WORKING PAPER Nº IDB-WP-01838

The Unmeasured Cost of Fiscal Execution: Payment Timing and Public Service Delivery

Wladimir Zanoni
Rodrigo López
Marcos A. Rangel

Inter-American Development Bank
Country Department Andean Group

June 2026

The Unmeasured Cost of Fiscal Execution: Payment Timing and Public Service Delivery

Wladimir Zanoni
Rodrigo López
Marcos A. Rangel

Inter-American Development Bank
Country Department Andean Group

June 2026

Keywords: budget execution, fiscal policy, public financial management, health service delivery, state capacity.

JEL Codes: H51, H61, O23

http://www.iadb.org

Rea_can@iadb.org

The Unmeasured Cost of Fiscal Execution: Payment Timing and Public Service Delivery

Wladimir Zanoni* Rodrigo López† Marcos A. Rangel‡

June 5, 2026

Abstract

When governments face fiscal stress, they frequently manage consolidation by delaying cash payments on legally incurred obligations rather than cutting explicit appropriations. This "execution wedge" reduces the real resources available for public production while leaving formal appropriations unchanged. We identify the causal effect of this adjustment margin using monthly administrative data from Ecuador's public health system, exploiting institutional features that make the cash execution pipeline observable and its frictions plausibly exogenous. A one-standard-deviation increase in the execution wedge reduces hospital discharges by 19.9 percent and increases the conditional inpatient mortality rate after 48 hours by 59 percent. Conversely, mortality within 48 hours—an indicator of patient severity at arrival, is unaffected. These findings demonstrate that payment timing operates as an active fiscal policy variable with first-order welfare costs that are unmeasured in standard fiscal policy analysis.

Keywords: budget execution, fiscal policy, public financial management, health service delivery, state capacity.
JEL Codes: H51, H61, O23

*Inter-American Development Bank (IDB).
†Universidad San Francisco de Quito (USFQ).
‡Duke University.

1 Introduction

When governments consolidate under fiscal stress, what is the real cost of adjusting through the timing of payments rather than through explicit spending cuts? Standard fiscal analysis typically tracks appropriations, obligations, deficits, and realized expenditures. Yet in many public financial management systems, the authorization of spending and the release of cash are institutionally distinct and do not necessarily occur contemporaneously. When liquidity becomes scarce, governments can preserve formal appropriations while slowing the conversion of legally incurred obligations into cash payments. In other words, they can use the timing of payment execution as an adjustment margin within the fiscal plumbing that turns authorized obligations into paid inputs.

This paper examines that margin. The execution wedge — the gap between legally incurred obligations and actual cash payments — is the object of measurement; its deliberate management under liquidity stress, which we call execution wedge manipulation, is the phenomenon under study. This wedge matters because public production of goods and services depends not only on what governments authorize or legally obligate, but on what they pay reliably and on time. If fiscal adjustment operates partly through managing the wedge, standard expenditure aggregates may mismeasure the resources effectively available for public production and neglect an important channel through which fiscal policy, when under liquidity stress, reaches frontline services. Directing attention to this institutional practice shifts the locus of fiscal policy from what formal rules indicate the budget should do to what it does, emphasizing that liquidity management and payment timing are neglected yet salient margins of fiscal adjustment.

To discipline measurement and interpretation, we develop a simple conceptual framework in which a government facing a fiscal gap chooses between two adjustment margins. A government facing a fiscal gap can reduce appropriations explicitly, or it can preserve formal budgets while slowing the conversion of obligations into cash payments. The second margin is less visible in standard fiscal aggregates, but it can matter for public production. When service-delivery units cannot self-finance and services must be produced in real time, payment delays can constrain usable inputs during the production window, and later settlement of the obligation may not fully recover the foregone service episode. This compression falls disproportionately on flexible operating inputs — personnel wages are institutionally and legally difficult to reduce at short notice in most public sector settings, so cash shortfalls concentrate on the residual non-wage margin. The framework delivers three empirical implications that organize the analysis below: an increase in the execution wedge should reduce public output; it should operate through compression of timely operating payments rather than through formal appropriations alone; and patient outcomes should deteriorate most along margins that depend on sustained input availability during the service episode.

While the wedge margin of adjustment is well recognized, measuring it precisely and identifying its causal effects on public service delivery have remained

elusive. Payment delays are recorded in administrative financial systems, while their consequences are felt in service-delivery units. Linking the two requires high-frequency data that connect the fiscal pipeline to frontline outcomes. Such linked datasets are rarely available for research or performance monitoring because financial execution records and service-delivery records are typically maintained in separate administrative systems, with limited interoperability and restricted access.

Besides data availability, identifying the causal effect of the wedge on public service delivery is an equally demanding challenge. The budget execution wedge is not randomly distributed: under centralized cash management, liquidity constraints force the Treasury to sequence payments across providers in ways that reflect service profiles, budget structures, and administrative capacity — not by design, but as a consequence of managing payment queues under fiscal stress. Separately, providers in operational distress may generate slower invoice processing upstream. Either channel produces correlation between the wedge and provider performance that does not reflect the causal effect of payment timing.

We address these threats using monthly administrative records from three Ecuadorian sources — Ministry of Economy and Finance (MEF) budget-execution records, Ministry of Public Health (MSP[1]) payroll and staffing records, and hospital activity and mortality records — during the 2015–2019 period. The setting is well suited to this purpose for two reasons. First, the 2014–2015 collapse in oil prices sharply tightened public liquidity in a dollarized economy with no monetary adjustment available, generating sustained and largely exogenous fiscal pressure. Second, Ecuador's institutional architecture (common to many emerging economies) separates the units that incur obligations from the Treasury that controls cash release, making the execution wedge directly observable at the facility level in administrative records.

We identify the causal effect of the wedge using an instrumental-variables design. The instrument is the interaction between monthly leadership turnover in key planning and administrative-management positions at the MSP and each hospital's predetermined budget rigidity. The logic of the first stage is as follows: leadership disruptions slow the upstream processing of payment authorizations, generating execution frictions that propagate into cash disbursements. Crucially, this propagation is not uniform — hospitals with more rigid budget structures have fewer uncommitted internal margins to draw on when authorizations stall, so the same upstream shock translates into a larger and more persistent wedge. This differential pass-through is the source of identifying variation, and we verify it directly in the data rather than assuming it.

The second stage estimates how the instrumented wedge affects hospital throughput, measured by discharges. We then examine the fiscal margin through which the wedge operates by estimating its effect on obligations, aggregate payments, and non-wage operating payments. Finally, we study mortality outcomes to assess both welfare consequences and mechanisms. If payment delays operate through compression of operating inputs, mortality should rise more among

[1]MSP–Ministerio de Salud Pública

patients whose survival depends on sustained care after admission than among patients whose outcomes are more closely tied to admission severity. The estimates should be interpreted as the effect of execution frictions generated by the differential pass-through of upstream administrative disruption, rather than as the average effect of all payment delays in all settings.

The evidence indicates that execution frictions reduce public service delivery. We validate the empirical results against the principal threats to identification — differential demand shocks, staffing changes, budget composition, inherited fiscal stress, and pre-existing hospital trends — and the estimate is stable across all checks. While the estimates are local to Ecuador's institutional setting, the mechanism generalizes to any context combining three features: a centralized treasury that controls cash release separately from the units that incur obligations, a public provider sector without meaningful self-financing capacity, and periodic liquidity stress that forces the treasury to manage payment queues. These conditions are most persistently present across Latin America, Sub-Saharan Africa, and South and Southeast Asia, but they can activate in advanced economies as well — US states operating under binding balanced-budget rules, European subnational governments facing austerity constraints, and eurozone sovereigns during the 2010–2015 debt crisis all share the institutional architecture through which execution wedge manipulation becomes a margin of adjustment.

The findings reveal that payment execution is not a back-office accounting detail. A one-standard-deviation increase in the execution wedge reduces hospital discharges by approximately 19.9 percent — equivalent to roughly 846 fewer discharges per hospital-year at the sample mean. This throughput loss is not driven by a contraction in aggregate payments. Instead, the wedge captures a disruption inside the budget-execution pipeline: legally incurred obligations continue to accumulate while cash payments for non-wage operating goods and services — the inputs most directly tied to day-to-day hospital production — fall sharply. The burden of adjustment falls disproportionately on the flexible operating margin because payroll is the most institutionally protected expenditure category and is less adjustable at monthly frequency. We distinguish deaths occurring within 48 hours of admission, which are more closely tied to severity at arrival and pre-hospital conditions, from deaths occurring after 48 hours, which are more likely to depend on the continuity of inpatient care. If the wedge harms patients through operating-input compression, the mortality response should be more pronounced in the over-48-hour category. This is the pattern we find. A one-standard-deviation increase in the execution wedge raises the conditional over-48-hour inpatient mortality rate by approximately 59 percent, while the under-48-hour mortality rate is statistically unchanged. This asymmetry is difficult to reconcile with a general demand or admission-severity explanation and is instead consistent with a deterioration in the continuity of inpatient care generated by the compression of operating inputs.

This paper contributes to three related literatures. The first is the macro-fiscal literature on the effects of government spending, which estimates how output responds to discretionary changes in appropriations or recorded expen-

ditures (Ramey, 2011; Ilzetzki et al., 2013; Auerbach and Gorodnichenko, 2012). This literature often treats authorized or recorded spending as the relevant measure of the resources available for public production. Our work is closer in spirit to research that questions this equivalence. Checherita-Westphal et al. (2016) document that delayed public payments and the build-up of arrears affect private-sector liquidity and macroeconomic outcomes in European Union countries, while Vieira and Santos (2018) shows how Brazil's federal authorities use unpaid commitments to meet formal fiscal targets while effectively deferring health expenditure. We extend this line of work by identifying the execution wedge—the gap between legally incurred obligations and actual cash payments—as a quantitatively important margin through which fiscal stress affects realized public production but that appropriations-based analysis cannot capture.

The second literature is public financial management research on how budgeted resources reach frontline service providers. Pattanayak (2016) and Flynn and Pessoa (2014a) conceptualize the budget cycle as a sequence of distinct controls—appropriation, commitment, verification, payment authorization, and cash execution—each of which can constrain the conversion of authorized resources into usable inputs. Applied descriptive work documents associations between payment delays, arrears, and resource flows to public health providers (Fritz et al., 2014; Piatti-Fünfkirchen and Schneider, 2018; Moses et al., 2021; Musiega et al., 2023). Our work is also related to Jordán (2025), who exploits a prompt-payment reform in Chile to show that enforcing payment discipline in framework agreements shortens payment lags and lowers procurement prices without observable quality losses, and to Dahis et al. (2026), who assemble transaction-level execution data for Brazilian municipalities and make payment timeliness a measurable dimension of fiscal capacity. We contribute to this literature by providing causal estimates linking high-frequency payment-execution frictions to hospital output and patient outcomes. We also open the institutional black box by showing that the wedge operates through operating-input compression: obligations continue to accumulate while payments for non-wage goods and services fall.

The third literature studies how bureaucratic institutions shape state capacity and public service delivery (Besley and Persson, 2009; Finan et al., 2017; Rasul and Rogger, 2018; Bandiera et al., 2021). This work has shown how personnel selection, autonomy, incentives, and contract design map into measurable performance, but it typically treats fiscal resources as either appropriated amounts or completed expenditures. The administrative mechanisms through which liquidity constraints become binding restrictions on frontline production—payment queues, treasury centralization, cash-flow rationing, and payment authorization—remain less developed empirically within this tradition, although qualitative work has long recognized their importance (Asante et al., 2006; Ahenkan and Azaare, 2018). Our results suggest that fiscal-execution institutions are an underappreciated dimension of state capacity. The relevant input into public service delivery is not only what has been allocated on paper

or legally committed, but what has been executed reliably and on time[2].

The remainder of the paper proceeds as follows. Section 2 describes the historical and institutional setting that motivates the empirical design. Section 3 presents the empirical strategy and identification approach. Section 4 describes the data and sample construction. Section 5 reports the main results: the first-stage evidence, the throughput estimates, the fiscal-pipeline mechanism, and the mortality evidence. Section 6 presents the identification validation. Section 7 discusses the findings and concludes.

2 Historical and Institutional Setting

Commodity-price downturns recurrently generate sustained fiscal pressure in commodity-dependent developing and middle-income economies, with the 2014–2015 oil-price collapse and the broader industrial-commodity slump as the most recent salient instance (Baffes et al., 2015; Gruss, 2014). In commodity-rich settings, such episodes are associated with procyclical fiscal contractions, particularly where institutional quality and rules-based fiscal frameworks are weaker (Frankel et al., 2013; Céspedes and Velasco, 2014).

Public financial management frameworks across these economies formally separate commitment control from payment authorization as distinct dimensions of internal control: governments may legally initiate, commit, or accrue obligations while delaying the cash payments that make inputs available for public production, particularly when cash is centrally managed through a Treasury Single Account. Under fiscal stress, this institutional separation creates scope for an adjustment margin operating through centralized cash management, in which formal appropriations and recorded obligations are preserved while payment timing absorbs the shock (Pattanayak, 2016; Pattanayak and Fainboim, 2011).

The resulting wedge between obligations and cash payments, and the related risk of arrears accumulation, is widely documented across low and middle-income countries in PEFA assessments, indicating that payment timing can operate as a recurrent rather than exceptional margin of adjustment under revenue stress (Flynn and Pessoa, 2014b). We study this general mechanism in the context of Ecuador's 2015–2019 fiscal-stress episode, which followed the oil-price collapse and placed sustained pressure on public liquidity in a dollarized economy with no monetary policy and with a high non-oil primary deficit and limited access to international financing (International Monetary Fund, 2019). Ecuador is a strategic empirical case rather than an idiosyncratic one: it combines high revenue exposure to commodity prices, a centralized treasury architecture broadly typical of public financial management systems in the region, and granular administrative data covering the full execution chain (López et al., 2026). This combination allows us to identify the wedge between legally initiated obligations and actual cash payments with a degree of precision that is

[2]For related institutional arguments in Andean contexts, see López et al. (2026, 2025)

rarely feasible elsewhere, so that the estimates obtained in this setting speak to a broader, cross-country phenomenon.

This section describes the features of that architecture necessary to explain our identification strategy. Four elements matter: the separation between administrative and financial execution, the centralization of public liquidity in the Treasury at MEF, the hierarchical processing of spending requests within the health system, and heterogeneity in hospital budget rigidity. Together, these features generate an observable wedge between legally initiated obligations and cash payments, and they make that wedge responsive to upstream administrative shocks in ways that vary systematically across hospitals.

2.1 Separation of administrative and financial execution

Ecuador's budget process is organized around an institutional separation between the agency that executes spending and the agency that authorizes payment[3]. The MEF defines the macro-fiscal envelope, opens institutional budgets in the integrated financial management system (ESIGEF), and retains legal authority over cash disbursements through the National Treasury, the subsecretariat within the MEF that operates the Treasury Single Account (*Cuenta Única del Tesoro*, CUT) on behalf of the central government. The MSP formulates the sectoral budget within those ceilings and administers intra-year execution. Hospitals are operational units where spending needs arise, but they do not control aggregate fiscal space or the timing of payments. Expenditures advance through a sequence of administrative steps (precommitment, commitment, accrual) that the MSP can complete without immediate cash disbursement. Once obligations have been legally incurred, settlement depends on Treasury decisions, financial programming, and cash availability. The wedge between precommitments and payments that we exploit empirically, defined formally in equation (3), is therefore not a measurement artifact but a structural feature of a system in which administrative and financial execution operate on different timelines.

