Capacity Constraints, Encounter Frequency, and the Paradox of Tightening

Endre J. Reite

Norwegian University of Science and Technology, NTNU Business School

Abstract

I study enforcement systems that observe regulated activity through discrete administrative encounters (filings, inspections, audits) under binding screening capacity. When agents can choose how to partition a fixed activity volume across encounters, per-encounter procedural burdens induce consolidation into fewer and larger encounters. If screening resources are allocated at the encounter level and scale sublinearly with encounter scope, detectability declines with encounter size. In that case, tightening procedural requirements can reduce equilibrium detection by triggering discrete reductions in encounter frequency, generating a non-monotonic relationship between stringency and effective enforcement. The analysis delivers comparative statics for consolidation thresholds and shows how design features that scale screening with encounter size—through sampling, automation, or size-contingent escalation—eliminate the paradox channel by preserving screening intensity per unit.

Keywords: enforcement, monitoring, administrative capacity, JEL: D82, H26, K42

Email address: endre.j.reite@ntnu.no (Endre J. Reite)

1. Introduction

Administrative enforcement is organized around discrete encounters: filings, inspections, audit episodes, case reviews, follow-up meetings, and other points of contact between the state and regulated agents. In many environments, the regulated party can influence the encounter process itself: how often to appear, how much to bundle into each encounter, and when to engage. These organizational choices shape what the regulator observes, how screening resources are allocated, and how enforcement effectiveness is measured in administrative data. This paper studies enforcement under binding administrative capacity when encounter frequency and encounter scope are endogenous.

Question and object. I ask how changes in *per-encounter* procedural requirements affect (i) encounter frequency, (ii) encounter scope, and (iii) equilibrium detectability when screening capacity is allocated at the encounter level. The key modeling choice is to hold the underlying activity volume fixed at $Y > 0$ and focus on how that fixed workload is partitioned across encounters. This isolates an encounter-organization margin that is typically treated as exogenous in enforcement and compliance frameworks. Standard deterrence effects (changes in Y) and regulator optimization can be layered on without removing the channel studied here.

Core mechanism: consolidation under per-encounter burden. Encounters impose two costs on agents. First, there is a convex within-encounter processing cost (congestion, coordination, and complexity) that makes very large encounters internally expensive. Second, there is a per-encounter *exposure cost*—the overhead of appearing—capturing documentation, procedural friction, follow-ups, managerial attention, and the risk of escalating scrutiny triggered by repeated contact. When per-encounter exposure rises, agents economize on appearances by consolidating the fixed workload into fewer encounters. Because the encounter count is discrete, this consolidation occurs in steps: as exposure increases, agents switch from n to $n-1$ encounters at thresholds, raising average

encounter scope from Y/n to $Y/(n-1)$.

Architecture and the paradox of tightening. The central implication depends on enforcement architecture: how screening resources scale with encounter scope. In many settings, screening time, attention, and processing capacity are allocated *per encounter* and do not scale proportionally with the amount bundled into that encounter. I model this as a fixed encounter-level screening capacity $B > 0$, so screening intensity per unit is $e(y) = B/y$ for an encounter of scope y, and detectability is $p(y) = \pi(B/y)$ with diminishing returns in screening intensity. Under this architecture, consolidation dilutes screening intensity per unit. This yields a paradox:

> **Paradox of tightening.** Increasing per-encounter procedural stringency can reduce equilibrium detection when agents respond by consolidating activity into fewer, larger encounters and detectability declines with encounter scope because screening capacity is allocated at the encounter level.

Detection falls discretely at consolidation thresholds: moving from n to $n-1$ encounters reduces the number of independent screening opportunities and lowers per-encounter detectability by increasing encounter scope. When this dilution effect dominates, the overall probability of detection declines even though the regime has become "stricter" in the sense of higher per-encounter requirements.

Main results. The analysis delivers three sets of results.

1. **Endogenous encounter frequency and consolidation thresholds** With convex within-encounter processing costs and non–decreasing marginal exposure costs, the optimal encounter count exists and is weakly decreasing in per-encounter burden: tightening induces consolidation into fewer encounters and larger average scope.

2. **Persistence under history-dependent scrutiny** When exposure depends on accumulated scrutiny from past encounters (risk scoring, follow-ups, escalation), encounter patterns display persistence: temporary increases in scrutiny can produce sustained reductions in encounter frequency as agents "cool off" while scrutiny decays.
3. **Non-monotonic detection under capacity dilution.** When detectability declines with encounter scope because screening capacity does not scale proportionally, equilibrium detection can be non-monotonic in procedural stringency. In particular, at consolidation thresholds, tightening can reduce total detection—establishing the paradox of tightening and characterizing when it arises.

Interpretation for measurement. A direct implication is that administrative events (filings, audits, cases) are equilibrium objects. A decline in encounter counts need not reflect a decline in underlying activity; it may reflect consolidation driven by higher exposure or tighter procedures. Likewise, higher per-encounter scrutiny can coincide with lower detection per unit of underlying activity if activity is bundled into fewer, harder-to-screen encounters. Empirical work that uses encounter counts or audit episodes as proxies for activity or enforcement intensity should therefore track encounter scope and timing jointly, not only counts.

Policy design: encounter architecture as an instrument. The paper emphasizes that tightening "rules" is not the same as improving observability. If per-encounter burdens rise without scaling screening capacity with encounter scope, agents rationally consolidate, potentially undermining detection. The natural complement is architectural design: rules that (i) preserve encounter frequency (lightweight periodic contact), (ii) prevent excessive bundling (caps or minimum frequency requirements), or (iii) scale screening with scope (within-encounter sampling, automation, escalation protocols that allocate more capacity to larger encounters). These design levers operate directly on the encounter process and can mitigate capacity dilution.

Empirically, the paper provides a small illustrative exhibit rather than a causal policy evaluation. The purpose is to demonstrate how the model's core objects—encounter frequency, encounter scope, and their joint dynamics—appear in administrative data when *per-encounter* procedural burden shifts. Using discrete changes in internal procedural requirements, the exhibit documents patterns consistent with exposure-driven consolidation and persistence, without relying on eligibility cutoffs or exogenous assignment. Accordingly, robustness checks target mechanical timing and transition artifacts around regime changes, not causal treatment effects.

Roadmap. Section 2 relates the mechanism to enforcement models, administrative burden, threshold behavior, and capacity-constrained monitoring. Section 3 develops the baseline partitioning model and comparative statics, then adds history-dependent exposure and size-sensitive detectability to derive the paradox of tightening. Section 4 interprets the results through the lens of enforcement architecture and derives design implications under capacity constraints. Section 6 summarizes empirical signatures and measurement guidance, and Section 7 concludes.

2. Related Literature and Contribution

This paper sits at the intersection of enforcement/compliance models and work on monitoring design. There is an extensive literature on how penalties, audit probabilities, and enforcement intensity affect compliance and underlying activity, typically taking the timing or frequency of regulatory encounters as given. A complementary line of research emphasizes state dependence, targeting, or follow-up that evolves with past interactions, creating dynamic incentives that can generate persistence in observed outcomes. The present paper focuses on a distinct margin that becomes central whenever encounter timing or partitioning is discretionary: the organization of a fixed workload across encounters. I organize the discussion around encounter timing/frequency, administrative burden, monitoring design, and capacity constraints.

2.1. Enforcement with administrative burden and capacity constraints

Much of the enforcement literature studies behavior *conditional on appearing* (e.g., reporting, compliance effort), while treating the timing of encounters as exogenous. In contrast, the present paper highlights a complementary margin: when scrutiny depends on past interactions (risk scoring, escalation, follow-ups), agents can respond by reducing encounter frequency and retiming activity, generating persistence in the encounter process even after the temporary shock dissipates.

This perspective is consistent with models of state-dependent enforcement and dynamic audit intensity, and aligns with hysteresis logic in which temporary shocks produce lasting changes because the state evolves with past actions. In this paper, persistence arises because accumulated scrutiny makes returning more costly, so agents rationally 'cool off' by delaying encounters until the scrutiny stock decays.

A key premise of this study is that enforcement costs are not limited to expected fines. Encounters can impose real procedural burdens related to, for example, documentation, follow-up tasks, time, and cognitive managerial attention. The administrative-burden framework formalizes these interaction costs as learning, psychological, and compliance burdens that arise in encounters with the state (Moynihan, Herd, and Harvey, 2015).

I treat exposure costs in reduced form to isolate how procedural frictions reshape the encounter process rather than institution-specific details, consistent with enforcement systems that allocate scarce scrutiny using risk information and targeted capacity.

2.2. Monitoring Design, Thresholds, and Strategic Responses

The paper's monitoring-design results connect naturally to settings with discontinuities and size-based regimes. Evidence on tax and regulatory thresholds shows that kinks or notches can reshape observed distributions (bunching) and that behavioral responses need not map one-for-one to real activity changes (Kleven, 2016). Audit experiments and information-based enforcement further

show that reporting and measured compliance clearly respond to audit probabilities and information trails (Kleven et al., 2011; Pomeranz, 2015). Size-dependent monitoring can induce strategic responses around eligibility cutoffs (Almunia and Lopez–Rodriguez, 2018; Kleven and Waseem, 2013).

