aid\_5yr_i) + \beta_4 resource_i + \beta_x X_i + \varepsilon_i$$

	(A4)	(A5)
PEFA_budg	-0.304**	-0.344**
	(0.141)	(0.143)
|gdp_shock|	-0.007	-0.009
	(0.017)	(0.017)
ln(aid_5yr)		0.061
		(0.071)
resource		0.401***
		(0.144)
ln(gni_pc)	-0.050	0.054
	(0.090)	(0.127)
polity2	-0.010	-0.008
	(0.013)	(0.013)
GEI	-0.006	-0.005
	(0.005)	(0.005)
constant	2.945***	2.620***
	(0.349)	(0.384)
no. obs.	*319*	*314*
R-squared	*0.055*	*0.089*
RESET p-value	*0.267*	*0.689*
White's p-value	*0.985*	*0.986*
mean VIF	*1.40*	*1.77*

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Standard errors in parentheses.

ANNEX TABLE 7. ECONOMIC COMPOSITIONAL VARIANCE, OLS RESULTS USING PEFA SUB-SCORES

(A6) $PI2ii\_var_i = \beta_0 + \beta_1 PEFA_i + \beta_2 |exp\_dev_i| + \beta_x X_i + \varepsilon_i$

where $PEFA_i$ can be any of the six sub-scores

	(A6)					
	PEFA_tran	PEFA_mgmt	PEFA_budg	PEFA_exec	PEFA_acct	PEFA_audit
PEFA score	-2.688	-5.265***	-2.293	-6.780**	-2.676	-5.014***
	(2.837)	(1.988)	(2.047)	(2.871)	(1.867)	(1.848)
|exp_dev|	0.739***	0.759***	0.758***	0.719***	0.712***	0.671***
	(0.094)	(0.084)	(0.083)	(0.088)	(0.105)	(0.098)
ln(gni_pc)	2.439	2.161	2.106	2.700*	2.150	1.540
	(1.559)	(1.577)	(1.642)	(1.365)	(1.567)	(1.582)
polity2	-0.313	-0.333	-0.324	-0.428**	-0.344	-0.461**
	(0.252)	(0.219)	(0.238)	(0.210)	(0.223)	(0.192)
GEI	-0.193***	-0.156***	-0.213***	-0.135**	-0.221***	-0.183***
	(0.069)	(0.057)	(0.060)	(0.054)	(0.050)	(0.056)
constant	19.464***	23.075***	20.114***	28.274***	21.254***	26.583***
	(5.886)	(4.688)	(5.943)	(6.981)	(5.929)	(5.705)
no. obs.	*95*	*95*	*95*	*95*	*95*	*95*
R-squared	*0.679*	*0.689*	*0.676*	*0.706*	*0.680*	*0.705*
RESET p-value	*0.090*	*0.126*	*0.114*	*0.081*	*0.065*	*0.018*
White's p-value	*0.001*	*0.824*	*0.231*	*0.016*	*0.494*	*0.662*
mean VIF	*1.42*	*1.40*	*1.41*	*1.47*	*1.38*	*1.36*

Notes: *, **, and *** imply statistical significance at 10%, 5%, and 1%, respectively. Standard errors in parentheses; robust standard errors for variations of model with the issue of heteroskedasticity – i.e., using $PEFA\_tran$ and $PEFA\_exec$.