2.2 Centralized liquidity and the operational bite of payment timing

The separation described above would be of limited consequence if hospitals could finance ongoing operations from their own cash flows while waiting for the Treasury to settle accrued obligations. In Ecuador, they cannot. Public sector liquidity is administered through the Treasury Single Account introduced above, and held at the Central Bank and operated by the MEF. Public health care in Ecuador is provided free of charge at the point of service, and MSP

[3]Public financial management frameworks distinguish commitment control and payment authorization as separate dimensions of internal control (PEFA Secretariat, 2016; Pattanayak, 2016), and weaknesses in the conversion of obligations into timely payments have been documented as constraints on health service delivery in developing-country settings (Piatti-Fünfkirchen and Schneider, 2018).

hospitals do not generate revenue from patient care. Their resources are financed entirely through budgetary allocations channeled by the General State Budget[4]. Their treatment under the Central Unity of Treasury (CUT) regime implies that, in operational terms, hospitals do not hold meaningful cash balances of their own and cannot draw on their own funds to honor obligations whose payment has been deferred. The wedge therefore does not measure a narrow payment-processing delay. It is the aggregate operational manifestation of all execution frictions that become binding under the CUT regime — procurement slowdowns, authorization backlogs, request-routing disruptions, and supplier-payment bottlenecks all appear in the wedge the moment they constrain input acquisition. In an institutional setting where every operationally binding transaction must clear the payment pipeline, the wedge and the operational disruption are the same thing measured at the point where fiscal and operational reality converge. The exclusion restriction in Section 3 follows from this architecture rather than being imposed on it.

As a result, even constitutionally protected or legally consolidated spending commitments remain contingent on the timing of Treasury disbursements during execution. This institutional design has three implications for our empirical analysis. First, it gives the wedge E_{it} its operational meaning: because hospitals cannot self-finance, a delay in the conversion of accrued obligations into payments is a binding constraint on inputs available for service delivery, not an accounting reclassification.

Second, it eliminates a margin of local response that would otherwise contaminate the pass-through from T_t into the observed wedge. If hospitals retained operational cash balances, they could partially offset upstream administrative delays by self-financing operating inputs, and the hospital-level wedge would reflect not only the disruption itself but also heterogeneous liquidity-management decisions across hospitals. Under the CUT regime, no such margin exists, so the pass-through from T_t into the wedge is governed by the budgetary structure captured by $R_{i,t-1}$ rather than by hospital-level cash management.

Third, it identifies the lagged budget-rigidity measure $R_{i,t-1}$, formally defined in equation (1), as the relevant moderator of pass-through. When the only feasible local response to an upstream delay is reallocation within the codified budget, the hospital's capacity to absorb the shock depends on the share of its budget that remains flexibly reallocable.

Because all operationally binding input acquisition must pass through the commitment and payment pipeline, and because hospitals have no off-pipeline financing alternative under the CUT regime, the gap between obligations and payments captures the aggregate downstream consequence of upstream administrative frictions regardless of their specific origin. Procurement delays, autho-

[4]The only exception corresponds to self-generated resources (*recursos de autogestión*) arising from inter-institutional account settlements within the public health network, when patients affiliated to another subsystem of the public health network are attended at an MSP facility. They do not constitute operational revenue for the hospital and, even when they materialize, they are deposited into the Central Unity of Treasury rather than retained as facility-level cash balances.

rization slowdowns, request-sequencing disruptions, and supplier-payment bottlenecks are therefore not separate from the wedge once they become binding for input acquisition; they are precisely the events that widen it.

2.3 Hierarchical transmission of administrative shocks

Hospital execution outcomes depend not only on local operational needs but also on how spending requests are processed upstream within the administrative hierarchy of the MSP. Budget allocation, prioritization, and the sequencing of execution are coordinated by a central MSP office that oversees how operational requests move through the health system. Personnel occupying key planning and administrative-management positions within the central MSP administration therefore influence how operational needs are translated into validated requests and how those requests are sequenced across the territorial network. Changes in these positions can disrupt the processing chain even when no contemporaneous operational change occurs within a given hospital. Administrative routines may be interrupted, institutional knowledge may be temporarily lost, and the prioritization or sequencing of requests may shift across hospitals and regions.

Because these positions operate upstream from hospitals and serve the entire health network, the resulting disruption is common across hospitals in a given month and outside the control of any individual establishment. This is the variation captured by T_t in our empirical design.

The transmission of these shocks is not purely mechanical. Territorial administrative nodes mediate how central disruptions are expressed operationally at the hospital level. The same upstream disruption may therefore generate different execution conditions across hospitals depending on how requests are processed through these shared layers and on the hospital's own capacity to absorb delays through internal budgetary adjustments. The degree to which a given hospital can absorb those delays depends on the flexibility of its own budget structure, to which we now turn.

2.4 Heterogeneous budget rigidity and differential pass-through

The same upstream disruption does not generate uniform execution frictions across hospitals. The intensity of pass-through depends on the composition of the hospital's budget at the time the disruption occurs. Hospitals whose budgetary items are concentrated in inflexible categories, such as personnel obligations[5], have fewer internal margins to reallocate spending or smooth the timing of execution. When upstream processing slows, these hospitals have less room to absorb the delay before it appears as a gap between precommitments and

[5]Personnel obligations are protected by labor regulations and civil-service rules and are effectively non-adjustable at the monthly frequency relevant for our analysis. Beyond their legal protection, payroll adjustments carry social and political costs that delays in non-wage operating payments do not, so governments facing fiscal stress tend to preserve payroll and absorb adjustment elsewhere in the expenditure pipeline.

payments. Within this structure, non-wage operating expenditure, the budget category that covers goods and services required for day-to-day hospital production, is the natural margin of adjustment when liquidity tightens, because it can be reprogrammed or delayed at higher frequency than payroll or contractual obligations[6]. Hospitals with greater budgetary flexibility can buffer part of the same disruption through reallocation across spending categories, so a smaller share of the upstream shock is transmitted into the wedge. Budget rigidity is the hospital-level moderator of pass-through. It varies across hospitals because facilities differ in size, care level, and expenditure composition, and it evolves within hospitals as commitments accumulate during the fiscal year. This is the variation captured by $R_{i,t-1}$ in equation (1), measured before current MSP disruptions and current MEF payment decisions are realized.

These four features jointly motivate the empirical strategy that follows. The hospital-level wedge is the downstream manifestation of a process that begins outside the hospital and that hospitals cannot bypass through their own liquidity, which limits the concern that it is driven by contemporaneous hospital-level demand or local managerial decisions. MSP-level disruption T_t is common to all hospitals and originates upstream of the payment queue, in the administrative processing that the MSP completes before obligations reach the Treasury for cash settlement. Because T_t does not operate through MEF treasury decisions about which hospitals to pay, it is plausibly orthogonal to hospital-specific service shocks once aggregate time effects are absorbed. Lagged hospital rigidity $R_{i,t-1}$ is predetermined relative to the current wedge and to current hospital output. The interaction $Z_{it} = T_t \times R_{i,t-1}$ therefore motivates an empirical design that isolates execution-friction variation generated when a common upstream shock is transmitted into hospital operations through pre-existing budgetary constraints. The next subsection uses these institutional features to derive the empirical predictions that guide the analysis; the following section then formalizes the identification strategy.

2.5 Conceptual Framework and Empirical Predictions

The institutional setting described above motivates a simple conceptual framework. A government facing fiscal stress can adjust through two margins: it can reduce appropriations explicitly, or it can preserve formal budgets while slowing the conversion of legally incurred obligations into cash payments. The second margin is less visible in standard fiscal aggregates but consequential for public production whenever two conditions hold. First, service-delivery units cannot self-finance — they have no independent access to liquidity that would allow them to bridge the gap between approved and paid resources. Second, services must be produced in real time — output foregone in one period cannot be recovered in the next. When both conditions hold, payment delays reduce the

[6]This expenditure category corresponds to MEF expenditure Group 53 in the eSIGEF nomenclature, which is the operating margin used in the fiscal-pipeline mechanism evidence in Section 5.4.

inputs available during the production window even if the obligation is eventually settled, and the foregone service episode is irreversible. The institutional features documented in Sections 2.2 and 2.4 confirm that both conditions characterize Ecuador's public hospitals: the CUT regime eliminates self-financing, and inpatient care is non-storable by nature. The framework is formalized in Appendix D; here we state the three empirical predictions that organize the analysis.

The first prediction concerns throughput. The wedge rises when authorized spending accumulates faster than cash is released. Because hospitals cannot finance operations outside the Treasury pipeline, this is not merely an accounting delay — it is a binding constraint on the timely availability of inputs. Hospital throughput, measured by completed discharges, should therefore fall when payment execution slows.

The second prediction concerns the fiscal mechanism. If the wedge reflects cash-execution frictions rather than formal budget cuts, obligations may continue to accumulate even as timely payments fall. Because payroll and other rigid expenditure categories are institutionally and legally protected from short-notice adjustment, the burden of cash rationing should fall disproportionately on flexible non-wage operating inputs — the goods and services required for day-to-day hospital production. A widening wedge should therefore compress operating payments while leaving obligations largely unchanged.

The third prediction concerns patient outcomes, and it is the most discriminating test of the mechanism. If payment delays operate through compression of operating inputs, the consequences should be strongest for patients whose survival depends on the sustained availability of those inputs during the inpatient episode. This motivates a mortality split that is not symmetric in its theoretical content. Deaths occurring within 48 hours of admission are more closely tied to severity at arrival, pre-hospital conditions, and the initial clinical response, while deaths occurring after 48 hours are more likely to depend on the hospital's ability to sustain care during the inpatient episode. Deaths occurring after 48 hours depend on the hospital's ability to sustain pharmaceutical supply, nursing coverage, monitoring, and consumables throughout the inpatient stay — precisely the inputs compressed when non-wage operating payments slow. The prediction is therefore directional: over-48-hour mortality should respond to the wedge, while under-48-hour mortality should not. This asymmetry cannot be reconciled with a general demand shock or admission-severity story, both of which would affect both margins proportionally. It is the signature of operating-input compression, and it is what we test in Section 5.5. The empirical analysis follows these three predictions in sequence: throughput, fiscal mechanism, mortality.

3 Empirical Strategy

The empirical challenge is that budget-execution frictions are not randomly assigned. Hospitals with weaker management conditions may both accumulate

larger execution gaps and produce fewer discharges; central authorities may prioritize payments to hospitals facing greater service pressure; and unobserved demand or capacity shocks may affect both budget execution and hospital output. A simple regression of hospital performance on the execution wedge would therefore combine the causal effect of execution frictions with endogenous fiscal management, persistent hospital differences, and contemporaneous shocks to service demand. We address this problem by constructing an instrument based on the interaction between two time-varying institutional variables that together determine how central administrative disruptions are transmitted into hospital-level execution frictions.

The two components of the instrument

MSP-level administrative disruption. The first component captures disruption in the central administrative chain through which spending obligations are planned, processed, and converted into payments. Let T_t denote the monthly count of leadership changes in key MSP positions involved in planning, administrative management, and the sequencing of budget execution. The baseline index is constructed from turnover in four positions: the administrative-financial coordinator, the planning coordinator, the administrative director, and the finance director. Higher values of T_t indicate greater disruption at the MSP in month t. This variable is common to all hospitals in a given month: it reflects events at the ministry that are outside any individual hospital's control and that vary over time as political and administrative conditions change.

Hospital budget rigidity. The second component captures how constrained each hospital is in absorbing administrative disruptions through budgetary reallocation as the budget cycle unfolds. We define lagged budget rigidity as

$$R_{i,t-1} \equiv 1 - \frac{\text{Flexible Budget}_{i,t-1}}{\text{Codified Budget}_{i,t-1}}, \tag{1}$$

where higher values indicate that a smaller share of hospital i's codified budget was available for reallocation in month $t-1$. Inflexible expenditure categories include budget lines that are difficult to adjust in the short run, such as personnel obligations, fixed service contracts, and mandatory transfers. More flexible categories include expenditure lines that are more amenable to reprogramming or delay as execution conditions change.

Crucially, $R_{i,t-1}$ is neither fixed across time nor common across hospitals. It varies across hospitals because facilities differ in size, care level, staffing structure, and expenditure composition. It also varies within hospitals over time as commitments evolve during the fiscal year. This variation is measured before the current MSP disruption and before current MEF payment decisions are realized. Lagged rigidity is therefore predetermined with respect to current execution frictions and current hospital output, while still allowing hospital budget structures to evolve over time.

The instrument

We instrument the execution wedge with the interaction between MSP-level administrative disruption and lagged hospital budget rigidity:

$$Z_{it} = T_t \times R_{i,t-1}. \tag{2}$$

The identifying variation comes from whether common MSP-level disruptions generate larger execution wedges in hospitals that entered the month with fewer flexible budgetary margins.

The first-stage logic is institutional. The same MSP-level disruption should generate larger execution frictions in hospitals whose pre-existing budget structure leaves fewer margins for adjustment. A hospital with a high share of rigid expenditure commitments has less room to absorb disruption when the budget process moves from planning and precommitment to accrual and payment. In that setting, administrative disruption is more likely to appear as a gap between precommitments and payments. By contrast, a hospital with more flexible margins can partially buffer the same disruption with less impact on measured execution. The instrument therefore captures the portion of execution-friction variation generated by the interaction between a common ministerial disruption and a predetermined hospital-level budget constraint.[7]

The execution wedge

The endogenous variable is the hospital-level execution wedge. For hospital i in month t, we define

$$E_{it} \equiv \log\left(\frac{\text{Precommit}_{it} + \epsilon}{\text{Paid}_{it} + \epsilon}\right), \tag{3}$$

where Precommit_{it} denotes monthly precommitments, Paid_{it} denotes monthly cash payments, and ϵ is a small constant used to retain observations with zero values. Higher values of E_{it} indicate that legally initiated spending obligations are accumulating relative to payments, capturing tighter execution conditions inside the budget pipeline.

The wedge has a precise institutional interpretation that bears on how the exclusion restriction should be understood. Because hospitals operate under the CUT regime and cannot self-finance, every operationally binding input acquisition must clear the commitment and payment pipeline before cash is released. There is no off-pipeline financing alternative. This means that upstream administrative frictions — whether they originate in procurement processing, request routing, authorization sequencing, or supplier-payment management — do not constitute separate channels from the wedge once they become binding for input acquisition. They are precisely the events that cause obligations to accumulate relative to payments and therefore widen E_{it}. The wedge is therefore not a narrow proxy for one specific type of friction; it is the aggregate downstream fiscal

[7] Alternative formulations, including distributed-lag IV models, high-versus-low rigidity TWFE comparisons, and Bartik-style shift-share exposure designs, are not well aligned with the institutional structure of the execution wedge.

expression of all pipeline-based execution delays, regardless of their administrative origin. The exclusion restriction should accordingly be understood as requiring that the instrument affects hospital discharges only through channels that do not pass through this pipeline — such as differential patient demand, direct staffing shocks, or clinical-capacity changes — rather than requiring orthogonality to every conceivable upstream administrative disruption.

Estimation

The first-stage equation is

$$E_{it} = \pi Z_{it} + \alpha_i + \lambda_m + \tau_y + u_{it}. \tag{4}$$

where α_i are hospital fixed effects, λ_m are calendar-month fixed effects, and τ_y are year fixed effects. Hospital fixed effects absorb time-invariant differences across hospitals, including permanent differences in care level, geographic location, and baseline management conditions. Calendar-month fixed effects absorb seasonal patterns common to all hospitals, and year fixed effects absorb aggregate fiscal trends. The baseline specification estimates this equation without additional controls; the validation analysis then adds lagged demand, staffing, budget-composition, and inherited fiscal-state controls to evaluate specific identification threats.

The second-stage equation is

$$Y_{it} = \beta \widehat{E}_{it} + \alpha_i + \lambda_m + \tau_y + \varepsilon_{it}. \tag{5}$$

The main outcome is hospital throughput, measured by the log number of hospital discharges. The parameter β captures the effect of execution frictions on hospital output during Ecuador's 2015–2019 fiscal-stress episode. Standard errors are clustered at the hospital level. Mechanism specifications use the same instrumented execution wedge to study the fiscal-pipeline response and the associated patient-outcome consequences documented below.