Finally, the monitoring-redesign approach is consistent with evidence that size-contingent regulation distorts organizational scale and observed distributions (Garicano, Lelarge, and Van Reenen, 2016), reinforcing the broader point that the enforcement architecture can change behavior *and* the informational content of administrative data.

The capacity-dilution mechanism in Subsection 3.10 is consistent with bounded-attention and processing-constraint foundations in which effective scrutiny is limited by scarce cognitive or organizational capacity (Sims, 2003; Matějka and McKay, 2015; Gabaix, 2014; Maćkowiak, Matějka, and Wiederholt, 2023).

2.3. Contribution relative to the literature

Taken together with history-dependent scrutiny and size-sensitive detectability, the framework generates persistence in encounter patterns and identifies conditions under which stricter per-encounter requirements can reduce detection by inducing consolidation into fewer, larger encounters. The partitioning problem also mirrors classic batching and lumpy-adjustment logic (Baumol, 1952; Tobin, 1956; Caballero and Engel, 1999); I keep extended discussion and additional references in Appendix A.

3. Model

3.1. Environment

Consider a risk-neutral agent operating over a fixed horizon (one period). The agent must conduct a fixed total volume of economic activity, denoted by $Y > 0$, that is subject to regulatory oversight. Each time the agent engages with the regulator (for example, by filing a report, responding to an inquiry, or undergoing an inspection) constitutes an interaction.

The agent chooses the following:

- an integer number of interactions $n \in \{1, 2, 3, \ldots\}$, and
- an allocation of activity across interactions $(y_1, \ldots, y_n)$ with $y_i \geq 0$ for all i.

All activity must be processed or reported within the following horizon:

$$\sum_{i=1}^{n} y_i = Y. \tag{1}$$

The model isolates a margin that is often treated as exogenous in enforcement frameworks: the agent can choose how often to interact and, therefore, the size of each interaction.

Scope note. Holding total activity fixed isolates the organizational margin, how activity is distributed across regulatory encounters, rather than the level of that activity. This assumption is appropriate in settings where agents face fixed demand, mandated output, or contractual obligations. Allowing activity to be elastic would introduce standard deterrence effects; these would coexist with, rather than eliminate, the consolidation mechanism studied here.

3.2. Capacity Constraints in Administrative Enforcement

In many enforcement environments, screening resources are allocated at the encounter level and are subject to binding time and processing constraints. When activity is consolidated into larger encounters without proportional scaling of screening capacity, effective scrutiny per unit can fall. Encounter size y refers to the scope bundled into a single administrative interaction—most naturally, the number of issues/periods or line items examined within a tax audit (and in customs, the number of units/containers inspected)—while Y denotes total underlying activity over the relevant horizon.

My capacity-dilution assumption captures these settings by modeling a bounded screening capacity per encounter. The mechanism applies when per-encounter screening time is capped and when agents can reorganize activity across encounters. It is less relevant in regimes that automatically escalate large encounters into separate, uncongested processes or specialized units. In Section 4, I treat

encounter design as a policy lever and describe how capacity constraints shape optimal administrative rules.

3.3. Cost Structure and Cumulative Exposure

Total cost has two components: an internal (batching) cost and an enforcement (exposure) cost. In many regulatory environments—such as voluntary filing and self-reporting regimes, discretionary transactions, and systems that allow extensions or amendments—agents have considerable leeway over when and how often they interact. The model abstracts from hard periodic mandates to isolate the extreme case of endogenous timing, clarifying what is possible when agents can bundle or delay interactions.

(i) Internal batching cost. Processing a large amount of activity in a single interaction is internally costly (administrative congestion, coordination, and complexity). Then let *g(y)* denote the internal cost of processing amount y in one interaction.

Assumption 1 (Convex internal cost)**.** $g(0) = 0$*,* g *is increasing, and* g *is strictly convex on* $\mathbb{R}_+$*.*

This captures the idea that making a single interaction very large becomes increasingly costly, reflecting the increase in internal challenges as activity is bundled.

(ii) Cumulative exposure cost. Each interaction creates regulatory exposure related to documentation, follow-up requests, managerial attention, and the risk of triggering enhanced scrutiny. Let *c(k)* denote the incremental exposure cost of the *k-th* interaction in the period. Interpret $c(k)$ as the fixed procedural overhead plus the expected follow-up burden induced by an additional encounter (documentation cycles, case handling, and escalation risk), in reduced form. The total exposure cost from n interactions is:

$$C(n) = \sum_{k=1}^{n} c(k). \tag{2}$$

Assumption 2 (Escalating marginal exposure)**.** *For all* $k \geq 1$*,* $c(k+1) \geq c(k)$*.*

This allows later interactions to be weakly more burdensome than earlier ones.

Escalating exposure is plausible because encounters are not independent. Each interaction expands the record and can trigger a follow-up, update risk scores, and raise review intensity in later encounters. This captures routine enforcement practice in settings with targeted monitoring and state dependence (e.g., risk scoring, follow-ups, escalation). I treat (c(k)) as a reduced form for this history dependence.

3.4. Agent's Problem

The agent must carry out total activity of $Y > 0$ over the horizon. They choose an integer $n \in \mathbb{N}$ and a partition $\{y_i\}_{i=1}^n$ with $y_i \geq 0$ and $\sum_{i=1}^n y_i = Y$. The total cost is

$$\text{TotalCost}(n, \{y_i\}_{i=1}^n) = \sum_{i=1}^n g(y_i) + C(n), \tag{3}$$

where $g(\cdot)$ captures the within-encounter adjustment costs and $C(\cdot)$ captures the exposure costs that increase with the number of encounters. The agent solves

$$\min_{n \in \mathbb{N},\, y_i \geq 0} \sum_{i=1}^n g(y_i) + C(n) \quad \text{s.t.} \quad \sum_{i=1}^n y_i = Y. \tag{4}$$

In this problem, internal economies of scale are traded off against exposure: increasing n lowers the convex cost component $\sum_i g(y_i)$, but raises $C(n)$ by adding encounters.

3.5. Optimal Allocation Conditional on Interaction Count

Lemma 1 (Equal-split allocation)**.** *Fix $n \geq 1$. Under Assumption 1, the unique solution to*

$$\min_{\{y_i\}_{i=1}^n} \sum_{i=1}^n g(y_i) \quad \textit{s.t.} \quad \sum_{i=1}^n y_i = Y, \qquad y_i \geq 0,$$

is $y_i^\star = Y/n$ for all $i = 1, \ldots, n$. Consequently, the reduced-form objective is

$$\Phi(n) = n\, g\left(\frac{Y}{n}\right) + C(n). \tag{5}$$

Proof. By Jensen's inequality,

$$\frac{1}{n}\sum_{i=1}^{n} g(y_i) \;\geq\; g\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = g(Y/n),$$

so $\sum_{i=1}^{n} g(y_i) \geq n\, g(Y/n)$. Strict convexity implies equality only if $y_1 = \cdots = y_n$, which, together with $\sum_i y_i = Y$ yields $y_i^\star = Y/n$ for all i. Substituting the optimizer into equation (3) gives equation (5). □

3.6. Existence and Static Comparative Statics

Assumption 3 (Coercivity)**.** *$C(n) \to \infty$ as $n \to \infty$.*

(This holds for many schedules, including constant c(k) > 0.)

Proposition 1 (Existence of an optimal interaction count)**.** *Under Assumptions 1 and 3, the reduced objective, equation* (5)*, attains a minimum at some finite $n^\star \in \mathbb{N}$.*

Proof. Because $g(\cdot) \geq 0$, equation (5) implies $\Phi(n) \geq C(n)$ for all n. Assumption 3 gives $C(n) \to \infty$ as $n \to \infty$, hence $\Phi(n) \to \infty$. Since Φ is defined on the positive integers, it attains a minimum on a finite set $\{1, 2, \ldots, N\}$ for some sufficiently large N, yielding a finite minimizer, $n^\star$. □

Definition 1 (Stricter exposure regime)**.** *Regime B is (weakly) stricter than regime A if $c_B(k) \geq c_A(k)$ for all $k \geq 1$, with strict inequality for at least one k.*

Proposition 2 (Consolidation under higher marginal exposure)**.** *Let g satisfy Assumption 1 and let $C_r(n) = \sum_{k=1}^{n} c_r(k)$ for regimes $r \in \{A, B\}$. If regime B is stricter than regime A in the sense of Definition 1, then the smallest minimizer*

$$\underline{n}_r \in \arg\min_{n \in \mathbb{N}} \left\{ n\, g\left(\frac{Y}{n}\right) + C_r(n) \right\}$$

satisfies $\underline{n}_B \leq \underline{n}_A$.