Identifying assumption and validation

The identifying assumption is that, conditional on hospital, calendar-month, and year fixed effects, the interaction Z_{it} affects hospital discharges through the execution wedge rather than through other contemporaneous determinants of hospital output. The assumption is weaker than requiring MSP-level administrative disruption itself to be unrelated to all aggregate shocks: calendar-month and year fixed effects absorb shocks common to all hospitals. The identifying variation comes from whether the same MSP-level disruption translates into larger execution frictions in hospitals with higher lagged rigidity. The assumption would be violated if, precisely in months of MSP disruption, hospitals with higher lagged rigidity also experienced differential unobserved demand shocks, staffing changes, budget-composition shifts, inherited fiscal stress, or structural trends that affected discharges independently of the execution wedge.

The interpretation of the IV estimate is local to the execution frictions generated when MSP-level planning and administrative-management disruptions interact with heterogeneous, time-varying hospital budget rigidity. The design does not require all execution delays to be random. During fiscal stress, payment timing may itself be an administratively available margin of adjustment. The empirical strategy isolates the plausibly exogenous component of that margin: variation generated when a common ministerial disruption is transmitted differently across hospitals because their lagged budget structures leave them with different margins to absorb delays. This source of variation is plausibly distinct from deliberate treasury prioritization decisions because it arises from the interaction between a common MSP-level disruption and predetermined hospital-level rigidity, rather than from contemporaneous MEF choices about which hospitals to pay.

This variation approximates the experimental contrast of comparing otherwise similar hospitals facing different levels of execution friction. The instrument shifts the execution wedge through institutional variation that is not directly chosen in response to contemporaneous hospital demand or service production. Under the exclusion restriction, the resulting variation in hospital performance identifies the causal effect of the execution wedge for the component of execution frictions shifted by the instrument.

4 Data and Sources

The empirical analysis combines administrative records from MEF, MSP, and the National Institute of Statistics and Census (INEC). We construct a hospital-month panel covering Ecuador's 2015–2019 fiscal-stress episode, with the hospital in a calendar month as the unit of observation. The MEF budget-execution system provides monthly hospital-level allocations and execution variables, including codified budgets, precommitments, accrued obligations, and cash payments, disaggregated by expenditure category. These records are used to construct the execution wedge, defined as the log ratio of precommitments to payments, and the fiscal-pipeline variables used in the mechanism analysis: precommitments and financial obligations, total monthly payments, operating goods and services payments, and operating goods and services accrued obligations. Operating goods and services correspond to MEF expenditure Group 53, the non-wage operating margin most directly related to day-to-day hospital production.

We harmonize these fiscal records with MEF/MSP personnel and administrative data, which identify monthly staffing levels by broad occupational categories and changes in key MSP planning and administrative-management positions. These records are used to construct the MSP-level administrative disruption index used in the instrument and to build lagged staffing controls for the validation analysis. We then merge the fiscal and personnel data with MSP and INEC hospital statistics on discharges and mortality. The main service-delivery outcome is hospital throughput, measured as the log number of hospi-

tal discharges; mortality records allow us to distinguish deaths within and after 48 hours of admission. Hospital identifiers and monthly dates are harmonized across all sources. The resulting panel spans January 2015 through December 2019, and the main discharge estimation sample comprises 6,811 hospital-month observations from 120 hospitals. The final dataset includes the execution wedge, the instrument, hospital discharges, mortality measures, fiscal-pipeline variables, staffing controls, bed-capacity measures, budget-composition controls, and inherited fiscal-state controls. Appendix A reports the sample-construction flow and annual coverage.

5 Empirical Results

This section proceeds in five steps. We first show that the execution wedge is a policy-relevant margin and that administrative turbulence at the MSP raises execution frictions more in hospitals with more rigid budget structures. We then estimate the causal effect of those frictions on hospital performance, showing that a larger wedge reduces throughput. We next examine the fiscal-pipeline mechanism, showing that obligations continue to accumulate while operating payments fall. We then use mortality outcomes to test whether this operating-payment compression affects patient welfare in the way implied by the mechanism: mortality should rise among patients whose outcomes depend on sustained inpatient care, but not among patients whose deaths are more likely determined by condition severity at arrival. Finally, we examine the validity of this interpretation through demand, staffing, fiscal-state, timing, and predetermined-capacity checks.

5.1 Administrative Disruptions, Budget Rigidity and the Execution Wedge

Figure 1 provides descriptive evidence on the relationship between MSP-level administrative disruption, lagged hospital budget rigidity, and the execution wedge during the 2015–2019 fiscal-stress episode. Using the notation introduced above, the figure plots turbulence T_t together with the deseasonalized execution wedge, E_{it}, separately for hospitals in the low- and high-rigidity parts of the $R_{i,t-1}$ distribution. Specifically, the low- and high-rigidity series correspond to hospitals at the 25th and 75th percentiles of lagged budget rigidity.

For each month, we compute the average execution wedge among low-rigidity hospitals and among high-rigidity hospitals. We then residualize these two wedge series, together with the MSP disruption index T_t, with respect to month-of-year fixed effects, standardize them using the 2015–2019 distribution, and smooth them using a three-month moving average. The figure is descriptive rather than a formal first-stage estimate, but it shows whether the time-series movement in E_{it} is consistent with the institutional logic of the instrument $Z_{it} = T_t \times R_{i,t-1}$.

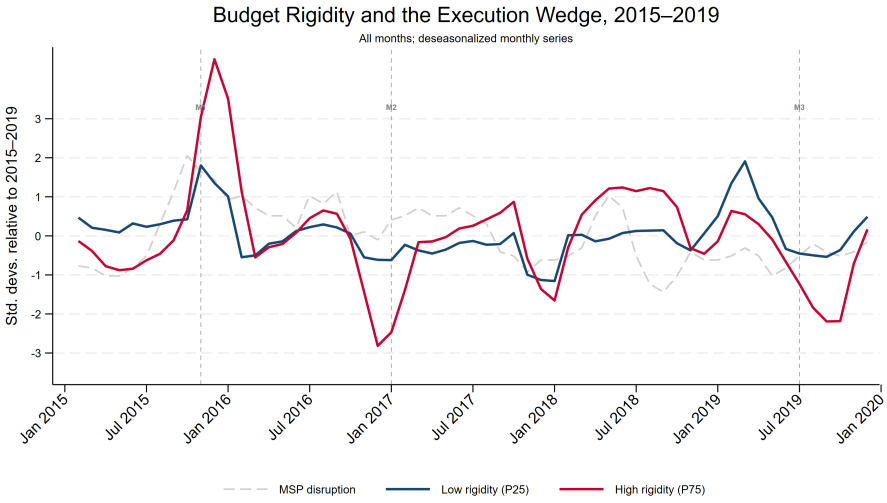

Figure 1: Execution Wedge and MSP Disruption, 2015–2019

Notes: The figure plots deseasonalized monthly series for MSP disruption and the execution wedge among hospitals with low and high lagged budget rigidity. MSP disruption is T_t, the monthly count of leadership changes in key MSP planning and administrative-management positions. The execution wedge is E_{it}, defined as the log ratio of precommitments to payments. Low- and high-rigidity series correspond to hospitals in the lower and upper tails of the lagged budget-rigidity distribution, $R_{i,t-1}$. All series are residualized with respect to month-of-year fixed effects, standardized using the 2015–2019 distribution, and smoothed using a three-month moving average. Vertical lines denote MSP ministerial transitions.

The figure shows that movements in the execution wedge accompany movements in MSP disruption, and that this relationship is stronger among hospitals with more rigid budget structures.[8] The high-rigidity series displays larger movements in the wedge than the low-rigidity series, particularly during periods in which MSP disruption is elevated. This pattern is consistent with the first-stage logic: a common disruption in the MSP administrative chain is more likely to translate into execution frictions where the hospital's lagged budget structure leaves fewer margins to absorb delays through reallocation.

5.2 The First Stage: The Instrument Predicts the Wedge

We now turn from descriptive evidence to the formal first stage. The identifying logic of the design is that the same upstream administrative shock should

[8] Appendix Figure B.1 replicates the exercise restricting the sample to March through October, excluding November through February, when budget execution and payment behavior are more likely to reflect year-end and year-start administrative patterns. The same qualitative relationship remains visible: the wedge is more responsive among high-rigidity hospitals than among low-rigidity hospitals. This suggests that the pattern is not driven solely by recurrent end-of-year or beginning-of-year bunching in budget execution.

generate different execution frictions depending on the hospital's predetermined degree of budget rigidity. If this mechanism is present in the data, the interaction between MSP turbulence and lagged budget rigidity should strongly predict the hospital-level execution wedge.

Table 1: The First Stage: Turbulence and Rigidity Predict the Wedge

	(1) Pooled IV first stage	(2) Low rigidity (P25)	(3) High rigidity (P75)
MSP disruption × lagged rigidity	0.243***		
	(0.051)		
MSP disruption		0.019	0.495***
		(0.096)	(0.104)
First-stage F-statistic	23.02	0.04	22.53
Observations	6811	1697	1702

Notes: The dependent variable is the hospital-level execution wedge, defined as the log ratio of precommitments to payments. All specifications include hospital fixed effects, calendar-month fixed effects, and year fixed effects. Robust standard errors are clustered at the hospital level. Columns (2) and (3) restrict the sample to hospitals at the 25th and 75th percentiles of lagged budget rigidity.

Table 1 confirms that this is the case. In the pooled specification, the interaction between MSP turbulence and lagged rigidity is a strong predictor of the wedge. Controlling for hospital fixed effects, calendar-month effects, and year effects, the coefficient on the instrument is 0.243, with a first-stage F-statistic of 23.02. Because the dependent variable is the log ratio of precommitments to payments, this estimate implies that a one-unit increase in the instrument is associated with roughly a 27.5 percent increase in the precommitment-to-payment ratio[9]. The instrument, therefore, has substantial predictive power for the execution wedge.

This first-stage relationship is not uniform across hospitals. Column (1) reports the formal pooled first stage from equation (4), where the execution wedge is regressed on the instrument $Z_{it} = T_t \times R_{i,t-1}$, controlling for hospital fixed effects, calendar-month fixed effects, and year fixed effects. Columns (2) and (3) then provide a diagnostic split-sample exercise: they estimate the association between MSP disruption T_t and the execution wedge separately for hospitals in the lower and upper quartiles of the lagged rigidity distribution.

The results show that the relationship between MSP disruption and the execution wedge is concentrated among more rigid hospitals. In the lower-rigidity group, the coefficient on MSP disruption is small and statistically weak. In the upper-rigidity group, the coefficient is larger and more precisely estimated. Thus, consistent with Figure 1, the same upstream administrative disruption is more strongly associated with execution frictions where lagged budget rigidity leaves hospitals with fewer margins to absorb delays.

[9]This semi-elasticity is estimated as $100 \times (\exp(0.243) - 1) \approx 27.5\%$.

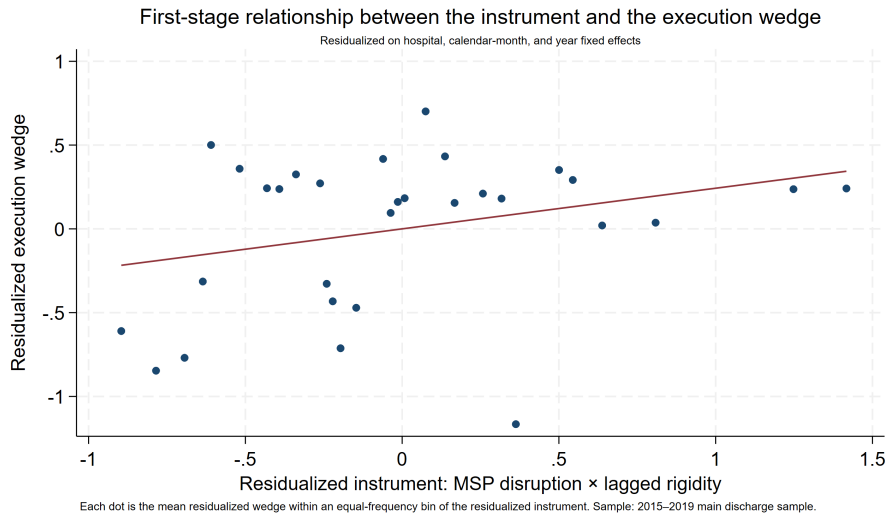

Figure 2: First-stage relationship between the instrument and the execution wedge

Notes: The figure plots the residualized hospital-level execution wedge against the residualized instrument, defined as MSP turbulence interacted with lagged budget rigidity. Both variables are residualized with respect to hospital fixed effects, calendar-month fixed effects, and year fixed effects. Each dot represents the mean residualized wedge within an equal-frequency bin of the residualized instrument; the solid line is the fitted linear relationship.

Figure 2 provides a visual representation of this first stage. After residualizing both the instrument and the wedge with respect to hospital fixed effects, calendar-month effects, and year effects, the figure shows a clear positive relationship between the two variables. This pattern is consistent with the regression results and reinforces the central mechanism of the design: administrative turbulence at the MSP predicts hospital-level execution frictions, and that pass-through is amplified by predetermined budget rigidity.

Taken together, these results provide strong support for the relevance condition of the instrument. The interaction between MSP turbulence and lagged rigidity is not only statistically significant in the pooled sample, but also behaves exactly as the institutional mechanism would predict: it has limited predictive power among the most flexible hospitals and substantially greater predictive power among the most rigid ones.

5.3 Execution Frictions and Hospital Throughput

We now examine whether liquidity-driven execution frictions translate into measurable losses in hospital throughput. Table 2 presents the main throughput estimates. Columns (1) and (2) report OLS benchmarks, without and with the full set of validation controls. Column (3) reports the reduced-form relation-

ship between the instrument and hospital throughput. Column (4) presents our preferred two-stage least squares estimate, with the corresponding first-stage F-statistic reported in the table. The first-stage relationship itself is documented in Table 1.

Table 2: Execution Frictions and Hospital Throughput

	(1) OLS	(2) OLS + controls	(3) Reduced form	(4) IV
Dependent variable	Log discharges	Log discharges	Log discharges	Log discharges
Execution wedge	-0.0046	-0.0044	–	-0.0613
	(0.0017)	(0.0014)	–	(0.0261)
p-value	0.0078	0.0027	–	0.0189
Instrument	–	–	-0.0149	–
	–	–	(0.0054)	–
p-value	–	–	0.0065	–
First-stage F-statistic	–	–	–	22.62
Observations	6811	6811	6811	6811
Hospitals	120	120	120	120
One-SD effect (%)	-1.63	-1.59	–	-19.86

Notes: The table reports estimates using the 2015–2019 hospital-month panel. The main outcome is log hospital discharges. Columns (1) and (2) report OLS estimates of the relationship between the execution wedge and log discharges, without and with the full set of validation controls. Column (3) reports the reduced-form effect of the instrument on log discharges. Column (4) reports the preferred IV estimate. The execution wedge is defined as the log ratio of precommitments to payments. The instrument is MSP-level administrative disruption interacted with lagged hospital budget rigidity. All specifications include hospital fixed effects, calendar-month fixed effects, and year fixed effects. Standard errors are clustered at the hospital level. The first-stage F-statistic corresponds to the excluded instrument in the IV specification.

The OLS benchmarks in Columns (1) and (2) show a negative and statistically significant association between the execution wedge and hospital discharges. The point estimate is stable after adding controls, moving only modestly from -0.0046 to -0.0044. These controls include lagged demand, staffing, budget-composition, and inherited fiscal-state variables. The stability of the OLS coefficient indicates that the negative wedge-throughput relationship is not mechanically driven by observable changes in recent service pressure, measured production capacity, the scale and composition of hospital budgets, or pre-existing fiscal distress. These estimates are useful descriptive benchmarks, but they do not address the endogeneity of the observed execution wedge.

Column (3) reports the reduced form. The institutional variation that predicts larger execution wedges in Table 1 also predicts lower hospital discharges. This is the most direct evidence that the identifying variation shifts hospital output in the expected direction: it requires no scaling by the first stage and is independent of assumptions about the magnitude of the wedge effect. Column (4) reports the preferred IV estimate. The coefficient on the instrumented execution wedge is -0.0613, with a standard error of 0.0261 and a p-value of 0.019. The first-stage F-statistic reported in the table confirms that the IV estimate is based on a strong first stage. Evaluated at the standard deviation of the

execution wedge in the IV estimation sample, a one-standard-deviation increase in the wedge reduces hospital discharges by approximately 19.9 percent. The first stage shows that the instrument shifts the execution wedge; the reduced form shows that the same variation reduces discharges; and the IV estimate quantifies the implied throughput loss associated with execution frictions.

The OLS estimates are considerably smaller in magnitude than the IV estimate. Three factors help explain this gap. First, the observed wedge combines harmful execution frictions with endogenous execution activity. Hospitals facing higher service demand or possessing stronger administrative capacity may generate more precommitments while simultaneously maintaining higher throughput, weakening the raw association between the wedge and discharges. Second, high-frequency administrative budget records may contain timing mismatches between the recording of obligations and the recording of payments in ESIGEF, which would further attenuate OLS. Third, the IV estimate isolates a narrower and more policy-relevant source of variation: the component of the execution wedge generated by upstream administrative disruption interacting with predetermined hospital budget structure. This variation is more closely aligned with the institutional mechanism of interest: disruptions in the budget-execution pipeline that translate into local cash-execution stress.

Appendix Table B.1 further clarifies the role of budget rigidity in the design. An interaction-IV specification allowing the effect of the execution wedge to vary with standardized lagged rigidity shows little evidence of heterogeneous second-stage effects. The interaction coefficient is small and statistically insignificant, and the marginal effects at one standard deviation below and above mean rigidity are nearly identical. This suggests that rigidity primarily shapes exposure to execution frictions through the first stage, rather than making hospitals differentially sensitive to a given realized wedge.

Taken together, Table 2 and Appendix Table B.1 establish the core empirical finding of the paper. Delays in the conversion of legally initiated spending obligations into cash payments reduce realized hospital output. This execution margin is economically meaningful and is not captured by standard fiscal aggregates based only on approved budgets, obligations, or aggregate expenditure. The relevant operational input for public service delivery is not only what is authorized or legally initiated, but what is paid reliably and on time.

5.4 Fiscal Pipeline Mechanism: Obligations Accumulate While Operating Payments Fall

The first-stage results show that MSP-level administrative disruption, interacted with lagged hospital budget rigidity, predicts the execution wedge. We now examine what this wedge captures institutionally. The key question is whether the wedge reflects a simple contraction in aggregate payments, or instead a disruption inside the budget-execution pipeline: obligations continue to move forward, but timely operating payments fail to keep pace. Table 3 estimates the effect of the instrumented execution wedge on four fiscal margins: precommitments and

obligations, total monthly payments, operating goods and services payments, and accrued obligations.

Table 3: Fiscal Pipeline Mechanism

Outcome	Beta	SE	p-value	First-stage F	N
Precommitments / obligations	0.8351	0.0690	0.0000	22.62	6811
Total monthly payments	0.0192	0.0561	0.7318	22.62	6811
Operating goods and services payments	-0.7210	0.1558	0.0000	22.62	6811
operating goods and services accrued obligations	0.1652	0.0790	0.0366	22.62	6811

Notes: The table reports IV estimates using the 2015–2019 hospital-month panel. The endogenous variable (Beta) is the execution wedge, defined as the log ratio of precommitments to payments. The instrument is MSP-level administrative disruption interacting with lagged hospital budget rigidity. All specifications include hospital fixed effects, calendar-month fixed effects, and year fixed effects. Standard errors are clustered at the hospital level. Operating goods and services correspond to MEF expenditure Group 53.

The results show that the execution wedge does not simply measure lower aggregate payments. When the wedge increases, precommitments and obligations rise sharply: the coefficient in the first row is positive and precisely estimated. At the same time, total monthly payments do not fall significantly. The coefficient on total payments is close to zero and statistically insignificant. This distinction is important. It suggests that the execution friction captured by the wedge is not an across-the-board reduction in cash payments to hospitals.

Instead, the adjustment is concentrated inside the operating-payment margin. Operating goods and services payments fall substantially, while operating goods and services accrued obligations rise. This pattern indicates that hospitals continue to generate and record obligations, but the conversion of those obligations into timely payments deteriorates. The fiscal pipeline therefore remains active on the obligation side, while the payment side becomes compressed for the operating inputs most directly connected to day-to-day service delivery.

This mechanism is central to the paper's interpretation. The execution wedge captures a disruption in the timing and composition of budget execution, not merely a decline in aggregate resources. Fiscal adjustment can therefore occur invisibly inside the execution pipeline: obligations accumulate, total payments do not necessarily contract contemporaneously, but operating payments fail to keep pace. Table 3 thus links the main throughput result to the paper's conceptual claim that cash-execution frictions are a policy-relevant margin through which fiscal stress reduces public-sector output.

The fiscal-pipeline evidence also generates a testable implication for patient outcomes. If the wedge operates through operating-payment compression, its consequences should be strongest for outcomes that depend on inputs used during the hospital stay. By contrast, outcomes determined primarily by severity at arrival should be less responsive. This distinction motivates the mortality analysis below. The mortality results are therefore not a separate welfare exercise; they are a downstream mechanism test of whether the fiscal shock operates through the hospital production function.

5.5 Mortality as Mechanism: Patient Consequences of Operating-Input Compression

Table 4 uses mortality outcomes to test whether the fiscal-pipeline mechanism has patient-level consequences. The evidence in Table 3 shows that the execution wedge operates through operating-payment compression: precommitments and obligations accumulate while payments for non-wage operating goods and services fall. This is precisely the margin that should affect the availability of drugs, medical supplies, pharmaceuticals, diagnostic consumables, and other inputs required to sustain inpatient care. Mortality therefore provides a mechanism test, not only a welfare calculation. If execution frictions impair care through operating-input shortages, the mortality response should be concentrated among patients whose outcomes depend on the quality and continuity of care after admission.

Table 4: Throughput and mortality responses

Outcome	Beta	SE	p-value	First-stage F	N	One-SD effect
Hospital discharges	-0.0613	0.0261	0.0189	22.62	6811	-19.86
Deaths, total	0.0485	0.0390	0.2134	22.62	6811	19.12
Deaths over 48h	0.0720	0.0363	0.0471	22.62	6811	29.69
Deaths under 48h	-0.0516	0.0358	0.1496	22.62	6811	-17.00
Mortality rate over 48h	0.2007	0.0822	0.0146	22.62	6811	0.72
Mortality rate under 48h placebo	-0.0352	0.0388	0.3643	22.62	6811	-0.13

Notes: The table reports IV estimates using the 2015–2019 hospital-month panel. The endogenous variable is the execution wedge, defined as the log ratio of precommitments to payments. The instrument is MSP-level administrative disruption interacted with lagged hospital budget rigidity. Death counts are measured in logs. Mortality rates are measured in levels and expressed as deaths per 100 discharges. All specifications include hospital fixed effects, calendar-month fixed effects, and year fixed effects. Standard errors are clustered at the hospital level. The one-standard-deviation effect is computed using the standard deviation of the execution wedge in each estimation sample.

The table is organized to separate the quantity margin from the patient-outcome margin. The discharge row captures the contraction in completed hospital output. The mortality rows ask whether the same fiscal shock changes outcomes among patients who are admitted. The distinction between deaths within and after 48 hours is central. Under-48-hour mortality is more likely to reflect severity at arrival or pre-hospital conditions. Over-48-hour mortality is more likely to reflect the hospital's ability to sustain treatment once the patient is inside the facility. This split therefore turns mortality into a diagnostic test of the operating-input channel.

To assess this, we estimate the effect of the instrumented execution wedge on inpatient deaths and mortality rates, distinguishing between deaths that occur within 48 hours of admission and deaths that occur after 48 hours. This distinction provides a within-table mechanism test. Deaths within 48 hours are more likely to reflect severity at arrival and pre-hospital conditions, and are therefore unlikely to be sensitive to the availability of operating inputs during the hospital stay. Deaths after 48 hours are more plausibly related to the

hospital's ability to sustain care over the course of the inpatient episode. If execution frictions harm patients through the operating-payment compression documented in Table 3, the mortality response should be concentrated in the over 48-hours category. The under-48-hours category serves as a placebo-style outcome: it captures deaths that are determined before operating-input constraints can plausibly affect patient trajectories.

The results are consistent with this prediction. Deaths over-48-hours increase significantly when the execution wedge widens: the coefficient is positive and statistically significant ($p = 0.047$), with a one-standard-deviation effect of approximately 29.7 percent. Deaths under-48-hours show no significant response ($p = 0.150$), with a negative point estimate consistent with the throughput reduction documented in Table 2. This asymmetry is informative. It is not consistent with a general mortality shock, which would raise both categories, nor with a pure throughput story, in which fewer admissions mechanically reduce death counts in both categories equally. It is consistent specifically with a deterioration in care quality during the admission along the operating-input margin that the fiscal-pipeline evidence identifies as the primary channel of adjustment.[10]

The mortality-rate specifications sharpen this interpretation by conditioning on the volume of hospital activity, addressing the concern that fewer admissions mechanically reduce death counts even if care quality is unchanged. The over-48-hours mortality rate, expressed as deaths per 100 discharges, increases significantly ($p = 0.015$). A one-standard-deviation increase in the execution wedge is associated with an increase of 0.72 deaths per 100 discharges in the over-48-hours mortality rate. Against a sample mean of approximately 1.23 deaths per 100 discharges, this corresponds to a roughly 59 percent increase in the conditional probability of inpatient death after 48 hours. The under-48-hours mortality-rate placebo is small and statistically insignificant ($p = 0.364$), with a one-standard-deviation effect of -0.13 deaths per 100 discharges, confirming that the over-48-hours result is not a general mortality shift but a specific deterioration in outcomes for patients whose care depends on sustained input availability during the admission.

Taken together, these results make mortality the patient-level counterpart of the fiscal-pipeline mechanism. Execution frictions reduce completed hospital output, but they also raise the conditional probability of inpatient death among patients whose outcomes depend on sustained input availability during the hospital stay. The over-48-hour mortality response, combined with the null under-48-hour placebo, is difficult to reconcile with a pure admissions or demand story. It is instead consistent with a deterioration in inpatient care quality generated by operating-input compression. The costs of this type of fiscal adjustment are therefore not limited to fewer services being delivered; they also appear in patient survival margins that are invisible in standard fiscal aggregates but measurable once the execution pipeline is linked to hospital records.

[10] Total deaths are positive but not statistically significant at conventional levels ($p = 0.214$), consistent with the positive over-48-hours and negative under-48-hours responses partially offsetting in the aggregate count.

5.6 Identification Validation

Table C.1 evaluates the main threats to the exclusion restriction. The baseline IV estimate of −0.0613 is the reference point for all exercises. The identifying assumption requires that, conditional on hospital, calendar-month, and year fixed effects, the instrument $Z_{it} = T_t \times R_{i,t-1}$ affects hospital discharges only through the execution wedge rather than through other contemporaneous determinants of hospital output. The validation exercises ask whether specific observable confounders, inherited fiscal conditions, administrative-capacity differences, or structural trends could independently account for the discharge effect. Full exposition of each exercise is provided in Appendix C.

Measurement and observable confounders. A first concern is that the estimate could be driven by a small number of hospital-months with unusually large wedge values arising from timing mismatches in administrative records rather than genuine execution stress. Winsorizing the wedge at the 1st and 99th percentiles leaves the coefficient virtually unchanged (−0.0624, row 2). Adding lagged demand controls — including lagged patient age and medical visits — does not attenuate the estimate; if anything, the coefficient becomes slightly more negative (−0.0746, row 3), confirming that the discharge effect is not explained by observable shifts in service pressure or patient composition. Adding lagged staffing controls yields a coefficient of −0.0578 (row 4), and adding lagged budget-composition controls yields −0.0646 (row 5), both statistically significant and close to the baseline. Observable differences in hospital demand, production capacity, and budget composition therefore do not account for the main result.

Inherited fiscal stress and joint confounders. A related concern is that the instrument could be selecting hospitals that entered the period with worse inherited fiscal positions. Adding a control for lagged December accrued obligations relative to the codified budget leaves the coefficient at −0.0643 (row 6). When all control blocks are included simultaneously — demand, staffing, budget-composition, and inherited fiscal-state controls — the coefficient remains −0.0681 with a first-stage F-statistic of 23.32 (row 7). Two reduced-form diagnostics confirm that the instrument does not systematically predict predetermined fiscal-state variables: the coefficients on lagged accrued obligations in levels and as a share of the codified budget are both small and statistically insignificant (Panel B, rows 10–11).

Differential trends and common shocks. The estimate could reflect differential time paths across hospital types or regions rather than execution frictions. Adding differential year effects by hospital care level, service class, and province leaves the coefficient at −0.0614 (row 8), essentially identical to the baseline. Replacing separate calendar-month and year fixed effects with full month-year fixed effects — absorbing any shock common to all hospitals in a given month — yields a coefficient of −0.0663 (row 9). Precision declines, as expected, since this specification removes substantially more time-series variation and restricts

identification to within-month, cross-hospital differences in pass-through. The point estimate's stability supports the interpretation that neither differential structural trends nor common monthly shocks drive the main result.

Timing and predetermined administrative capacity. Panel C reports timing-placebo diagnostics. Future values of the instrument do not predict current hospital discharges (rows 14–15), supporting the interpretation that the discharge result is not driven by smooth pre-trends or forward-looking dynamics in hospital performance. Future values do predict the current execution wedge (rows 12–13), consistent with short-run persistence in execution conditions and temporal clustering of administrative disruptions — a feature of the institutional environment rather than a violation of the exclusion restriction. Panel D tests whether the instrument predicts predetermined non-fiscal characteristics. Four of five coefficients are statistically insignificant and close to zero. The exception — lagged management and finance staff share — is economically small (less than two-tenths of one percentage point) and does not alter the second-stage coefficient when added as a control, confirming that this association does not account for the estimated throughput effect.

The OLS–IV gap and timing mismatch. Table C.2 examines the OLS–IV gap by comparing estimates across three sample partitions under full month-year fixed effects: the full sample, March–October (low administrative bunching), and November–February (high bunching). In the March–October subsample, where timing mismatch is least severe, the first-stage F-statistic rises to 54.55, the OLS estimate triples in magnitude, and the OLS–IV ratio falls to 1.46 — well within the range consistent with standard endogeneity rather than instrument invalidity. In the November–February subsample, where bunching is most severe, the OLS estimate is small and insignificant and the ratio reaches 43.59. This monotonic pattern is consistent with the timing-mismatch interpretation: the instrument predicts the true signal in the wedge but not the administrative noise, so OLS is attenuated in noisy months while IV is not. The convergence of OLS and IV in the clean months supports the structural interpretation of the main estimate. This table and its full exposition are reported in Appendix C.

Overall assessment. Across all exercises in Table C.1, the coefficient remains negative, statistically significant, and close to the baseline. No single observable channel — and no combination of them — materially shifts the estimate. The validation evidence therefore supports the interpretation that the discharge loss reflects execution frictions operating through the fiscal pipeline, rather than demand shocks, staffing changes, budget-composition differences, inherited fiscal stress, administrative-capacity heterogeneity, or differential structural trends.