Proof. Define $H(n) \equiv n\, g(Y/n)$ and $\Phi_r(n) \equiv H(n) + C_r(n)$. For each regime r,

$$\Phi_r(n+1) - \Phi_r(n) = \big(H(n+1) - H(n)\big) + c_r(n+1).$$

By Assumption 1, g is strictly convex, which implies $H(n)$ is convex in n on $\mathbb{N}$ and, hence, the discrete difference $H(n+1) - H(n)$ is nondecreasing in n. If $c_r(\cdot)$ is nondecreasing (Assumption 2), then $\Phi_r(n+1) - \Phi_r(n)$ is nondecreasing in n. Therefore, the minimizing set is an interval of integers and the smallest minimizer is characterized by the first index where the forward difference is nonnegative.

Since $c_B(k) \geq c_A(k)$ for all k, we have $\Phi_B(n+1) - \Phi_B(n) \geq \Phi_A(n+1) - \Phi_A(n)$ for all n, so the crossing to nonnegativity occurs slightly earlier under B, implying $\underline{n}_B \leq \underline{n}_A$. □

3.7. Static Benchmark Illustration

This partitioning problem mirrors classic inventory and transaction models in which agents trade off convex adjustment or processing costs against fixed or per-occurrence costs, generating endogenous batching and lumpy timing (Baumol, 1952; Tobin, 1956; Caballero and Engel, 1999). I make this trade-off concrete by considering a simple parameterization with total activity $Y = 100$ and internal processing cost $g(y) = y^2$. Compare two exposure schedules:

- A baseline schedule with moderate per-interaction exposure costs.
- A stricter schedule with higher exposure costs in which additional interactions are more heavily penalized.

For each candidate interaction count n, the reduced cost is:

$$\Phi(n) = n \cdot g\left(\frac{Y}{n}\right) + C(n). \tag{6}$$

3.8. Dynamic Scrutiny, Persistence, and Hysteresis

Next, I extend the benchmark by allowing regulatory exposure to depend on past interactions. The goal is not to add strategic behavior on the part of the regulator to the model but to capture the empirically common feature that scrutiny is persistent: past contact affects the expected burden of future contact.

Time is indexed discretely by $t = 1, 2, \ldots, T$. Let $s_t \geq 0$ denote a scrutiny stock summarizing accumulated information, attention, or follow-up obligations associated with the agent at the start of period t. Scrutiny evolves according to:

$$s_{t+1} = (1 - \delta)s_t + \psi \, 1\{\text{interaction at } t\}. \tag{7}$$

where $\delta \in (0, 1)$ governs decay ('cooling off') and $\psi > 0$ captures the incremental effect of an interaction on future scrutiny.

An interaction at time t imposes an exposure cost $c(s_t)$, where $c(\cdot)$ is increasing. Internal processing costs are as before. The agent allocates a fixed total activity Yacross interaction dates and sizes over the horizon, taking the law of motion for s_t as given.

This formulation nests the static benchmark as the special case $\psi = 0$ (or $\delta = 1$), but differs in one key respect: the marginal cost of an additional interaction depends on its timing relative to past interactions.

Proposition 3 (Persistence following a temporary scrutiny shock)**.** *Consider two periods. Let the agent's period-t objective be* $\Phi(n_t; s_t) = n_t \, g(Y/n_t) + C(n_t; s_t)$*, where* $C(\cdot; \cdot)$ *has increasing differences in* (n, s)*. Scrutiny follows*

$$s_{t+1} = (1 - \delta)s_t + a_t, \qquad 0 \leq \delta < 1, \tag{8}$$

with $a_1 > a_2$ *and* s_1 *given. Even if* a_t *reverts in period 2, the elevated state* s_2 *induced by* a_1 *weakly reduces the agent's optimal interaction count in period 2 relative to the no-shock path where* $a_1 = a_2$*.*

3.9. Monitoring Design and Strategic Encounter Avoidance

The persistence mechanism above takes scrutiny as exogenous to policy design. This subsection adds a minimal strategic layer to clarify how enforcement rules interact with endogenous interaction choices.

Consider a two-player game between a regulator and the agent. The regulator commits to a monitoring rule before activity commences. According to the baseline rule, scrutiny depends only on the following state:

$$a\left(s_t\right) \text{ or } d\left(s_t\right) \tag{9}$$

where $a(\cdot)$ denotes audit probability and $d(\cdot)$ denotes audit depth. The agent then chooses the timing and size of interactions, anticipating how scrutiny evolves as set out in Section 3.8.

Under this encounter-only monitoring regime, increasing the intensity of scrutiny has a clear behavioral effect.

Proposition 4 (Strategic consolidation with encounter monitoring)**.** *When monitoring intensity depends only on the scrutiny state s_t, an increase in per-encounter scrutiny weakly reduces encounter frequency and weakly increases average encounter size, holding total activity constant.*

The reason is mechanical. A higher expected burden per–encounter raises the shadow cost of appearing. Because total activity is fixed, the agent economizes on appearances by bundling activity into fewer interactions.

Next, I consider an alternative monitoring design. Suppose the regulator conditions scrutiny on both history *and* the size of the interaction:

$$a(s_t, y_t) \text{or} d(s_t, y_t), \tag{10}$$

with monitoring intensity increasing in y_t.

Proposition 5 (Size-contingent monitoring limits consolidation)**.** *Suppose size-contingent monitoring enters the agent's objective as an additional per-encounter burden that is increasing in encounter size. Specifically, let total cost be*

$$\sum_{i=1}^{n} g(y_i) \;+\; C(n) \;+\; \kappa \sum_{i=1}^{n} m(y_i),$$

where $m : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing and $\kappa \geq 0$ indexes the strength of size-contingent monitoring. Under Assumption 1, if m is convex, the equal-split allocation remains optimal conditional on n, and the reduced problem is

$$\min_{n \in \mathbb{N}} \Big\{ n\, g(Y/n) + C(n) + \kappa\, n\, m(Y/n) \Big\}.$$

Assume the discrete function $T(n) \equiv n\, m(Y/n)$ is decreasing and convex on $\mathbb{N}$. Let $\underline{n}(\kappa)$ denote the smallest minimizer of the reduced problem. Then $\underline{n}(\kappa)$ is weakly increasing in κ. In particular, $\underline{n}(\kappa) \geq \underline{n}(0)$ for all $\kappa \geq 0$.

Proof. See Appendix B.2. □

This extension has policy implications but does not establish an additional mechanism. When encounters are endogenous, enforcement rules that shift scrutiny entirely to the act of appearing can reduce information flow. Allowing monitoring intensity to vary with interaction size, or maintaining low-burden frequent contact, mitigates strategic avoidance without weakening deterrence. Additional derivations are collected in Appendix B.

Table 2 summarizes the agent's optimal interaction frequency under encounter-only and size-contingent monitoring, holding total activity and internal processing costs constant. Figure 2 plots the total cost as a function of the interaction count under each monitoring rule. In the calibration shown, the discrete optimum shifts from $n^* = 5$ under encounter-only monitoring to $n^* = 6$ under size-contingent monitoring.

The comparison shows that conditioning scrutiny on interaction size raises the marginal cost of bundling activity and shifts the cost–minimizing interaction frequency upward.

3.10. Endogenous Observability and Information Collapse

The analysis thus far has treated regulatory encounters as costly but equally informative. In many settings, however, how activity is organized affects what regulators learn. Larger submissions, filings, or transactions can be harder to screen, audit, or interpret, even if the level of scrutiny per-encounter is high. I then show how the organizational responses studied above can, therefore, reduce both exposure and observability (Maćkowiak et al., 2023; Gabaix, 2014).

Suppose that a violation $v \in \{0, 1\}$ occurs with some exogenous probability. Conditional on an interaction of size y, the regulator detects a violation with probability $p(y)$. Detection probabilities are independent across interactions.

Larger encounters can be harder to screen effectively. Auditors have limited attention and processing capacity: reviewing one massive submission may have diminishing returns relative to the review of several smaller ones. This 'haystack effect' is consistent with rational inattention and observed audit failures in complex organizations, where scale overwhelms scrutiny.

Assumption 4 (Capacity-constrained detection)**.** *Each enforcement encounter is subject to a fixed screening capacity $B > 0$ (time, staff–hours, or processing resources) allocated at the encounter level. For an encounter of size $y > 0$, the screening intensity per unit is*

$$e(y) \equiv \frac{B}{y}.$$

Detection probability is a function of screening intensity,

$$p(y) \equiv \pi(e(y)) = \pi\left(\frac{B}{y}\right),$$

where $\pi'(\cdot) > 0$ and $\pi''(\cdot) < 0$.

Remark. (Sublinear scaling.) The dilution channel persists if encounter-level capacity scales *sublinearly* with scope, e.g. $B(y) = B_0 y^{\alpha}$ with $\alpha \in [0, 1)$, so that $e(y) = B(y)/y = B_0 y^{\alpha-1}$ declines in y. The knife-edge case $\alpha = 1$ corresponds to proportional scaling (no mechanical dilution), which removes the paradox channel.

Assumption 4 captures 'capacity dilution': holding the encounter-level capacity fixed, bundling activity into larger encounters lowers the screening intensity per unit, reducing detectability. The mechanism is most relevant when per-encounter screening time is capped and escalation thresholds are absent or binding.

Given n interactions of equal size $y = Y/n$, the total probability of detecting a violation is

$$P(n) = 1 - \left(1 - p\left(\frac{Y}{n}\right)\right)^n. \tag{11}$$

For a small $p(Y/n)$, $P(n) = 1 - (1 - p(Y/n))^n \approx n\, p(Y/n)$,

which can be approximated by $P(n) \approx n \cdot p(Y/n)$ when detection probabilities are small.

Theorem 1 (Paradox of tightening)**.** *Suppose Assumption 4 holds and the agent chooses an encounter frequency optimally, as in Section 3. Then there exist values of the exposure schedule at which a tightening (a higher per-encounter procedural burden) reduces the regulator's equilibrium probability of detection.*

Table 3 reports a numerical illustration. As the optimal encounter size rises from 20 to 25, the internal processing cost per encounter increases from $g(20) = 400$ to $g(25) = 625$ (under $g(y) = y^2$). In the illustration, the per-encounter exposure parameter increases from 0.40 to 0.50, inducing a discrete drop in the optimal encounter count from 5 to 4 and reducing total detection from 4.90% to 2.18%.

Interpretation. The mechanism operates through capacity-constrained screening: tighter regimes induce consolidation into fewer, larger encounters, which can reduce detectability (a capacity-dilution channel). When this size dependence is strong, detection can become non-monotonic in scrutiny; when it is weak, consolidation still counteracts stricter screening and attenuates marginal gains from increased scrutiny.

Design corollary. The paradox channel requires that per-encounter screening resources do not scale proportionally with encounter scope. If the agency adopts within-encounter sampling, automated checks that scale with volume, or size-contingent escalation such that effective capacity becomes $B = B(y)$ with $B'(y) > 0$ and near-proportional scaling over the relevant range, then $p(y)$ need not decline with y and the non-monotonic detection region shrinks or disappears.

4. Enforcement Architecture Under Capacity Constraints

This section links the model to administrative design. Building on Assumption 4, screening capacity is bounded at the encounter level, so screening intensity per unit is $e(y) = B/y$ and detection probability is $p(y) = \pi(B/y)$. The

regulator takes the total activity, Y, as given and chooses rules for how activity is partitioned across encounters.

4.1. Objective and Implementable Instruments

Let the agent allocate the total activity Y across n encounters with sizes $\{y_i\}_{i=1}^n$, $y_i > 0$, $\sum_{i=1}^n y_i = Y$. Given encounter-level detection probabilities $p(y_i)$, the probability of at least one detection over the horizon is

$$D(\{y_i\}_{i=1}^n) \equiv 1 - \prod_{i=1}^{n} \big(1 - p(y_i)\big). \tag{12}$$

I treat $D(\cdot)$ as the regulator's benchmark objective and focus on two administratively common instruments:

1. **Encounter size caps:** impose $y_i \leq \bar{y}$ for all i.
2. **Frequency requirements:** impose $n \geq n_{\min}$, where $n_{\min} \in \mathbb{N}$ is a policy parameter (equivalently, an upper bound on the reporting interval).

These instruments are distinct from reforms that increase penalties or raise per-encounter procedural requirements. In the model, higher per-encounter stringency raises exposure costs and can induce bundling, thereby reducing the number of encounters and increasing their size.

4.2. A Design Result: Encounter Caps and Detection

Proposition 6 below formalizes how encounter architecture can improve detection when screening capacity is bounded.

Proposition 6 (Encounter capping weakly increases detection)**.** *Suppose detection is capacity-constrained as in Assumption 4 and $p(y) \in (0,1)$ for all $y > 0$. Define*

$$\ell(y) \equiv -\log\big(1 - p(y)\big).$$

A standard sufficient condition for "risk reductions" to be concave in encounter size is that $\ell(\cdot)$ is concave on $(0,\infty)$. Fix total activity Y and consider any feasible allocation $\{y_i\}_{i=1}^n$ with $\sum_{i=1}^n y_i = Y$. Let $\bar{y} \in (0, Y]$ and construct

$\{\tilde{y}_j\}_{j=1}^{\tilde{n}}$ *by splitting any encounter with* $y_i > \bar{y}$ *into* $\lceil y_i/\bar{y} \rceil$ *encounters that sum to* y_i *and satisfy* $\tilde{y}_j \le \bar{y}$. *Then*

$$D(\{\tilde{y}_j\}_{j=1}^{\tilde{n}}) \;\ge\; D(\{y_i\}_{i=1}^{n}),$$

where $D(\cdot)$ *is defined in* (12). *If* $\ell(\cdot)$ *is strictly concave and at least one encounter is split nontrivially, the inequality is strict.*

Proof. Write the probability of no detection under allocation $\{y_i\}_{i=1}^{n}$ as

$$S(\{y_i\}) \equiv \prod_{i=1}^{n} \big(1 - p(y_i)\big), \qquad \text{so that} \qquad D(\{y_i\}) = 1 - S(\{y_i\}).$$

Consider splitting a single encounter of size $y > 0$ into $m \ge 2$ parts $\{y^{(1)}, \ldots, y^{(m)}\}$ with $\sum_{j=1}^{m} y^{(j)} = y$. Taking logs,

$$-\log\left(\prod_{j=1}^{m} (1 - p(y^{(j)}))\right) = \sum_{j=1}^{m} \ell(y^{(j)}).$$

By concavity of ℓ, for each j,

$$\ell(y^{(j)}) \;\ge\; \frac{y^{(j)}}{y}\, \ell(y),$$

so summing over j yields $\sum_{j=1}^{m} \ell(y^{(j)}) \ge \ell(y)$. Exponentiating implies

$$\prod_{j=1}^{m} (1 - p(y^{(j)})) \;\le\; 1 - p(y).$$

Thus splitting weakly reduces the no-detection probability for that encounter. Applying the same argument to every encounter larger than $\bar{y}$ shows $S(\{\tilde{y}_j\}) \le S(\{y_i\})$, hence $D(\{\tilde{y}_j\}) \ge D(\{y_i\})$.

If ℓ is strictly concave and at least one split is nontrivial, at least one inequality above is strict, so $S(\{\tilde{y}_j\}) < S(\{y_i\})$ and therefore $D(\{\tilde{y}_j\}) > D(\{y_i\})$. □

4.3. Architecture, Sampling, and Scalable Screening

Encounter caps and frequency requirements are extreme forms of a more general principle: screening intensity should be preserved per unit when activity is consolidated. Flexible implementation can be accomplished through

within-encounter sampling or *size-contingent screening*. For large encounters, the regulator can apply a rule in terms of which a subset of subcomponents (e.g., line-item audits, random container checks, targeted deep dives) is inspected. This effectively replaces $p(y) = \pi(B/y)$ by

$$p(y) = \pi\big(B/y\big) \text{ with } B = B(y),$$

where $B(y)$ increases with encounter size because automated checks, analyst time, or inspection slots scale with volume. In this sense, scalable screening technologies and sampling rules mitigate the capacity-dilution problem without necessarily imposing hard caps on y.

4.4. Policy Tradeoff and Evaluation Criterion

The model highlights a tradeoff between (i) *procedural exposure*–the per-encounter burden that discourages frequent encounters—and (ii) *informational preservation*—the benefits of more frequent, smaller encounters when screening is capacity-constrained. A practical criterion for evaluating reforms follows: reforms that increase the per-encounter procedural burden should be paired with either (a) a reporting architecture that preserves encounter frequency (e.g., lightweight periodic micro-reporting with occasional deep dives), or (b) screening technologies that scale with encounter size, so that effective screening intensity per unit does not collapse under consolidation. Otherwise, stricter rules can mechanically shift activity into forms that are harder to screen. I discuss design implications and boundary conditions in Section 5, and empirical signatures and falsification strategies in Section 6.

4.5. Policy Counterfactuals Under Capacity Constraints

This subsection states the policy counterfactual that follows from the model when screening capacity is allocated at the encounter level (Assumption 4). The comparison is between reforms that increase per-encounter burdens and reforms that change encounter architecture, holding total activity Y fixed.

Counterfactual comparison. A reform that increases per-encounter procedural burden shifts the exposure schedule upward and induces consolidation

(Proposition 2). Under encounter-level capacity allocation (Assumption 4), consolidation increases encounter scope and can reduce equilibrium detectability when screening does not scale with scope (Theorem 1).

Architectural reforms act on the encounter process directly: encounter size caps $y_i \leq \bar{y}$, minimum frequency requirements $n \geq n_{\min}$, and scalable screening via sampling or automation (i.e., $B = B(y)$ with $B'(y) > 0$). These instruments preserve screening intensity per unit and limit the scope for information collapse.

5. Discussion: Administrative Design under Capacity Constraints

The analysis yields a policy-relevant design implication: when screening capacity is binding, regulators can improve detection by shaping encounter architecture rather than relying solely on higher penalties or stricter per-encounter requirements. I therefore characterize optimal encounter design under capacity constraints. In particular, I show that policies imposing an upper bound on encounter size or a minimum encounter frequency weakly increase detection and can strictly dominate increases in per-encounter stringency when detection is sufficiently concave in screening intensity. This provides an economic rationale for administrative rules such as filing frequency requirements, staggered reporting requirements, and shipment size caps in settings where enforcement agents face congestion and limited screening resources.