Fiscal calendar predeterminedness test. The time-varying nature of $R_{i,t-1}$ raises an identification concern that standard shift-share designs address by requiring shares to be fixed: does the within-hospital evolution of lagged rigidity

reflect exogenous budget accumulation, or does it track hospital-specific performance trajectories that independently predict discharge outcomes? We address this by exploiting the mechanical within-year commitment-accumulation path driven by Ecuador's fiscal calendar — procurement cycles, payroll schedules, and contract renewal dates set administratively rather than by hospital conditions — to construct a purified instrument $Z_{it}^{*} = T_t \times \hat{R}_{i,t-1}$, where $\hat{R}_{i,t-1}$ is the calendar-predicted component of lagged rigidity. Panel E of Table C.1 reports the results. Replacing observed rigidity with its calendar-predicted component moves the main IV estimate from -0.0613 to -0.0559, a change of 8.9 percent and well within one standard error, while the first-stage F-statistic rises from 22.6 to 26.9 — consistent with the calendar instrument discarding idiosyncratic noise while preserving the identifying signal. Three formal predeterminedness tests reinforce this finding: lagged rigidity does not predict MSP disruption timing after fixed effects (row 22, $p = 0.201$), changes in the residual component of rigidity do not predict discharge outcomes (row 23, $p = 0.673$), and future rigidity does not predict current outcomes conditional on current rigidity (row 24, $p = 0.566$). Taken together, these results establish that the time-varying share component of the instrument satisfies a predeterminedness condition analogous to Bartik share exogeneity — tested rather than assumed — and that the endogenous component of lagged rigidity is not driving the main throughput result.

5.7 Interpreting the magnitudes.

Table 5 translates the preferred IV estimates into operational and welfare-relevant magnitudes. These calculations are intentionally back-of-the-envelope. They apply the estimated effects of a one-standard-deviation increase in the execution wedge to sample means from the estimation sample. The objective is not to provide a full welfare valuation, but to express the estimated effects in units that are meaningful for policy and operational decisions.

The magnitudes are economically large. A one-standard-deviation increase in the execution wedge reduces hospital discharges by approximately 19.9 percent. At the sample mean of about 355 discharges per hospital-month, this corresponds to roughly 70 fewer discharges per hospital-month, 846 fewer discharges per hospital-year, and approximately 101,500 foregone discharges per year across the 120 hospitals in the estimation sample. These figures should be interpreted as approximate translations of the IV estimate rather than precise accounting totals. They nevertheless show that disruptions in the conversion of obligations into payments are large enough to generate substantial losses in realized hospital output.

Table 5: Back-of-the-Envelope Magnitude Calculations

Statistic	Value
Standard deviation of execution wedge	3.609
Preferred IV estimate	-0.0613
One-SD discharge effect (%)	-19.86
Mean monthly hospital discharges	354.8
Implied monthly discharge loss	70.5
Implied annual discharge loss	845.7

Notes: The table reports back-of-the-envelope magnitude calculations based on the preferred IV estimate. The one-standard-deviation effect is computed using the standard deviation of the execution wedge in the IV estimation sample. Monthly and annual discharge losses translate this percentage effect using the sample mean of monthly hospital discharges. These calculations are intended to provide policy-relevant scale and should not be interpreted as a full welfare valuation.

The discharge losses are accompanied by direct welfare costs. The mortality rate results in Table 4 imply that a one-standard-deviation increase in the execution wedge is associated with approximately 2.55 additional inpatient deaths per hospital-month among patients admitted for more than 48 hours. This is the patient group whose outcomes are more likely to depend on care quality during the hospital stay rather than only on condition severity at arrival. Scaled across the 120 hospitals in the estimation sample, the point estimate corresponds to roughly 3,680 additional deaths per year. The 95 percent confidence interval ranges from approximately 725 to 6,630 additional system-level deaths per year. Both bounds are positive, so the uncertainty concerns the magnitude of the mortality cost rather than its direction. These calculations should be interpreted as scale translations of the preferred IV estimates under the maintained extrapolation assumption that the local mortality effect identified by the instrument applies proportionally across the 120-hospital estimation sample.[11]

[11] System-level mortality calculations proceed as follows. The IV estimate of the effect of the execution wedge on the over-48-hour mortality rate is $\hat{\beta}_m = 0.2007$, with standard error 0.0822 and $p = 0.0146$. The mortality rate is expressed as deaths per 100 discharges. Table 4 reports a one-standard-deviation effect of 0.72, implying 0.72 additional deaths per 100 discharges. At the sample mean of 354.8 discharges per hospital-month, this implies

$$0.72 \times \frac{354.8}{100} = 2.55$$

additional deaths per hospital-month, or $2.55 \times 12 = 30.65$ per hospital-year. Scaled across 120 hospitals, the system-level point estimate is $30.65 \times 120 \approx 3{,}679$ additional deaths per year. The 95 percent confidence interval is constructed by applying the same calculation to the lower and upper bounds of the mortality coefficient:

$$\hat{\beta}_m \pm 1.96 \times \hat{\sigma}_m = [0.2007 - 1.96 \times 0.0822,\ 0.2007 + 1.96 \times 0.0822] = [0.0396,\ 0.3618].$$

Using the standard deviation of the execution wedge implied by the reported one-standard-deviation mortality effect,

$$\sigma_E = \frac{0.72}{0.2007} \approx 3.59,$$

these coefficient bounds correspond to mortality-rate effects of approximately [0.142, 1.298] additional deaths per 100 discharges. Multiplying by the sample mean of 354.8 discharges per

The mortality translation is equally important because it is conditioned on completed hospital activity. The over-48-hour mortality rate increases by 0.72 deaths per 100 discharges after a one-standard-deviation increase in the execution wedge. Relative to a baseline of 1.23 deaths per 100 discharges, this represents an increase of approximately 59 percent in the conditional probability of dying after 48 hours. Unlike the discharge effect, this margin is not simply a reduction in service volume. It captures a deterioration in outcomes among patients who remain inside the hospital system and whose survival depends on the continuity of inpatient care.

Taken together, these translations show that execution frictions have consequences that are large in operational and welfare terms. A one-standard-deviation increase in the wedge implies a system-wide service contraction equivalent to the annual output of a mid-sized hospital, together with a measurable increase in over-48-hour inpatient mortality. These costs are invisible in approved budget aggregates but become observable once the execution pipeline is linked to hospital records. They also clarify the policy relevance of the estimates: improving cash management systems, strengthening commitment controls, and making payment timing more reliable could recover meaningful hospital output and reduce mortality risk without requiring additional formal budget allocations.

6 Discussion and Conclusions

This paper studies a form of fiscal adjustment that is rarely observed in standard public finance data: the separation between legally initiated spending obligations and actual cash execution. Using monthly administrative data from Ecuador's public health system, we show that this type of adjustment takes the form of an execution wedge between precommitments and payments, and that this wedge is not an accounting artifact. It is systematically predicted by administrative turbulence at the Ministry of Public Health, especially in hospitals with more rigid budget structures, and it translates into measurable losses in public service delivery. A one-standard-deviation increase in the execution wedge reduces hospital discharges by approximately 19.9 percent — equivalent to roughly 846 fewer discharges per hospital-year at the sample mean. It also raises the over-48-hour inpatient mortality rate by approximately 59 percent,

hospital-month, by 12 months, and by 120 hospitals yields system-level bounds of approximately [725, 6,632] additional deaths per year.

The discharge-loss calculation follows analogously. Table 4 reports a one-standard-deviation effect of -19.86 percent for hospital discharges. At the sample mean of 354.8 discharges per hospital-month, this corresponds to

$$0.1986 \times 354.8 = 70.5$$

fewer discharges per hospital-month, or $70.5 \times 12 = 845.8$ fewer discharges per hospital-year. Scaled across the 120-hospital network, this implies approximately $845.8 \times 120 \approx 101{,}500$ foregone discharges annually. All calculations apply the preferred IV estimates from Tables 2 and 4 to sample means from the main estimation sample. Subscripts d and m denote the discharge and mortality specifications, respectively.

while leaving the under-48-hour mortality placebo statistically unchanged. This timing pattern is central to the interpretation of the results. It shows that fiscal adjustment through managing the timing of cash disbursements does not merely reduce the quantity of hospital output; it also weakens the quality and continuity of inpatient care for patients whose outcomes depend on resources available during the admission. The impacts arise not from cuts to approved budgets but from frictions inside the execution pipeline that are invisible to standard fiscal aggregates.

The mechanism evidence clarifies why execution frictions are operationally binding rather than merely accounting noise. Under Ecuador's Treasury Single Account regime, delays in settling legally accrued obligations weaken the credibility of future payments to suppliers. As suppliers anticipate payment risk, they become less willing to participate in new procurement processes or to deliver operating inputs under normal conditions. The wedge, therefore, constrains the immediate future availability of hospital inputs by weakening supplier participation and disrupting procurement continuity, rather than reflecting a simple contemporaneous absence of budgeted resources. Consistent with this interpretation, when the execution wedge widens, precommitments and accrued obligations continue to accumulate, while operating payments—the cash flows most directly tied to day-to-day service delivery—fall sharply. Payroll, being the most institutionally protected expenditure category, adjusts less than non-wage operating inputs in the short run. The burden of adjustment therefore falls disproportionately on non-wage operating margins, compressing the inputs that hospitals need to convert authorized budgets into realized care. This is not a story about resource scarcity in the aggregate; it is a story about the institutional machinery through which authorized resources reach the point of production.

The mortality evidence shows that this compression reaches the patient level. The rise in over-48-hour mortality is consistent with a failure to sustain care once patients are admitted, while the absence of an under-48-hour mortality response helps rule out the interpretation that the results are driven by changes in severity at arrival. Mortality therefore provides the clinical counterpart to the fiscal mechanism: the payment wedge first appears as operating-payment compression, then as reduced throughput, and finally as worse outcomes for patients whose treatment requires sustained hospital inputs.

These findings reframe a question at the intersection of public financial management and state capacity. A large literature documents that state capacity shapes the quality of public goods provision, emphasizing management practices, bureaucratic quality, and the fidelity of policy implementation. Our results suggest that fiscal institutions — specifically, the architecture governing commitment controls, cash management, and payment authorization — constitute an underappreciated dimension of this capacity. The relevant input into public service delivery is not what has been allocated on paper or even legally precommitted, but what has been executed reliably and on time. When administrative and financial execution operate on different timelines, approved budgets and legally incurred obligations are not measures of effective resources available

for production. They are claims on future cash. Treating them as productive capacity leads standard fiscal aggregates to mismeasure both the incidence of adjustment and the real operating capacity of the state.

This reframing has direct implications for reform. The conventional toolkit for improving public service delivery focuses on increasing budget allocations, improving provider incentives, or strengthening management at the facility level. Our results suggest a complementary and largely overlooked margin: the timely and reliable conversion of spending commitments into operating payments. Improving cash management systems, tightening commitment controls, and reducing the administrative processing delays that generate execution wedges could recover a nontrivial share of hospital throughput and reduce avoidable mortality risk without requiring additional formal budget allocations. For governments and development institutions operating under fiscal constraints — precisely the conditions under which execution frictions tend to worsen — this margin deserves considerably more attention than it currently receives. Two limitations bound the scope of these conclusions. First, the IV estimate is local to the execution frictions generated when MSP-level administrative disruptions interact with heterogeneous hospital budget rigidity. Whether the magnitude generalizes to other sources of execution delay, or to institutional settings where hospitals retain greater operational liquidity, is an open question. Second, the welfare cost of the discharge losses we document depends on the clinical composition of affected cases — whether frictions suppress elective procedures or constrain acute care — a distinction our data do not resolve. These are natural extensions. What the evidence suggests is that payment execution timing is a first-order determinant of public sector productivity, and that reforms to the administrative machinery of budget execution can generate meaningful service-delivery gains even in the absence of new fiscal resources.

References

Ahenkan, A. and Azaare, J. (2018). Managerial implications of delayed reimbursement of national health insurance claims: The case of two hospitals in northern Ghana. *Journal of Management and Research*, 5(2):1–29.

Asante, A. D., Zwi, A. B., and Ho, M. T. (2006). Getting by on credit: How district health managers in Ghana cope with the untimely release of funds. *BMC Health Services Research*, 6:105.

Auerbach, A. J. and Gorodnichenko, Y. (2012). Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, 4(2):1–27.

Baffes, J., Kose, M. A., Ohnsorge, F., and Stocker, M. (2015). The great plunge in oil prices: Causes, consequences, and policy responses. Policy Research Note PRN/15/01, World Bank, Washington, DC.

Bandiera, O., Best, M. C., Khan, A. Q., and Prat, A. (2021). The alloca-

tion of authority in organizations: A field experiment with bureaucrats. *The Quarterly Journal of Economics*, 136(4):2195–2242.

Besley, T. and Persson, T. (2009). The origins of state capacity: Property rights, taxation, and politics. *American Economic Review*, 99(4):1218–1244.

Céspedes, L. F. and Velasco, A. (2014). Was this time different?: Fiscal policy in commodity republics. *Journal of Development Economics*, 106:92–106.

Checherita-Westphal, C., Klemm, A., and Viefers, P. (2016). Governments' payment discipline: The macroeconomic impact of public payment delays and arrears. *Journal of Macroeconomics*, 47(Part B):147–165.

Dahis, R., Nascimento, L., et al. (2026). Fiscal capacity and execution at the local level: New evidence from Brazil. *The World Bank Economic Review*. Forthcoming.

Finan, F., Olken, B. A., and Pande, R. (2017). The personnel economics of the developing state. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments, Volume 2*, pages 467–514. Elsevier.

Flynn, S. and Pessoa, M. (2014a). Prevention and management of government arrears. Technical Notes and Manuals TNM/14/03, International Monetary Fund, Fiscal Affairs Department, Washington, DC.

Flynn, S. and Pessoa, M. (2014b). Prevention and management of government arrears. Technical Notes and Manuals TNM/14/03, International Monetary Fund, Fiscal Affairs Department, Washington, DC.

Frankel, J. A., Végh, C. A., and Vuletin, G. (2013). On graduation from fiscal procyclicality. *Journal of Development Economics*, 100(1):32–47.

Fritz, V., Sweet, S., and Verhoeven, M. (2014). Strengthening public financial management exploring drivers and effects.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110:2586–2624.

Gruss, B. (2014). After the boom–commodity prices and economic growth in Latin America and the Caribbean. IMF Working Paper WP/14/154, International Monetary Fund, Washington, DC.

Ilzetzki, E., Mendoza, E. G., and Végh, C. A. (2013). How big (small?) are fiscal multipliers? *Journal of Monetary Economics*, 60(2):239–254.

International Monetary Fund (2019). Ecuador: Staff report for the 2019 Article IV consultation and request for an extended arrangement under the extended fund facility. IMF Country Report No. 19/79, International Monetary Fund, Western Hemisphere Department, Washington, DC.

Jordán, F. (2025). Prompt payment enforcement on framework agreements for public hospitals: Evidence from Chile. *Fiscal Studies*. Online first.

López, R., Peñaranda, D., Prat, J., and Zanoni, W. (2025). ¿cómo se presupuesta en salud y educación en la Región Andina? IDB Monograph IDB-MG-1313, Inter-American Development Bank, Andean Group Country Department, Washington, DC.

López, R., San Roman Vucetich, C., and Zanoni, W. (2026). ¿cómo se presupuesta la salud y la educación en Ecuador? Technical report, Inter-American Development Bank, Washington, DC.

Moses, M. W., Korir, J., Zeng, W., Musiega, A., Oyasi, J., Lu, R., Chuma, J., and Di Giorgio, L. (2021). Performance assessment of the county healthcare systems in Kenya: A mixed-methods analysis. *BMJ Global Health*, 6(6):e004707.

Musiega, A., Tsofa, B., Nyawira, L., Njuguna, R. G., Munywoki, J., Hanson, K., Mulwa, A., Molyneux, S., Maina, I., Normand, C., Jemutai, J., and Barasa, E. (2023). Examining the influence of budget execution processes on the efficiency of county health systems in Kenya. *Health Policy and Planning*, 38(3):351–362.

Pattanayak, S. (2016). Expenditure control: Key features, stages, and actors. Technical Notes and Manuals 16/02, International Monetary Fund.