The model isolates a margin that is often treated as fixed in enforcement analyses, allowing its implications to be examined transparently. It holds total activity fixed and focuses on its organization across regulatory encounters. That choice isolates a margin that is easy to overlook: agents can adjust both what happens within a filing or inspection *and* how often those encounters occur. The framework employed here complements deterrence models by highlighting a margin, interaction frequency, that is typically taken as exogenous. When detectability depends on interaction size, this reorganization also affects what regulators can learn, not just how often agents appear.

Boundary conditions and scope. The mechanism is most relevant when

regulated parties have discretion over the timing or partitioning of activity, and when the marginal burden of an additional encounter is nontrivial (procedural exposure, scrutiny, or follow-up costs). If encounter timing is fixed by statute or technologically constrained, enforcement operates primarily by reducing aggregate activity over the horizon rather than by reshaping encounter structure. However, many policy environments combine both margins: penalties affect the level of activity, while the procedural burden affects how activity is distributed across dates or channels, even when the particular frequency is mandated. The 'paradox' channel additionally requires that detectability declines sufficiently with encounter size or complexity; it is muted if screening resources scale proportionally with encounter size or if regulators can sample within large encounters such that effective detection does not fall mechanically with consolidation. Empirically, the mechanism is most plausible when enforcement shocks shift the joint distribution of encounters, encounter size, and spacing, not only aggregates. The key takeaway is that enforcement can change observability by changing the encounter process itself, and this can alter how enforcement effectiveness appears in administrative data.

6. An Illustrative Empirical Exhibit

This section provides a small empirical exhibit to illustrate the encounter-organization margin emphasized by the model. The goal is not causal evaluation or policy quantification. Rather, the exhibit shows how the model's objects—encounter frequency, conditional encounter size, and aggregate volume—map into administrative data, and how changes in per-encounter procedural burden can be reflected in observed outcomes even when underlying activity is comparatively stable. The institutional meaning of the procedural regime indicators and the corresponding changes in per-encounter requirements are documented in Appendix A.5. We exploit data on changes in client–bank interactions (i.e., number of transfers, and volume per transfer) in cross–border payment when KYC requirements change.

Setting and definition of procedural tightening. The exhibit draws on a panel of regulated customers observed over time in an administrative monitoring environment where interactions occur through discrete encounters. Periods of heightened procedural stringency are defined by institutional changes that increased fixed per-encounter requirements—such as additional documentation, validation steps, or follow-up obligations—without an explicit increase in screening capacity per encounter. These changes are treated as regime shifts in per-encounter burden and are coded once using discrete regime indicators. No attempt is made to exploit sharp eligibility thresholds or randomized assignment.

Mapping from the model to observables. In the model, agents allocate a fixed activity volume Y across a chosen number of encounters n, generating an average encounter scope Y/n. In the data, the corresponding objects are: (i) the number of encounters per period, (ii) total processed volume, and (iii) conditional mean encounter size among periods with positive activity. The consolidation mechanism predicts a joint response: following a tightening that raises per-encounter exposure, encounter counts fall and conditional encounter size rises, even if aggregate volume does not change proportionally.

Illustrative regression evidence. Table 4 reports within-customer regressions of encounter counts, aggregate volume, conditional encounter size, and a concentration measure on the tightening regime indicators. The specification includes customer fixed effects and sector–year controls, so the estimates reflect within-unit adjustments around the regime changes. The table is intended to summarize directional responses rather than establish causality.

Tightening regimes are associated with lower encounter counts and higher conditional encounter size, consistent with consolidation. I further find that aggregate volume responds weakly, suggesting re-partitioning rather than a uniform contraction in underlying activity. Interaction terms indicate that these adjustments are more pronounced for customers with greater prior exposure to the monitoring process (i.e., more payments performed during a strict regime),

but the baseline patterns are present even without conditioning on heterogeneity.

Counterfactual interpretation. To visualize how the estimated coefficients translate into time paths, Figure 4 plots the observed evolution of encounter frequency together with two model-implied counterfactuals. The first removes the policy regime indicators, yielding an implied path in the absence of procedural tightening. The second retains the regime timing but removes exposure-related interactions, illustrating the role of endogenous reorganization. These counterfactuals are constructed mechanically from the fitted linear predictor while holding customer fixed effects and controls constant; they do not extrapolate beyond the support of the data.

The figure highlights that observed encounter frequency lies between the two counterfactual paths: tightening reduces encounters relative to the no-policy benchmark, but adaptive reorganization attenuates the decline. Read together with Table 4, the exhibit illustrates the central mechanism of the paper: higher per-encounter burden can reshape the encounter process—fewer encounters with larger scope—without requiring a large change in total activity. When screening capacity is allocated at the encounter level, this reorganization has direct implications for observability and effective detection, as discussed in Section 3.10.

Scope and interpretation. The exhibit is intentionally narrow. It does not attempt to measure enforcement effectiveness or to identify treatment effects under strong assumptions. Its purpose is to show that the objects emphasized in the model—encounter frequency and conditional encounter size—move in administrative data in ways consistent with exposure-driven consolidation, and that focusing on encounter counts alone can be misleading when agents can reorganize activity across encounters.

7. Conclusion

This paper has considered a simple margin that enforcement models often treat as fixed: how frequently regulated agents interact with authorities. In

many settings, the main burden of enforcement is not the expected penalty per unit of activity, but the cost of repeated exposure in the form of preparation, documentation, review, and the possibility of follow-up. When past contact informs future scrutiny, a return encounter can be more costly than the first. Agents then have a clear incentive to reorganize activity to appear less often, even if this does not reduce the total activity.

The model formalizes this idea by letting an agent with fixed total activity choose how to partition that activity across encounters. Internal processing costs make very large batches costly, while exposure costs rise with the number of encounters. The resulting trade-off delivers a sharp prediction: when the marginal burden of an additional encounter increases, the agent chooses fewer encounters and larger average size per-encounter. If encounters are time-indexed, this mechanically generates retiming and clustering, with longer gaps followed by bursts. These outcomes arise under conditions of full rationality; temporary deterrence, biased beliefs, or other behavioral assumptions are not required.

This analysis has underscored a fundamental tension in modern enforcement systems: policies that increase scrutiny through repeated interactions may succeed in disciplining behavior, but they can also lead agents to reduce contact and consolidate activity, thereby complicating oversight. Treating encounter frequency as an endogenous choice helps explain lumpy post-enforcement patterns as rational adaptations and reveals how enforcement reshapes the organization of economic activity without necessarily altering its scale.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used ChatGPT 5.2 for code editing, structural assistance, and LaTeX formatting. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

Allingham, Michael G., and Agnar Sandmo. 1972. "Income Tax Evasion: A Theoretical Analysis." *Journal of Public Economics*, 1 (3–4): 323–338.

Almunia, Miguel, and David Lopez-Rodriguez. 2018. "Under the Radar: The Effects of Monitoring Firms on Tax Compliance." *American Economic Journal: Economic Policy*, 10 (1): 1–38.

Baumol, William J. 1952. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Quarterly Journal of Economics*, 66 (4): 545–556.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 76 (2): 169–217.

Caballero, Ricardo J., and Eduardo M. R. A. Engel. 1999. "Explaining Investment Dynamics in U.S. Manufacturing: A Generalized (S,s) Approach." *Econometrica*, 67 (4): 783–826.

Choudhary, Keshav, and Bhanu Gupta. 2025. "Dynamics of Evasion around Tax Thresholds: Evidence from Indian Firms." *Journal of Public Economics*, 252, 105535.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *Quarterly Journal of Economics*, 128 (4): 1499–1545.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 129 (4): 1661–1710.

Garicano, Luis, Claire Lelarge, and John Van Reenen. 2016. "Firm Size Distortions and the Productivity Distribution: Evidence from France." *American Economic Review*, 106 (11): 3439–3479.

Gray, Wayne B., and Jay P. Shimshack. 2011. "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence." *Review of Environmental Economics and Policy*, 5 (1): 3–24.

Harrington, Winston. 1988. "Enforcement Leverage When Penalties Are Restricted." *Journal of Public Economics*, 37 (1): 29–53.

Heyes, Anthony, and Neil Rickman. 1999. "Regulatory Dealing—Revisiting the Harrington Paradox." *Journal of Public Economics*, 72 (3): 361–378.

Kleven, Henrik J. 2016. "Bunching." *Annual Review of Economics*, 8: 435–464.

Kleven, Henrik J., Martin B. Knudsen, Claus T. Kreiner, Søren Pedersen, and Emmanuel Saez. 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica*, 79 (3): 651–692.

Kleven, Henrik J., and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Quarterly Journal of Economics*, 128 (2): 669–723.

Maćkowiak, Bartosz, Filip Matějka, and Mirko Wiederholt. 2023. "Rational Inattention: A Review." *Journal of Economic Literature* 61 (1): 226–273.