Pattanayak, S. and Fainboim, I. (2011). Treasury single account: An essential tool for government cash management. Technical Notes and Manuals TNM/11/04, International Monetary Fund, Fiscal Affairs Department, Washington, DC.

PEFA Secretariat (2016). Public expenditure and financial accountability (pefa) framework.

Piatti-Fünfkirchen, M. and Schneider, P. (2018). From stumbling block to enabler: The role of public financial management in health service delivery in Tanzania and Zambia. *Health Systems & Reform*, 4(4):336–345.

Ramey, V. A. (2011). Can government purchases stimulate the economy? *Journal of Economic Literature*, 49(3):673–685.

Rasul, I. and Rogger, D. (2018). Management of bureaucrats and public service delivery: Evidence from the Nigerian civil service. *The Economic Journal*, 128(608):413–446.

Vieira, F. S. and Santos, M. A. B. d. (2018). Contingenciamento do pagamento de despesas e restos a pagar no orçamento federal do SUS. *Revista de Administração Pública*, 52(4):731–739.

A Panel Construction and Sample Definition

This appendix documents the construction of the hospital-month panel used in the main analysis. We harmonize hospital identifiers and monthly dates across budget-execution records, personnel and administrative leadership records, and hospital activity records. The final unit of observation is a hospital-month. The main discharge estimation sample is defined by the availability of the execution wedge, the instrument, log hospital discharges, and the fixed-effect identifiers required by the empirical specification.

Table A.1 reports the sample-construction flow. The final discharge estimation sample contains 6,811 hospital-month observations from 120 hospitals. Table A.2 reports annual coverage over the 2015–2019 period. The panel is unbalanced because not all hospitals have complete fiscal and activity records in every month.

Table A.1: Construction of the Main Estimation Sample

Step	Hospital-months	Hospitals
All observed hospital-months, 2015–2019	7,260	121
Nonmissing execution wedge	6,963	120
Nonmissing execution wedge and instrument	6,811	120
Nonmissing wedge, instrument, and log discharges	6,811	120
Nonmissing fixed-effect identifiers	6,811	120
Final discharge estimation sample	6,811	120

Notes: The table reports the construction of the main discharge estimation sample. The final sample is defined by the availability of the execution wedge, the instrument, log discharges, and the fixed-effect identifiers required by the empirical specification.

Table A.2: Panel Coverage by Year

Year	Hospitals	Hospital-months	Coverage rate (%)
2015	113	1,226	90.4
2016	115	1,358	98.4
2017	118	1,402	99.0
2018	119	1,411	98.8
2019	118	1,414	99.9

Notes: The table reports annual coverage in the main discharge estimation sample. The sample is restricted to hospital-month observations with nonmissing execution wedge, instrument, log discharges, and fixed-effect identifiers. The coverage rate is computed as observed hospital-months divided by the maximum possible hospital-months among hospitals observed in each year.

B Descriptive Evidence: 2SLS First Stage

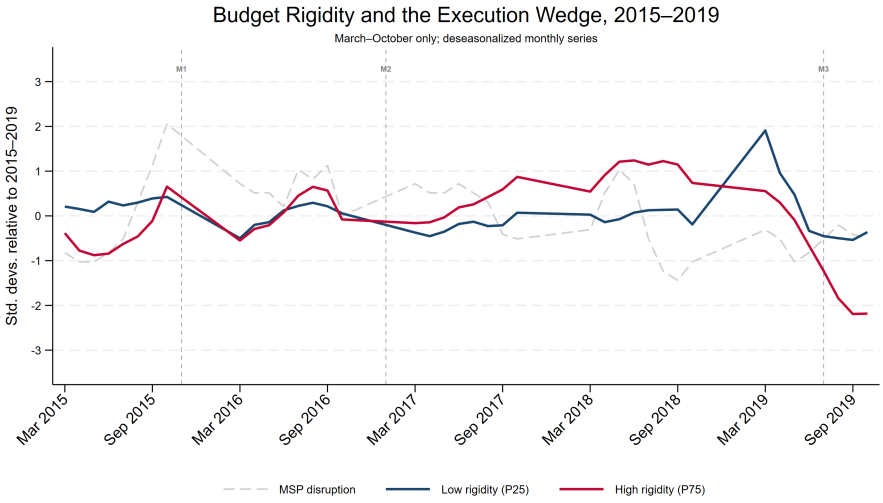

Figure B.1: Execution Wedge and MSP Disruption, 2015–2019

Notes: The figure replicates Figure 1 using only March–October observations, excluding the year-end and year-start budget-execution period. It plots deseasonalized monthly series for MSP disruption and the execution wedge among hospitals with low and high lagged budget rigidity. MSP disruption is T_t, the monthly count of leadership changes in key MSP planning and administrative-management positions. The execution wedge is E_{it}, defined as the log ratio of precommitments to payments. Low- and high-rigidity series correspond to hospitals in the lower and upper tails of the lagged budget-rigidity distribution, $R_{i,t-1}$. All series are residualized with respect to month-of-year fixed effects, standardized using the 2015–2019 distribution, and smoothed using a three-month moving average. Vertical lines denote MSP ministerial transitions.

B.1 Does the Effect of the Execution Wedge Vary with Budget Rigidity?

The preferred IV specification uses lagged budget rigidity to identify variation in the execution wedge: common MSP-level administrative disruptions should generate larger execution frictions in hospitals whose pre-existing budget structures leave fewer margins for adjustment. This raises a related question: does rigidity only affect the pass-through of administrative disruption into the wedge, or does it also make hospitals more sensitive to the wedge once the wedge materializes?

To examine this, we estimate an interaction-IV specification that allows the effect of the execution wedge to vary with standardized lagged budget rigidity. Specifically, we estimate

$$Y_{it} = \beta E_{it} + \gamma \left(E_{it} \times \widetilde{R}_{i,t-1} \right) + \alpha_i + \lambda_m + \tau_y + \varepsilon_{it},$$

where Y_{it} is log hospital discharges, E_{it} is the execution wedge, and $\widetilde{R}_{i,t-1}$ is lagged budget rigidity standardized to have mean zero and standard deviation one in the estimation sample. Because both E_{it} and $E_{it} \times \widetilde{R}_{i,t-1}$ are endogenous, we instrument them with Z_{it} and $Z_{it} \times \widetilde{R}_{i,t-1}$, respectively. The specification includes hospital fixed effects, calendar-month fixed effects, and year fixed effects, and standard errors are clustered at the hospital level.

Table B.1: Interaction-IV Estimates by Lagged Budget Rigidity

Parameter	Estimate	Standard error	p-value
Execution wedge at mean rigidity	-0.0606	0.0254	0.017
Wedge × standardized lagged rigidity	0.0025	0.0031	0.429
Marginal effect at $R_{i,t-1} = -1$ SD	-0.0631	0.0277	0.023
Marginal effect at $R_{i,t-1} = +1$ SD	-0.0581	0.0234	0.013
Difference: high minus low rigidity	0.0050	0.0063	0.429
Observations	6,811		

Notes: The table reports an interaction-IV specification in which both the execution wedge and its interaction with standardized lagged budget rigidity are treated as endogenous. The excluded instruments are the baseline instrument, Z_{it}, and its interaction with standardized lagged rigidity. All specifications include hospital fixed effects, calendar-month fixed effects, and year fixed effects. Standard errors are clustered at the hospital level. The marginal effects evaluate the effect of the execution wedge at one standard deviation below and above mean lagged rigidity.

Appendix Table B.1 shows that the interaction between the execution wedge and standardized lagged rigidity is small and statistically insignificant. The estimated marginal effects are also very similar across the rigidity distribution. At one standard deviation below mean rigidity, the marginal effect of the execution wedge is -0.063; at one standard deviation above mean rigidity, it is -0.058. The difference between these marginal effects is small and statistically indistinguishable from zero.

This result clarifies the role of budget rigidity in the empirical design. Rigidity operates primarily through the first stage: it determines how strongly common upstream administrative disruption is converted into a hospital-level execution wedge. Once the wedge materializes, however, the estimated effect of that wedge on hospital throughput is similar across more and less rigid hospitals. This is consistent with the institutional interpretation that rigidity identifies exposure to execution frictions, while the wedge itself carries the service-delivery effect.

C Empirical Results Validation

A useful way to interpret the exclusion restriction is that the execution wedge is not a narrow measure of a single administrative friction. It is the downstream

fiscal expression of frictions that delay the conversion of operational needs into paid inputs. In Ecuador's institutional setting, hospitals cannot self-finance under the Treasury Single Account regime, do not control the timing of cash payments, and must acquire budget-financed inputs through the commitment and payment pipeline. Therefore, upstream disruptions in procurement processing, request routing, administrative sequencing, or supplier-payment management are not independent channels once they become operationally binding: they are precisely the types of frictions that should appear as a widening gap between obligations and payments. The exclusion restriction would be violated if the instrument affected hospital output through non-fiscal channels that do not operate through the execution pipeline, such as differential patient demand, staffing shocks, clinical-capacity changes, or pre-existing hospital trends. The following validation exercises are therefore organized around the remaining threats: factors that could affect hospital output independently of the execution pipeline.

Table C.1 evaluates the main threats to the exclusion restriction. The baseline estimate, reported in the first row, is the reference point for the validation exercise. It reproduces the preferred IV estimate (row (1)) from Table 2: a one-unit increase in the instrumented execution wedge reduces log hospital discharges by 0.0613, with a p-value of 0.0189 and a first-stage F-statistic of 22.62. The remaining rows ask whether this estimate is materially altered when we address specific alternative explanations for the discharge effect.

Table C.1: Robustness of the IV Estimate to Identification Threats

	Specification	Estimate	p-value	First-stage F	N	Change from baseline (%)
Panel A. Robustness of the IV estimate						
(1)	Baseline IV	-0.0613	0.0189	22.62	6811	0.00
(2)	Winsorized wedge	-0.0624	0.0187	24.73	6811	-1.80
(3)	Demand controls	-0.0746	0.0099	21.46	6811	-21.66
(4)	Staffing controls	-0.0578	0.0305	24.11	6811	5.83
(5)	Budget-composition controls	-0.0646	0.0105	23.44	6811	-5.27
(6)	Inherited fiscal controls	-0.0643	0.0149	23.25	6811	-4.86
(7)	All controls	-0.0681	0.0121	23.32	6811	-11.05
(8)	Structural trends	-0.0614	0.0141	23.77	6811	-0.14
(9)	Full month-year fixed effects	-0.0663	0.0840	16.52	6811	-8.01
Panel B. Predetermined fiscal-state diagnostics						
(10)	Lagged accrued obligations	-0.0009	0.2708	–	5503	–
(11)	Lagged accrued-obligation share	0.0003	0.1474	–	5492	–
Panel C. Timing-placebo diagnostics						
(12)	Current wedge: Lead 1 instrument	0.2526	0.0000	–	6684	–
(13)	Current wedge: Lead 2 instrument	0.1625	0.0002	–	6557	–
(14)	Current log discharges: Lead 1 instrument	-0.0063	0.2070	–	6684	–
(15)	Current log discharges: Lead 2 instrument	-0.0047	0.3127	–	6557	–
Panel D. Predetermined non-fiscal operational and administrative-capacity diagnostics						
(16)	Lagged patient age	-0.0435	0.4979	–	6811	–
(17)	Lagged doctor share	-0.0004	0.5839	–	6681	–
(18)	Lagged management/finance staff share	-0.0016	0.0072	–	6682	–
(19)	Lagged human-resources staff share	-0.0001	0.2246	–	6682	–
(20)	Lagged planning staff share	0.0001	0.5442	–	6683	–
Panel E. Fiscal calendar predeterminedness test						
(21)	Purified calendar IV ($Z^*_{it} = T_t \times \hat{R}_{i,t-1}$)	-0.0559	0.0213	26.86	6811	-8.94
(22)	$R_{i,t-1}$ predicts T_t (share predicts shock)	-0.2539	0.2010	–	6811	–
(23)	$\Delta(R_{i,t-1} - \hat{R}_{i,t-1})$ predicts log discharges	-0.1299	0.6730	–	6811	–
(24)	$R_{i,t}$ predicts $Y_{it} \mid R_{i,t-1}$ (forward rigidity)	0.2165	0.5660	–	6811	–

Notes: The table reports IV estimates for log hospital discharges using the 2015–2019 hospital-month panel. The endogenous variable is the execution wedge, defined as the log ratio of precommitments to payments. The instrument is MSP-level administrative disruption interacted with lagged hospital budget rigidity. All specifications include hospital fixed effects, calendar-month fixed effects, and year fixed effects. Standard errors are clustered at the hospital level. The change column reports the percentage change in the coefficient relative to the baseline IV estimate. Predetermined fiscal diagnostics are reported as reduced-form checks of whether the instrument predicts inherited fiscal-state variables.

Extreme values in the execution wedge. The first concern is that the estimated effect could be driven by a small number of hospital-months with unusually large values of the execution wedge. This concern arises from the way the wedge is constructed. Because E_{it} is the log ratio of precommitments to payments, observations with very low payments relative to precommitments can generate large wedge values. In administrative fiscal data, such observations may reflect genuine execution stress, but they may also reflect timing mismatches, accounting delays, or unusually lumpy payment records. If the main IV estimate were driven by these extreme observations, the result would be less informative about the typical relationship between execution frictions and hospital output.

To evaluate this possibility, Table C.1 re-estimates the IV model after winsorizing the endogenous execution wedge at the 1st and 99th percentiles (see row (2)). This procedure limits the influence of the most extreme wedge observations while preserving the structure of the estimating sample and the same instrument. The coefficient remains virtually unchanged, moving from -0.0613 in

the baseline specification to −0.0624 after winsorization, with a similar p-value and a first-stage F-statistic of 24.73. The main discharge result is therefore not driven by a small number of unusually high or low execution-wedge observations.

Demand and patient-composition shocks. A second concern is that the estimated discharge effect could reflect changes in patient demand rather than execution frictions. This would be problematic if months in which the instrument is high also coincide with changes in the number or composition of patients seeking care. For example, hospitals exposed to larger instrument-induced wedges could also face changes in recent service pressure, patient age composition, or first-level demand conditions that independently affect the number of hospital discharges. In that case, the IV estimate would partly capture demand-side variation rather than the effect of budget-execution frictions on hospital output.

To evaluate this concern, row (3) of Table C.1 adds lagged demand controls to the baseline IV specification. These controls include lagged patient age and lagged medical visits, together with the missingness and tail-adjustment indicators generated by the diagnostic protocol. Using lagged controls is important because they capture pre-existing demand conditions without conditioning directly on contemporaneous outcomes that may themselves respond to execution frictions. The coefficient becomes more negative, moving from −0.0613 to −0.0746, and remains significant. In addition, there are no statistically significant differences between this coefficient estimate of the wedge effect and the baseline estimate in row (1). The estimate does not attenuate toward zero when observable lagged demand conditions are included, supporting the interpretation that the discharge effect is not explained by observable demand or patient-composition shocks.

Staffing and hospital production capacity. An additional concern is that the estimated discharge effect could reflect changes in hospital staffing rather than budget-execution frictions. This would be problematic if the instrument were correlated with short-run changes in doctors, nurses, or administrative personnel. In that case, lower discharges could arise because of changes in the inputs entering the production function of hospital performance, not because the execution wedge constrained their ability to convert budgeted obligations into timely operating capacity.

To evaluate this possibility empirically, in row (4) of Table C.1 we added lagged staffing controls to the baseline IV specification. These controls include lagged numbers of doctors, nurses, and administrative personnel, together with the missingness and tail-adjustment indicators generated by the diagnostic protocol. As with the demand controls, the use of lagged staffing variables is intended to capture pre-existing observable capacity without conditioning on contemporaneous inputs that may themselves respond to execution frictions. The coefficient remains negative and statistically significant, at −0.0578, very close to the baseline estimate. This indicates that the main discharge effect

is not explained away by observable changes in hospital staffing or measured production capacity.