Matějka, Filip, and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review*, 105 (1): 272–298.

Milgrom, Paul, and Chris Shannon. 1994. "Monotone Comparative Statics." *Econometrica*, 62 (1): 157–180.

Moynihan, Donald P., Pamela Herd, and Hope Harvey. 2015. "Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen–State Interactions." *Journal of Public Administration Research and Theory*, 25 (1): 43–69.

Nyborg, Karine, and Kjetil Telle. 2006. "Firms' Compliance to Environmental Regulation: Is There Really a Paradox?" *Environmental and Resource Economics*, 35 (1): 1–18.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy*, 115 (2): 200–249.

Polinsky, A. Mitchell, and Steven Shavell. 2000. "The Economic Theory of Public Enforcement of Law." *Journal of Economic Literature*, 38 (1): 45–76.

Pomeranz, Dina. 2015. "No Taxation without Information: Deterrence and Self–Enforcement in the Value Added Tax." *American Economic Review*, 105 (8): 2539–2569.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–690.

Shimshack, Jay P. 2014. "The Economics of Environmental Monitoring and Enforcement." *Annual Review of Resource Economics*, 6: 339–360.

Slemrod, Joel. 2019. "Tax Compliance and Enforcement." *Journal of Economic Literature* 57 (4): 904–954.

Stigler, George J. 1970. "The Optimum Enforcement of Laws." *Journal of Political Economy* 78 (3): 526–536.

Telle, Kjetil. 2013. "Monitoring and Enforcement of Environmental Regulation." *Journal of Public Economics*, 99: 24–34.

Tobin, James. 1956. "The Interest–Elasticity of Transactions Demand for Cash." *The Review of Economics and Statistics*, 38 (3): 241–247.

Topkis, Donald M. 1978. "Minimizing a Submodular Function on a Lattice." *Operations Research*, 26 (2): 305–321.

Zhou, Qi, Shen Qu, Qianzi Wang, Yunlei She, Yang Yu, and Jun Bi. 2023. "Sliding Window–Based Machine Learning for Environmental Inspection Resource Allocation." *Environmental Science & Technology*, 57 (44): 16743–16754.

Tables and figures

Table 1: Persistence of consolidation: illustrative two-period example

Period	Scrutiny level (s_t)	Optimal encounters (n_t^*)	Avg. size (Y/n_t^*)
1	High (s_H)	5	20.0
2	Low (s_L)	6	16.7

Notes: Total activity is normalized to $Y = 100$. Period 1 has higher scrutiny, inducing consolidation ($n_1^* = 5$). If scrutiny partially carries over, the period-2 optimum can remain below the low-scrutiny benchmark even when s_t reverts to s_L.

Table 2: Total cost by encounter count under encounter-only and size-contingent monitoring

n	Y/n	$\Phi_{\mathrm{EO}}(n)$	$\Phi_{\mathrm{SC}}(n)$
2	50.0	63.00	80.00
3	33.3	48.67	60.00
4	25.0	43.50	52.00
5	20.0	**42.00**	48.80
6	16.7	42.33	**48.00**
7	14.3	43.71	48.57
8	12.5	45.75	50.00
9	11.1	48.22	52.00
10	10.0	51.00	54.40

Notes: $Y = 100$. Encounter-only objective: $\Phi_{\mathrm{EO}}(n) = aY^2/n + c_0 n + \frac{\kappa}{2} n(n-1)$ with $a = 0.011$, $c_0 = 4$, $\kappa = 0$. Size-contingent objective adds $\tau\, n\, (Y/n)^{\beta}$ with $\tau = 0.0034$ and $\beta = 2$. Bold values denote the discrete minima (markers in Figure 2).

Table 3: The paradox of tightening: higher procedural burden can reduce detection (illustration)

Regime	Exposure cost s	n^*	Y/n^*	$p(Y/n^*)$	$D(n^*)$
A (moderate)	0.40	5	20.0	0.0100	0.0490
B (stricter)	0.50	4	25.0	0.0055	0.0218

Notes: Detectability $p(y)$ declines with encounter size y under capacity dilution. The tightening $s : 0.40 \to 0.50$ induces consolidation ($n^* : 5 \to 4$), increasing Y/n^* and lowering both $p(Y/n^*)$ and total detection $D(n^*)$.

Table 4: Regime tightening and encounter consolidation outcomes

	(1) log(1 + Transfers)	(2) asinh(Volume)	(3) log(MeanSize)	(4) HHI
R1_tightL1	-0.043^{***}	-0.002	0.055^{***}	-0.013^{***}
	(0.007)	(0.013)	(0.008)	(0.003)
R2_tightL2	-0.081^{***}	-0.023	0.084^{***}	-0.014^{***}
	(0.009)	(0.016)	(0.010)	(0.004)
R3_returnL1	-0.099^{***}	-0.069^{***}	0.062^{***}	-0.007
	(0.012)	(0.021)	(0.012)	(0.005)
NewClient	0.036^{***}	0.056^{***}	0.010	-0.006^{*}
	(0.008)	(0.014)	(0.008)	(0.003)
R1_tightL1 × ExperiencedStrict	0.001	-0.021^{*}	-0.022^{***}	0.025^{***}
	(0.007)	(0.012)	(0.007)	(0.003)
R2_tightL2 × ExperiencedStrict	-0.011^{*}	-0.049^{***}	-0.037^{***}	0.035^{***}
	(0.007)	(0.012)	(0.007)	(0.003)
R3_returnL1 × ExperiencedReturn	-0.027^{***}	-0.067^{***}	-0.034^{***}	0.033^{***}
	(0.009)	(0.017)	(0.010)	(0.004)
Customer fixed effects	Yes	Yes	Yes	Yes
Sector–year controls	Yes	Yes	Yes	Yes

Notes: Estimates are from within-customer regressions with sector–year controls. Standard errors clustered at the customer level are in parentheses. The table summarizes directional responses to increases in per-encounter procedural burden; coefficients are descriptive and not interpreted as causal treatment effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

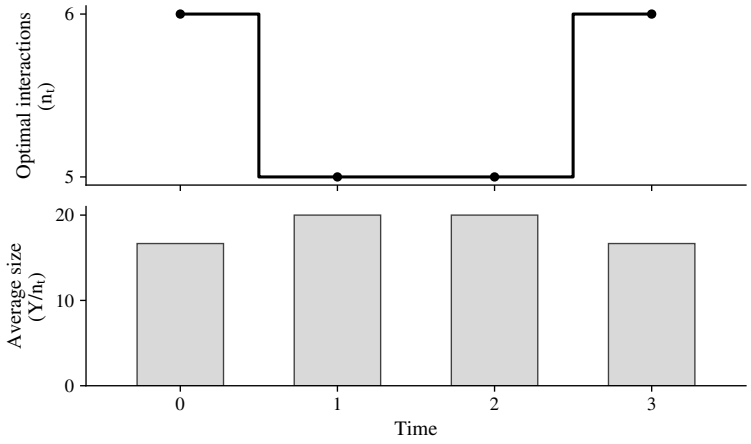


Figure 1: Optimal interaction frequency and average interaction size over time

Optimal interaction frequency (line) and average interaction size (bars) around a temporary scrutiny shock at $t = 1$.

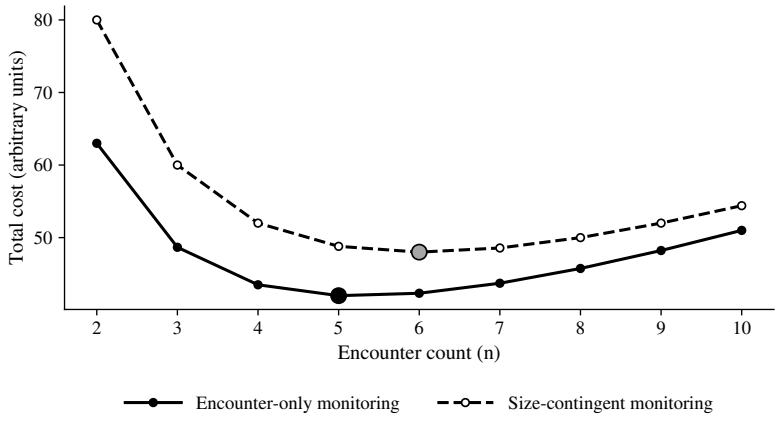


Figure 2: Total cost under encounter-only and size-contingent monitoring

Total cost as a function of the interaction count under encounter-only (solid) and size-contingent (dashed) monitoring. Markers denote optimal choices.