Budget scale and expenditure composition. A related concern is that the instrument could be capturing broader differences in hospitals' budget scale and expenditure composition rather than the transmission of administrative disruption into execution frictions. This concern is especially relevant because lagged budget rigidity enters directly into the instrument. Hospitals with more rigid budgets may also be larger, have a higher share of personnel obligations, rely less on flexible operating lines, or allocate a different share of their codified budget to operating goods and services. These differences could be correlated with discharge capacity even in the absence of execution frictions.

To evaluate this possibility, Table C.1 adds lagged budget-composition controls to the baseline IV specification (see row (5)). These include lagged total codified budget and lagged expenditure-composition measures, including personnel and operating-goods-and-services shares, together with the corresponding missingness and tail-adjustment indicators from the diagnostic protocol. The coefficient remains close to the baseline, at -0.0646, and statistically significant. This suggests that the estimated discharge effect is not simply comparing hospitals with different budget size or expenditure composition; it persists after accounting for observable differences in the scale and allocation of hospital budgets.

Pre-existing fiscal obligations. When hospitals enter the year with larger unpaid or accrued obligations, they may already face tighter operating conditions before the current execution shock occurs. This creates a concern that the instrument could be picking up pre-existing fiscal distress rather than current execution frictions: hospitals with larger inherited obligations might both experience larger execution wedges and produce fewer discharges, independently of the mechanism we seek to identify.

To rule out this possibility, row (6) of Table C.1 presents results from a specification that adds a control for lagged December accrued obligations relative to the hospital's codified budget. This measure captures the fiscal burden that each hospital carried into the current year, scaled by its budget size. The coefficient remains negative and statistically significant, at -0.0643, and remains close to the baseline estimate. This suggests that the estimated effect of current execution frictions is not merely proxying for fiscal stress inherited from the previous year.

Could multiple observable channels jointly explain the result? One might also question identification under the premise that no single observable channel explains the result, but several of them jointly do. For example, the instrument could be correlated with a combination of lagged demand conditions, staffing levels, budget composition, and inherited fiscal stress, even if none of these channels individually accounts for the main estimate.

We evaluate that possibility in row (7) of Table C.1 by estimating a joint-controls specification that includes all validated control blocks simultaneously. This specification adds the demand, staffing, budget-composition, and inherited-fiscal-state controls in the same IV regression. The coefficient remains negative and statistically significant, at -0.0681, with a first-stage F-statistic of 23.32. This is the most conservative observable-control specification in the table. Its stability indicates that the main discharge effect survives simultaneous adjustment for the principal observed threats to identification.

Could differential trends across hospital groups or regions explain the result? Finally, one last validation exercise evaluates whether the estimate could be driven by differential time paths across types of hospitals or regions rather than by execution frictions. Hospital fixed effects absorb permanent differences across facilities, but they do not rule out the possibility that hospitals of different care levels, service classes, or provinces were following different trajectories during the 2015–2019 period. If those trajectories coincided with the instrument, the baseline estimate could partly reflect structural differences in hospital trends rather than the causal effect of the execution wedge.

One way to address that possibility is to allow different groups of hospitals to have their own time paths. Table C.1 does this in row (8) by adding differential year effects by hospital care level, service class, and province. The coefficient remains almost identical to the baseline, at -0.0614, and remains statistically significant. This supports the interpretation that the result is not driven by differential structural trends across hospital groups or regions.

Absorbing common monthly shocks. A more demanding specification replaces the separate calendar-month and year fixed effects with full month-year fixed effects in row (9) of Table C.1. This absorbs any shock common to all hospitals in a given month, including aggregate fiscal conditions, health-system seasonality, and contemporaneous national policy changes. The identifying variation is therefore restricted to cross-hospital differences in the pass-through of the same monthly environment, operating through lagged budget rigidity. The coefficient remains very close to the baseline estimate, indicating that the main result is not driven by common monthly shocks. Precision declines, as expected, because the specification removes substantially more time-series variation and leaves only within-month, cross-hospital variation to identify the effect.

Thus, when all aggregate monthly shocks are absorbed nonparametrically, the estimated effect remains economically similar to the baseline — the coefficient moves from -0.0613 to -0.0663, a negligible change, though precision falls, reflecting the mechanical power cost of removing all time-series variation in the shock. Absorbing month-year fixed effects eliminates the time-series variation in T_t, leaving only cross-hospital differences in pass-through within each month to identify the effect. Given the expected precision loss, the point estimate's stability under this specification supports the interpretation that common monthly shocks do not drive the main result.

Timing mismatch and the OLS-IV gap. The gap between the OLS and IV estimates in Table 2 is large in the full sample. Three factors contribute to this gap, as discussed in Section 5.3: endogenous execution activity, timing mismatches in high-frequency fiscal records, and the IV isolating a more binding component of execution-friction variation.[12] To assess the quantitative contribution of timing mismatch specifically (i.e. the recording of obligations and payments in different calendar months due to administrative processing lags rather than genuine cash delays), we re-estimate both specifications using full month-year fixed effects and compare results across three sample partitions: the full sample, the March–October subsample that excludes the months most susceptible to year-end and year-start administrative bunching, and the November–February subsample that isolates those high-bunching months. Both OLS and IV use identical fixed-effect structures across all three columns, so the only source of variation across rows is the sample restriction.

Table C.2: OLS-IV Gap: Timing Mismatch Diagnostic

Sample	OLS	IV	OLS p	IV p	First-stage F	Abs. gap	Ratio
Full sample	−0.0052	−0.0663	0.0038	0.0840	16.52	0.0611	12.81
March–October	−0.0169	−0.0247	0.0012	0.2048	54.55	0.0078	1.46
November–February	−0.0023	−0.1005	0.1425	0.1210	5.93	0.0982	43.59

Notes: The table reports OLS and IV estimates of the effect of the execution wedge on log hospital discharges across three sample partitions. All specifications include hospital and full month-year fixed effects. The instrument is MSP-level administrative disruption interacted with lagged hospital budget rigidity. Standard errors are clustered at the hospital level. The absolute gap is the difference between the IV and OLS point estimates in absolute value. The ratio is the IV estimate divided by the OLS estimate in absolute value. The March–October sample excludes November through February to remove months most susceptible to year-end and year-start administrative bunching in fiscal records. The November–February sample isolates those high-bunching months.

The results reveal a monotonic pattern that is consistent with the timing-mismatch interpretation. In the November–February subsample, where administrative bunching is most severe, the OLS estimate is small and insignificant, the instrument is weak (first-stage $F = 5.93$), and the OLS-IV ratio reaches 43.59. In the full sample, which averages over bunching and non-bunching months, the ratio falls to 12.81. In the March–October subsample, where timing mismatch is least severe, the instrument is strong (first-stage $F = 54.55$), both OLS and IV are significant, and the ratio falls to 1.46 — well within the range consistent with standard endogeneity concerns rather than with instrument invalidity.

The convergence of OLS and IV in the clean months is the central finding of this diagnostic. When the execution wedge is measured with less administrative

[12] Timing mismatch in this context refers to the fact that in high-frequency administrative fiscal systems like eSIGEF, the recording of a spending obligation and the recording of its corresponding payment do not always fall in the same calendar month. An obligation incurred in December may be paid and recorded in January, or a payment recorded in March may settle an obligation that was logged in February.

noise, the OLS estimate triples in magnitude relative to the full sample, and the IV estimate falls toward the OLS, closing most of the gap. This pattern arises because timing mismatch generates noise in the wedge that is orthogonal to the instrument by construction: the instrument predicts the true signal in the wedge but not the noise, so OLS is attenuated by noise while IV is not, mechanically inflating the full-sample ratio. Removing the noisiest months reduces attenuation in OLS and reduces the inflation of the IV estimate simultaneously, producing convergence. The residual gap in the March–October sample — OLS of -0.0169 against IV of -0.0247 — is small and consistent with residual endogeneity in the observed wedge rather than with a structural disconnect between the two estimators. The convergence argument relies on the point estimates rather than on a precisely estimated IV in that subsample; the March–October IV is not significant at conventional levels, reflecting the reduced sample size and the demanding month-year fixed-effect structure rather than instrument weakness, as confirmed by the strong first-stage F of 54.55.[13]

Does the instrument select hospitals with inherited fiscal stress? The last two rows ask a slightly different question from the preceding robustness exercises. Instead of re-estimating the second-stage effect of the wedge on discharges, they test whether the instrument itself predicts fiscal conditions that were already in place before the current execution shock. This matters because the instrument would be less credible if high values of Z_{it} systematically occurred in hospitals that had entered the period with worse inherited fiscal positions. In that case, the estimated discharge effect could partly reflect pre-existing fiscal stress rather than current execution frictions.

To evaluate this possibility, Table C.1 reports reduced-form diagnostics in which the instrument is used to predict two predetermined fiscal-state variables: lagged accrued obligations as a share of the codified budget, and lagged accrued obligations in levels. These variables capture whether hospitals entered the period with larger unpaid or accrued obligations, either relative to their budget size or in absolute terms. The estimates are small and statistically insignificant in both cases. This provides additional reassurance that the instrument is not systematically selecting hospitals with worse inherited fiscal conditions. The main evidence on this issue comes from the inherited-fiscal-scale and joint-control specifications above; these reduced-form diagnostics complement that evidence by showing that the instrument does not directly predict the predetermined fiscal-stress measures themselves.

Overall validation evidence. Taken together, Table C.1 shows that the main discharge effect is stable across the principal threats to identification. The estimate is not driven by extreme values of the execution wedge, observable

[13] The IV estimate in the March–October sample is not significant at conventional levels ($p = 0.205$) despite a strong first stage ($F = 54.55$), reflecting the reduced sample size and the demanding month-year fixed-effect structure rather than instrument weakness. The point estimate's convergence with OLS is the relevant finding rather than its individual significance.

demand, or patient-composition differences, staffing changes, budget scale, or expenditure composition, inherited fiscal stress, or differential time paths across hospital groups and regions. Across these exercises, the coefficient remains negative, statistically significant, and close to the baseline estimate.

The validation ladder therefore supports the interpretation that the decline in hospital discharges reflects execution frictions rather than the main observable alternatives to the exclusion restriction. The exercise does not attempt to rule out every upstream administrative event; many such events are part of the execution-pipeline mechanism once they become operationally binding. Instead, the tests ask whether the instrument is correlated with factors that could affect hospital output independently of the execution pipeline. While no validation exercise can prove the exclusion restriction directly, the stability of the estimate across these targeted tests makes it less likely that the result is explained by outliers, observable confounders, inherited fiscal conditions, non-pipeline administrative-capacity differences, or structural trends unrelated to the execution wedge.

Do future values of the instrument predict current outcomes? A final timing-placebo exercise evaluates whether the instrument is proxying for broader latent trends rather than contemporaneous execution shocks. If future realizations of the instrument predicted current hospital output, this would raise concerns that the identifying variation reflects smooth pre-existing dynamics in hospital performance rather than the effect of current execution frictions. To assess this possibility, Panel C of Table C.1 regresses current values of the execution wedge and current log discharges on one- and two-month leads of the instrument, controlling for hospital fixed effects, calendar-month fixed effects, and year fixed effects.

The results are mixed in an informative way. Future values of the instrument significantly predict the current execution wedge, suggesting that execution conditions are persistent over short horizons or that administrative disruptions are temporally clustered. This is not surprising in the institutional setting: delays in budget execution can propagate across adjacent months, and leadership transitions may affect the execution pipeline before and after the month in which the turnover is recorded. For this reason, the wedge placebo should not be interpreted as a clean falsification test of the mechanism.

The more relevant placebo for the exclusion restriction is whether future values of the instrument predict current hospital output. They do not. The coefficients on the one- and two-month leads of the instrument are small and statistically insignificant in the log-discharge regressions. This supports the interpretation that the main discharge result is not driven by smooth pre-trends or forward-looking dynamics in hospital performance. Rather, the evidence is consistent with the instrument affecting hospital output through contemporaneous execution frictions, while the fiscal wedge itself displays short-run persistence inside the budget-execution pipeline.

Predetermined non-fiscal operational and administrative-capacity diagnostics. Panel D extends the falsification exercise to predetermined non-fiscal characteristics of hospital operations. A remaining concern is that the instrument could be correlated with pre-existing hospital features that affect throughput independently of execution frictions. In that case, the interaction between upstream administrative disruption and lagged budget rigidity could partly proxy for differences in hospital demand composition, production-function inputs, or local administrative capacity, rather than isolating variation in the execution wedge.

We test five predetermined variables: lagged patient age, lagged doctor share, lagged management and finance staff share, lagged human-resources staff share, and lagged planning staff share. The first two capture demand composition and the medical-input composition of the hospital production function. The last three capture administrative-capacity composition, which is the dimension most directly related to the non-pipeline-channel concern. If the instrument were selecting hospitals with systematically different administrative capacity, upstream MSP disruptions could affect throughput through non-fiscal differences in local administrative performance, rather than through execution frictions summarized by the wedge.

Four of the five coefficients are statistically insignificant and close to zero. The exception is lagged management and finance staff share, where the instrument has a statistically significant coefficient of -0.0016. The magnitude is economically small: interpreted as a share, the estimate corresponds to less than two-tenths of one percentage point. We therefore do not interpret Panel D as a formal proof of orthogonality. Instead, it serves as a diagnostic. The relevant robustness exercise is whether controlling for these predetermined administrative-capacity variables changes the IV estimate. When lagged management and finance staff share is added to the IV specification, the second-stage coefficient remains essentially unchanged. Thus, while the instrument is not perfectly orthogonal to every predetermined administrative-capacity measure, the only detectable association is small and does not account for the estimated throughput effect. Moreover, the instrument does not predict lagged planning staff share ($p = 0.5442$), a proxy for hospitals' dependence on local planning capacity. This provides additional reassurance that the instrument is not selecting hospitals that were differentially exposed to non-pipeline administrative-capacity constraints before the current execution shock.

Fiscal Calendar Predeterminedness Test. Our empirical design shares some resemblance with standard shift-share designs. Those designs address share exogeneity by requiring shares to be fixed and predetermined, so that their cross-sectional variation is uncorrelated with potential outcomes — a condition tested by checking whether shares predict pre-period outcomes Goldsmith-Pinkham et al. (2020). The time-varying nature of $R_{i,t-1}$ in our design raises an analogous concern: does the within-hospital evolution of lagged rigidity reflect exogenous budget accumulation, or does it track hospital-specific performance

trajectories that independently predict discharge outcomes?

We address this concern by exploiting a feature of Ecuador's fiscal calendar that has no analog in standard shift-share designs. Within each fiscal year, $R_{i,t-1}$ rises mechanically as commitments accumulate against the codified budget — a process driven by procurement cycles, payroll schedules, and contract renewal dates set administratively rather than by hospital performance. This within-year mechanical path, interacted with predetermined hospital characteristics that determine commitment accumulation rates, provides an instrument for $R_{i,t-1}$ itself. Replacing observed $R_{i,t-1}$ with its calendar-predicted component $\widehat{R}_{i,t-1}$ yields a "purified" instrument $Zit = T_t \times \widehat{R}_{i,t-1}$ in which the share variation is driven entirely by the fiscal calendar rather than by any hospital-specific factor.

Table C.1 reports the results. Replacing the observed share with its calendar-predicted component moves the main IV estimate from -0.0613 to -0.0559, a change of 8.9 percent and well within one standard error, while the first-stage F-statistic rises from 22.6 to 26.9 — consistent with the calendar instrument extracting a cleaner signal by discarding idiosyncratic noise. Three formal predeterminedness tests reinforce this finding. Lagged rigidity does not predict MSP disruption timing after fixed effects ($p = 0.201$), ruling out systematic co-movement between the share and the shock. Changes in the residual component of rigidity do not predict discharge outcomes ($p = 0.673$), ruling out a story in which hospitals becoming progressively more rigid are simultaneously on independent declining output trajectories. Future rigidity does not predict current outcomes conditional on current rigidity ($p = 0.566$), ruling out smooth hospital performance trends driving the rigidity-performance correlation. Taken together, these results establish that the time-varying share component of the instrument satisfies a predeterminedness condition that is tested rather than assumed.