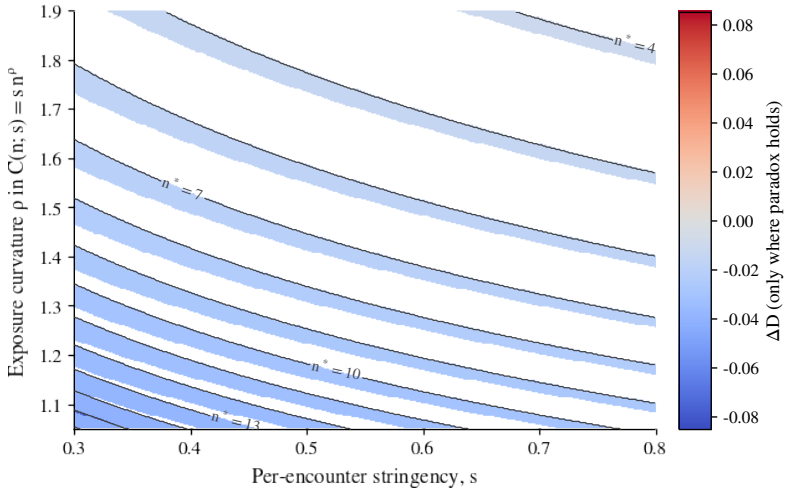


Figure 3: Paradox-of-tightening region and consolidation thresholds.

Detection $D(n) = 1 - (1 - p(Y/n))^n$; shading shows ΔD only where both n^* falls and detection declines under a tightening $s \to s + \Delta s$ (with $\Delta s = 0.05$ in the illustration).

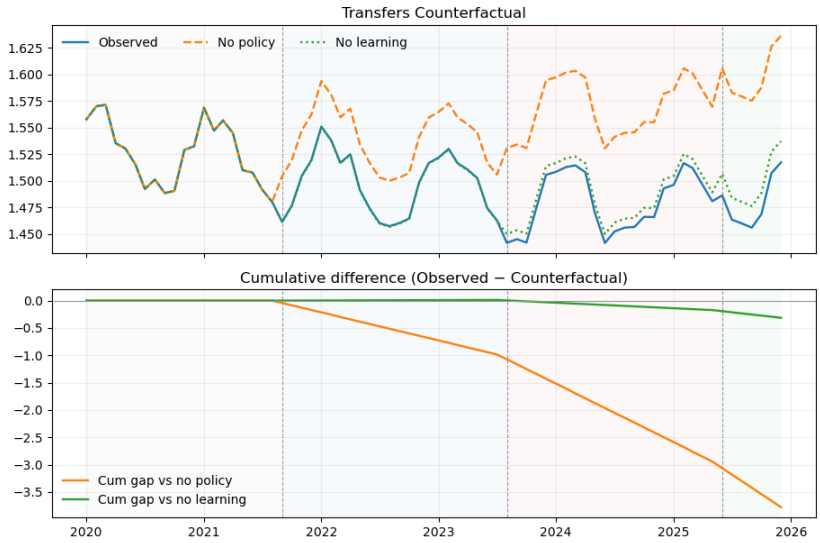


Figure 4: Counterfactual encounter frequency under procedural tightening

Observed encounter frequency is shown together with two model-implied counterfactual paths. The *no-policy* counterfactual removes regime indicators from the fitted linear predictor, while the *no-learning* counterfactual additionally removes exposure-related interaction terms. Customer fixed effects and sector–year controls are held constant. The figure illustrates how procedural tightening reshapes encounter frequency through consolidation rather than through a proportional reduction in aggregate activity.

Supplementary Material

This appendix collects supplementary material supporting the main text. Appendix A provides an extended mapping, examples, and discussion of alternative exposure schedules. Appendix B collects auxiliary results and derivations referenced in the text.

Appendix A. Extended Framework Mapping and Examples

Appendix A.1. Additional note on data-driven targeting

Capacity-constrained targeting can be framed as an administrative resource-allocation problem; data-driven tools are often used to prioritize high-risk cases under binding staffing constraints (Zhou et al., 2023).

Appendix A.2. Extended mapping: discrete regimes and reorganization around thresholds

A central implication of the paper is that monitoring design can induce reorganization of encounters around discrete regimes, altering the mapping from observed administrative events to underlying activity. When targeting relies on risk scores or cutoffs constructed from reported information, agents may change encounter timing, scope, or frequency in ways that shift measured risk without commensurate changes in underlying behavior. Choudhary and Gupta (2025), for instance, study dynamic responses around a size-based registration threshold; while their setting differs, the shared implication is that monitoring rules can change observables through reorganization and retiming.

Appendix A.3. Robustness to alternative exposure-cost schedules

The consolidation comparative static in Proposition 2 does not rely on a narrow functional form for the exposure schedule. To make this explicit, let the marginal exposure cost of the kth encounter be $c(k)$, so that $C(n) = \sum_{k=1}^{n} c(k)$. The key requirement is that $c(k)$ is nondecreasing over the relevant range (Assumption 2). The following stylized cases illustrate common institutional patterns:

1. **Constant marginal exposure.** $c(k) = \bar{c} > 0$ implies $C(n) = \bar{c}\, n$.
2. **Linear escalation.** $c(k) = c_0 + b(k-1)$ with $b > 0$ increases the marginal burden of additional encounters and shifts the optimum toward fewer encounters $n^\star$ and larger average size $Y/n^\star$.
3. **Convex escalation.** A convex $c(\cdot)$ produces sharper (discrete) reductions in $n^\star$ as policy increases the slope of marginal exposure around the pre-change optimum.
4. **Concave escalation / saturation.** $c(k) = c_0 + b(1 - e^{-\lambda(k-1)})$ with $\lambda > 0$ captures early escalation followed by saturation. Consolidation persists as long as $c(\cdot)$ is nondecreasing over the relevant range.
5. **Threshold escalation (trigger rules).** $c(k) = c_L$ for $k \leq K$ and $c(k) = c_H$ for $k > K$ with $c_H > c_L$ captures institutional triggers (repeat amendments, follow-ups, escalation rules) and implies discontinuous shifts in $n^\star$ and potentially strong bunching in encounter counts.

Magnitude. The size of the consolidation response depends on how much the marginal exposure schedule changes *around the pre-change optimum.* In discrete-choice problems, small policy changes can have large effects when they move the optimum across a switching threshold.

Appendix A.4. Batching logic and "ordeals"

The benchmark mechanism—partitioning a fixed total Y across a chosen number of encounters—is closely related to classic batching and lumpy-adjustment logic. In inventory, cash-management, and adjustment-cost models, agents trade off fixed or convex costs against the benefits of smoothing, generating discrete changes in frequency and batch size (Baumol, 1952; Tobin, 1956; Caballero and Engel, 1999). The present model applies this logic to administrative interactions, interpreting them as an endogenous *encounter process* rather than a fixed schedule.

A related conceptual link is "ordeal" design: when costs attach to repeated interactions, policies can screen or deter by increasing procedural burdens rather

than monetary penalties. This paper differs in emphasis by focusing on equilibrium reorganization of encounters and the resulting implications for what regulators observe.

Appendix A.5. Institutional background and definition of procedural regimes in the empirical exhibit

This appendix explains the institutional content of the procedural-regime indicators used in Section 6. The goal is to be explicit about what constitutes a "tightening" in this setting and why the regime shifts are best interpreted as increases in *per-encounter procedural burden*, rather than as changes in statutory penalties, mandated encounter frequency, or screening capacity.

What changes across regimes. The regime indicators capture discrete changes in internal administrative requirements governing how clients complete cross-border payments subject to the bank's compliance and KYC workflow. Documentation requirements tightened in 2021 and again in 2023, and in 2025 the bank reverted to (approximately) the documentation level introduced in the 2021 tightening, yielding three distinct regimes. Throughout, the statutory framework, sanction schedules, and formal external reporting obligations remained unchanged. The salient variation is therefore procedural: a higher fixed administrative load per interaction with the compliance process.

Concretely, the changes included additional purpose-of-transfer fields, separate declaration forms, expanded validation and consistency checks, stricter formatting or submission rules, and a higher likelihood of follow-up requests conditional on client interaction. These adjustments increase the time, attention, and coordination costs of completing an encounter, without mechanically increasing the number of encounters or imposing proportional requirements that scale one-for-one with the amount processed.

Why this maps to per-encounter exposure in the model. In the model, procedural burden is an exposure cost incurred each time an encounter occurs and, potentially, compounded by repeated interaction. The institutional changes

above align with this interpretation: they raise the fixed cost of engaging with the compliance workflow largely independent of encounter scope. Importantly, the bank did not introduce minimum-frequency requirements or rules that force activity to be split into more encounters. Agents therefore retained discretion over how to partition a given level of activity across encounters.

What did not *change: penalties and screening capacity.* A maintained assumption in the model is that screening capacity is allocated at the encounter level and does not scale proportionally with encounter scope. Consistent with this, the procedural shifts did not coincide with discrete increases in staffing, analyst hours, or automated screening capacity *per encounter*. Nor were size-contingent escalation rules introduced that would mechanically allocate more screening resources to larger encounters. Screening technology may have evolved over the sample, but such evolution appears gradual and not synchronized with the regime breakpoints captured by the indicators.

Coding of regime indicators. The empirical analysis distinguishes three periods. Regime 1 corresponds to the baseline procedural environment. Regime 2 denotes an initial tightening that increased fixed per-encounter requirements. Regime 3 denotes a subsequent adjustment that relaxed some procedural elements while leaving others in place. The indicators are coded at the period level and apply uniformly across agents. The analysis does not exploit sharp eligibility thresholds or differential implementation across units; the regime indicators are used only to mark periods with systematically higher or lower per-encounter burden.