D A Conceptual Model of Fiscal Adjustment with Execution Frictions

Standard models of fiscal adjustment treat expenditure appropriations as the operative policy variable: when governments consolidate, they cut budgets, and service delivery falls in proportion. This section develops a simple alternative framework in which governments have access to a second, less visible adjustment margin—payment timing—and characterizes when they use it, why it escapes standard fiscal measurement, and what it implies for public service delivery and welfare. The framework is general; it does not exploit institutional features specific to Ecuador. Those features, documented in Section 2, are what make this margin observable in our setting.

The model proceeds in four steps. We first derive the execution wedge as an equilibrium outcome of the government's adjustment problem under fiscal stress. We then show that standard expenditure data systematically mismeasure fiscal

stance when governments use this margin. We next trace the causal chain from payment timing to service delivery. Finally, we characterize the welfare cost of the adjustment and its distributional incidence.

D.1 The Government's Adjustment Problem

Setup. A government finances a public service network. In any period, it collects revenues R and faces legally incurred spending obligations $O \geq 0$. A *fiscal gap* $G = O - R > 0$ arises when revenues fall short of obligations—as occurs, for example, when commodity-dependent revenues collapse or expenditure obligations have grown faster than the revenue base. The government must close this gap through some combination of policy adjustments.

Two margins of adjustment. The government has access to two institutional margins for closing the gap:

(i) **Appropriations adjustment (transparent margin).** The government reduces the approved budget allocation from B to $B - \Delta$, where $\Delta \geq 0$ is an explicit cut. This reduction is recorded in formal fiscal accounts, requires administrative or legislative process, and is visible to citizens, legislatures, and international monitors. Let $\theta > 0$ denote the *political cost* of appropriation cuts, capturing the institutional friction of formal budget revision.

(ii) **Payment delay (neglected margin).** The government maintains the approved budget B but reduces cash disbursements to $C \leq B$. Obligations are legally incurred but not settled in cash. The gap between obligations and payments does not appear as a reduction in formal appropriations; it accumulates as arrears or execution shortfalls inside the budget pipeline. Because payment authorization and budget approval are institutionally distinct—as they are in any system with a centralized treasury managing cash separately from line ministries managing budgets—this margin is administratively available without triggering the formal processes required for an appropriations cut.

The *execution wedge* is the equilibrium outcome of the government's use of the neglected margin:

$$\omega \equiv \ln\left(\frac{B}{C}\right) \geq 0 \tag{6}$$

When $\omega = 0$, the budget is fully executed and cash equals appropriations. When $\omega > 0$, the government is converting legal obligations into cash more slowly than it is incurring them—a form of adjustment that leaves the approved budget intact but reduces the real resources available for service production.

Optimization. The government minimizes the total cost of closing the fiscal gap G using both margins. Appropriation cuts cost $c(\Delta) = \frac{\theta}{2}\Delta^2$, reflecting convex political costs of formal budget revision. Payment delays cost $d(\omega) = \frac{1}{2}\omega^2$, reflecting convex administrative and supplier-relationship costs of deferred payment. The government's problem is:

$$\min_{\Delta \geq 0,\, \omega \geq 0} \quad \frac{\theta}{2}\Delta^2 + \frac{1}{2}\omega^2 \quad \text{subject to} \quad \Delta + B\big(1 - e^{-\omega}\big) \geq G \tag{7}$$

The constraint requires that the combined resources freed by explicit cuts and payment delays be sufficient to close the fiscal gap.

Proposition D.1 (The Wedge as an Equilibrium Policy Variable)**.** *For any $G > 0$ and $\theta > 0$, the solution to (7) involves strictly positive use of both adjustment margins when the constraint binds. The equilibrium execution wedge $\omega^*(G, \theta)$ satisfies:*

$$\omega^*(G, \theta) > 0 \qquad \textit{and} \qquad \frac{\partial \omega^*}{\partial G} > 0, \quad \frac{\partial \omega^*}{\partial \theta} > 0 \tag{8}$$

The wedge is strictly increasing in the fiscal gap and strictly increasing in the political cost of transparent adjustment.

Proof. At an interior solution, the first-order conditions give $\theta\Delta^* = \lambda$ and $\omega^* \cdot Be^{-\omega^*} \cdot \frac{1}{2} = \lambda$ for multiplier $\lambda > 0$ on the fiscal constraint. Eliminating λ, the ratio $\theta\Delta^*/(\omega^* Be^{-\omega^*}) = 1$ pins the optimal mix. Since both c and d are strictly convex with $c(0) = d(0) = 0$, the unconstrained minimum of each is at the origin; the constrained solution requires both margins to be active whenever $G > 0$. The comparative statics follow by implicit differentiation of the binding constraint: a larger gap G requires more total adjustment, and since costs are symmetric in convexity, both margins expand; a higher θ raises the marginal cost of explicit cuts, rotating the optimal mix toward greater reliance on the neglected margin. □ □

Interpretation. Proposition D.1 establishes three things simultaneously. First, the execution wedge is not measurement error, accounting noise, or bureaucratic slippage—it is a policy variable that the government actively chooses in response to fiscal pressure. Second, the wedge responds systematically to fiscal stress: larger shocks and higher political costs of transparent adjustment both generate larger wedges. Third, because $\theta > 0$ captures any institutional or political friction that makes formal budget revision costly, the model applies broadly across institutional settings: any government with separated budget authorization and cash management will have $\omega^* > 0$ under fiscal stress.[14]

[14] The result generalizes to asymmetric cost structures. What matters is that $c(\cdot)$ and $d(\cdot)$ are both strictly convex and that $\theta > 0$, ensuring that the marginal cost of explicit cuts is strictly positive. Even when θ is very small, interior solutions exist for any $G > 0$.

D.2 Why Standard Expenditure Data Mismeasure Fiscal Stance

A researcher who observes only approved budgets B would conclude that fiscal stance is unchanged when ω rises, because appropriations appear constant. This section formalizes the mismeasurement that results.

Proposition D.2 (Mismeasurement of Fiscal Stance)**.** *Let $\Omega_t \equiv B_t e^{-\omega_t}$ denote the* effective fiscal stance*—the cash resources actually available for service production in period t. Standard expenditure data record B_t. The measurement error is:*

$$B_t - \Omega_t = B_t\big(1 - e^{-\omega_t}\big) > 0 \quad \textit{whenever} \quad \omega_t > 0 \tag{9}$$

This error is:

(i) ***Systematically positive****: approved budgets overstate effective resources whenever governments use the neglected adjustment margin.*

(ii) ***Increasing in fiscal stress****: by Proposition D.1, ω_t rises with the fiscal gap, so the mismeasurement is largest precisely when accurate measurement of fiscal stance matters most.*

(iii) ***Invisible to standard audits****: since appropriations are unchanged, conventional expenditure analysis and budget reviews detect no adjustment.*

Proof. Direct from definition (3): $C_t = B_t e^{-\omega_t} < B_t$ whenever $\omega_t > 0$, so $B_t - \Omega_t = B_t(1 - e^{-\omega_t}) > 0$. Properties (i) and (ii) follow from the strict monotonicity of $e^{-\omega}$ in ω and from the comparative statics in Proposition D.1. Property (iii) follows because B_t is recorded in formal accounts while C_t and ω_t require access to treasury cash management records that are institutionally separate from appropriations data. □ □

Implication for fiscal analysis. Proposition D.2 reframes a question central to the fiscal policy literature: what is the transmission of fiscal consolidation to real activity? The standard approach measures fiscal stance using appropriations or reported expenditures and estimates its effect on output. When governments use the denoted adjustment margin, this approach systematically understates the fiscal contraction, leading researchers to underestimate the real effect of consolidation episodes on service delivery. The execution wedge ω_t—the gap between legal obligations and cash payments—is the correction term that the standard approach omits. Identifying the causal effect of this term on service delivery, therefore, answers a question that appropriations-based fiscal analysis cannot.

D.3 The Causal Chain from Payment Timing to Service Delivery

Service delivery units—hospitals, schools, infrastructure agencies—acquire inputs to produce public services. Their production depends on resources that

are *actually paid*, not resources that are *formally approved*:

$$Y_t = f(\Omega_t) = f\left(B_t e^{-\omega_t}\right), \quad f' > 0 \tag{10}$$

The function f is any increasing production function mapping effective resources to output; the critical assumption is that production depends on cash received, not budgets approved. This is non-trivial in two respects. First, it requires that service-delivery units cannot self-finance—that they have no independent access to capital markets or own revenue streams that would allow them to bridge the gap between approved and paid resources. Second, it requires that inputs must be acquired contemporaneously with service delivery, so that a delay in payment is not merely a timing inconvenience but an input constraint.

Assumption D.1 (No Self-Financing and Non-Storability)**.** Service-delivery units have no access to off-budget financing and hold no independent cash balances. Services are non-storable: output foregone in period t due to input shortfalls cannot be produced in period $t+1$ and retroactively credited to period t.

Assumption D.1 is not a generic simplification; it is a precise characterization of how public service networks operate under centralized treasury regimes, documented for the Ecuador setting in Section 2. It is also the critical distinction between the public and private sectors: a private hospital can borrow against future revenues; a public hospital under a Treasury Single Account regime cannot.

Proposition D.3 (Causal Effect of Payment Timing on Service Delivery)**.** *Under Assumption D.1, holding the approved budget B_t constant, an increase in the execution wedge ω_t reduces service delivery:*

$$\left.\frac{dY_t}{d\omega_t}\right|_{B_t} = f'(\Omega_t) \cdot (-\Omega_t) < 0 \tag{11}$$

Moreover, this effect is:

(i) ***Invisible to appropriations-based analysis****: since B_t is unchanged, a researcher using only budget data would observe no policy change and attribute the output loss to non-fiscal factors.*

(ii) ***Irreversible****: delayed payments that eventually arrive in period $t+1$ do not recover period-t output, because services are non-storable and the delivery window has closed.*

(iii) ***Amplified by institutional payment hierarchies****: when a subset of input expenditures is institutionally protected from delay (e.g., public payroll), the full burden of the cash shortfall falls on the residual flexible inputs, generating a production loss larger than a proportional resource cut would imply.*

Proof. Equation (11) follows by differentiating (10): $dY/d\omega = f'(\Omega) \cdot d\Omega/d\omega = f'(\Omega) \cdot (-Be^{-\omega}) = -f'(\Omega) \cdot \Omega < 0$. Property (i) holds because (10) and (3) together imply that Y_t falls while B_t is constant; no information in B_t alone signals the change. Property (ii) follows from Assumption D.1: under no self-financing, the hospital cannot bridge the period-t input gap against the promise of period-$(t+1)$ settlement, and non-storability prevents retroactive production. Property (iii) follows from the structure of the production function: if protected inputs $\bar{X}$ are fixed and unresponsive to cash shortfalls, then $f(\Omega) = g(\bar{X}, \Omega - \bar{X})$ where $\Omega - \bar{X}$ is the residual flexible margin; the elasticity of f with respect to ω is amplified by the ratio $\Omega/(\Omega - \bar{X}) > 1$. □ □

The empirical question. Proposition D.3 establishes that the execution wedge has a causal, negative, and irreversible effect on service delivery—but it does not identify its magnitude. That magnitude is what the quasi-experimental analysis estimates. The challenge is that ω_t is endogenous: governments in fiscal stress may simultaneously delay payments and face other shocks that independently affect service delivery. Our instrumental-variables strategy, documented in Section 3, isolates variation in ω_t driven by the interaction of upstream administrative disruptions and predetermined hospital budget rigidity—a source of variation that operates through the fiscal pipeline and satisfies the exclusion restriction under the conditions stated there.

D.4 Welfare Implications

The preceding propositions characterize the fiscal adjustment through the wedge as a causal driver of service delivery losses. We now address the welfare question: how should we evaluate the costs of this adjustment mechanism relative to transparent alternatives, and on whom do those costs fall?

The welfare cost of hiding adjustment. A social planner observing the true fiscal gap G could, in principle, design an adjustment that minimizes welfare losses. Such a design would be informed by the marginal welfare value of different categories of public spending, and subject to democratic accountability mechanisms. The adjustment through the wedge bypasses both. The welfare cost of that margin therefore exceeds the welfare cost of an equivalent transparent adjustment for two reasons beyond the production losses already identified.

Proposition D.4 (The Wedge Adjustment Is Welfare-Inferior)**.** *Let $W(\Delta, \omega)$ denote the social welfare loss from a combination of transparent cut Δ and payments delay ω that together close fiscal gap G. For any adjustment path that closes G using only the hidden margin $(\Delta = 0, \omega > 0)$ relative to a path using only the transparent margin $(\Delta > 0, \omega = 0)$ with equal aggregate resource withdrawal, the path generates strictly higher welfare losses:*

$$W(0,\, \omega^{eq}) > W(\Delta^{eq},\, 0) \tag{12}$$

where ω^{eq} and Δ^{eq} each close G independently. This follows from two mechanisms that compound the production loss in Proposition D.3:

(i) ***Amplification****: By property (iii) of Proposition D.3, adjustment through the wedge falls disproportionately on flexible operating inputs due to payroll protection, generating a production loss larger than a proportional budget cut of equal magnitude would imply.*

(ii) ***Unaccountability****: Transparent cuts are subject to legislative scrutiny, public reporting, and international fiscal monitoring, all of which impose a social check on the depth of adjustment. The adjustment bypasses these mechanisms, enabling larger effective resource withdrawals than would survive political accountability if recorded in standard fiscal aggregates.*

Distributional incidence. The welfare costs of this type of fiscal adjustment are not distributed neutrally across the population. Two features of the mechanism concentrate the incidence.

Corollary D.5 (Incidence on Captive Users)**.** *The welfare cost of the fiscal adjustment in public health falls disproportionately on patients who:*

(i) Cannot substitute to alternative providers. Public health patients in developing-country settings typically have no private-sector exit option at the margin where public care quality deteriorates. The price of the public service is zero; the alternative is unaffordable.

(ii) Depend on sustained inpatient inputs during their admission. Patients admitted for acute or chronic conditions requiring continuous pharmaceutical treatment, monitoring, and nursing supply are more exposed than patients whose outcomes are determined by the initial clinical response at admission.

The mortality evidence in Section 5.5 identifies precisely this group: over-48-hour inpatient mortality rises significantly while under-48-hour mortality does not, concentrating the mortality cost on patients whose survival depends on sustained input availability during the hospital stay.

Implications for fiscal measurement and reform. Together, the four propositions and their corollary identify a coherent failure in how fiscal policy is measured, transmitted, and reformed. On measurement: approved budgets overstate effective resources when governments use the wedge margin of adjustment, so standard fiscal aggregates mischaracterize both the depth of consolidation and its welfare cost. On transmission: the channel from fiscal stress to service delivery runs through payment timing, not only through appropriations, so models that abstract from the timing of cash execution understate the real consequences of consolidation. On reform: improving cash management systems, tightening commitment controls, and making payment authorization more timely are policy interventions that can recover service delivery without

additional budget allocations—the relevant counterfactual is not more spending but more reliable and timely execution of what is already committed.

These implications are not specific to Ecuador's public health system. They apply to any government that (a) operates a centralized treasury managing cash separately from line agencies managing budgets, (b) finances public services that cannot self-fund, and (c) faces fiscal stress that makes the wedge manipulation adjustment margin attractive. This describes the institutional architecture of a large share of low- and middle-income countries, and it describes the fiscal episode we study. The contribution of our empirical analysis is to identify the magnitude of this mechanism where institutions make the execution pipeline observable and its frictions plausibly exogenous—not to claim that Ecuador is unique, but to provide the first credible estimate of a cost that is general.