Scope and limitations. This institutional account supports interpreting the empirical exhibit as reflecting exposure-driven consolidation rather than deterrence via reduced underlying activity. The regime changes are not treated as exogenous policy experiments, and the analysis does not attempt to cleanly separate procedural burden from other contemporaneous developments. Accordingly, robustness checks focus on ruling out mechanical timing artifacts around the breakpoint months. In particular, the baseline specification is re-estimated

after excluding windows around each regime change (Table A.5). The estimated associations in encounter frequency persist under these exclusions, which is more consistent with gradual and persistent reorganization than with a purely mechanical discontinuity at the implementation date. Overall, the appendix clarifies why the regime indicators plausibly map to the model's per-encounter exposure parameter, while stopping short of causal claims.

Appendix B. Mathematical Appendix

Appendix B.1. Monotone comparative statics for encounter consolidation

I illustrate a lattice-theoretic monotonicity condition that delivers the weak consolidation result invoked in the main text.

Let $s \in \mathcal{S} \subseteq \mathbb{R}$ index policy stringency and allow the exposure cost to depend on s. The agent chooses $n \in \mathbb{N}$ to minimize

$$V(n; s) \equiv n\, g(Y/n) + C(n; s). \tag{B.1}$$

Assume:

(i) (*Existence*) For each s, $V(n; s) \to \infty$ as $n \to \infty$.

(ii) (*Increasing differences / single crossing*) For all n and $s_2 > s_1$,

$$\Delta(n; s) \equiv V(n+1; s) - V(n; s) \;\text{ satisfies }\; \Delta(n; s_2) \geq \Delta(n; s_1), \tag{B.2}$$

with strict inequality for at least one n for any nontrivial increase in s.

Proposition 7 (Monotone comparative statics (weak consolidation))**.** *Let* $n^\star(s) \in \arg\min_{n\in\mathbb{N}} V(n; s)$. *Under* ***(i)****–****(ii)****, the smallest selection from* $\arg\min_n V(n; s)$ *is weakly decreasing in* s. *In particular, higher stringency weakly reduces the optimal number of encounters.*

Proof. Condition **(ii)** is increasing differences on the lattice $\mathbb{N} \times \mathcal{S}$ in discrete-increment form. By Topkis' monotonicity theorem for minimization on lattices, the smallest minimizer is weakly decreasing in s (Topkis, 1978). □

Table A.5: Donut robustness around regime transitions

	Donut window (months excluded around each break)		
	$k = 0$	$k = 1$	$k = 2$
Regime 1 (Tightening I)	-0.0433^{***}	-0.0295^{***}	-0.0318^{***}
	(0.0073)	(0.0089)	(0.0116)
Regime 2 (Tightening II)	-0.0806^{***}	-0.0596^{***}	-0.0509^{***}
	(0.0092)	(0.0112)	(0.0139)
Regime 3 (Partial return)	-0.0991^{***}	-0.0766^{***}	-0.0618^{***}
	(0.0115)	(0.0139)	(0.0168)
Regime 1 × Experienced strict	0.0007	0.0026	0.0023
	(0.0065)	(0.0069)	(0.0071)
Regime 2 × Experienced strict	-0.0114^{*}	-0.0114^{*}	-0.0140^{*}
	(0.0066)	(0.0069)	(0.0072)
Regime 3 × Experienced return	-0.0274^{**}	-0.0327^{**}	-0.0368^{***}
	(0.0092)	(0.0105)	(0.0116)
Observations	179,371	157,554	142,826
Customers	3,891	3,891	3,891
Months	72	63	57

Notes: The table reports fixed-effects estimates of encounter frequency (log number of transfers) under increasing exclusion windows around each regime breakpoint. k denotes the number of months excluded before and after each regime change. Standard errors clustered at the customer level are reported in parentheses. Larger exclusion windows (e.g., $k \geq 3$) remove substantial portions of the regime-defining periods and render some regime indicators weakly identified; these specifications are therefore not reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Appendix B.2. Size-contingent monitoring limits consolidation

Proof of Proposition 5. Fix $\kappa \geq 0$. Under strict convexity of g and convexity of m, the Jensen argument used for Lemma 1 applies to $g + \kappa m$, so conditional on n the unique minimizer is $y_i = Y/n$ for all i. The reduced objective is

$$V(n;\kappa) = H(n) + C(n) + \kappa T(n), \quad \text{where} \quad H(n) \equiv ng(Y/n),\ \ T(n) \equiv nm(Y/n).$$

Define the forward difference $\Delta(n;\kappa) \equiv V(n+1;\kappa) - V(n;\kappa)$:

$$\Delta(n;\kappa) = \underbrace{\big(H(n+1) - H(n)\big) + c(n+1)}_{\Delta_0(n)} + \kappa\ \underbrace{\big(T(n+1) - T(n)\big)}_{\Delta_T(n)}.$$

By Assumption 1, $H(n)$ is convex in n on $\mathbb{N}$, so $H(n+1) - H(n)$ is nondecreasing in n. By Assumption 2, $c(n+1)$ is nondecreasing in n, hence $\Delta_0(n)$ is nondecreasing. By the assumed convexity of $T(n)$ on $\mathbb{N}$, $\Delta_T(n)$ is also nondecreasing in n. Therefore, for each κ, $\Delta(n;\kappa)$ is nondecreasing in n, implying the set of minimizers is an interval and the smallest minimizer $\underline{n}(\kappa)$ is characterized by the first index where $\Delta(n;\kappa) \geq 0$.

Since $T(n)$ is decreasing, $\Delta_T(n) = T(n+1) - T(n) \leq 0$ for all n. Thus for any fixed n, $\Delta(n;\kappa)$ is weakly decreasing in κ. Increasing κ shifts the sequence $\{\Delta(n;\kappa)\}_n$ downward, so the crossing point where it becomes nonnegative occurs at a weakly larger n. Hence $\underline{n}(\kappa)$ is weakly increasing in κ, and in particular $\underline{n}(\kappa) \geq \underline{n}(0)$ for all $\kappa \geq 0$. □

Appendix B.3. Benchmark where encounter frequency is irrelevant

A useful benchmark clarifies why many models treat encounter timing and frequency as exogenous. Suppose expected enforcement costs are linear in total activity and independent of encounter history, so expected enforcement cost equals λY with $\lambda \geq 0$. Then expected enforcement cost is invariant to the partition $\{y_i\}_{i=1}^n$ and to the encounter count n. Encounter frequency becomes relevant only when exposure depends on the encounter process itself, e.g., through per-encounter overhead, history dependence, or state-dependent scrutiny.

Appendix B.4. Parametric illustration of the consolidation channel

To make the magnitude channel explicit, consider $g(y) = y^2$ and $C(n) = \alpha n^\rho$ with $\alpha > 0$ and $\rho \geq 1$. Then

$$V(n) = n\, g(Y/n) + C(n) = \frac{Y^2}{n} + \alpha n^\rho.$$

Treating n as continuous for approximation, the first-order condition yields

$$n^\star \approx \left(\frac{Y^2}{\alpha\rho}\right)^{\frac{1}{\rho+1}}.$$

Thus higher α or ρ lowers $n^\star$ and increases average encounter size $Y/n^\star$. In the discrete problem, the magnitude of the shift depends on how exposure increments change around the pre-change optimum.

Appendix B.5. Microfoundation for state-dependent exposure

A simple reduced-form microfoundation rationalizes escalating exposure costs. Let $s(t)$ denote a scrutiny state that summarizes accumulated information and attention about the agent. Between encounters, scrutiny decays as

$$\dot{s}(t) = -\delta s(t), \qquad \delta > 0.$$

Each encounter at time t_j increases scrutiny,

$$s(t_j^+) = s(t_j^-) + \eta, \qquad \eta > 0.$$

Let the per-encounter exposure cost be

$$c(t_j) = c_0 + \kappa s(t_j^-), \qquad c_0 > 0,\ \kappa > 0.$$

Then repeated encounters (especially in close succession) raise expected exposure costs even if statutory penalties do not change.

Appendix B.6. When equal splitting fails: order- and size-dependent exposure

In the baseline, exposure depends only on the number of encounters n, so equal splitting follows from convexity of $g(\cdot)$. If exposure depends on sequence position or encounter size, equal splitting need not hold. For example, let the

exposure cost of encounter i be $c(i)+\kappa y_i$, where $c(i)$ is increasing in i and $\kappa \geq 0$. Then the agent may optimally front-load activity, with $y^1 \geq y^2 \geq \cdots \geq y^n$, because shifting activity from later encounters (higher fixed exposure) to earlier encounters (lower fixed exposure) reduces total cost. This yields an additional empirical implication: under escalating scrutiny, activity may shift earlier in the encounter cycle even when total activity is unchanged.

Appendix B.7. Proof of Theorem 1

Proof. Let the agent choose $n \in \mathbb{N}$ to minimize $\Phi(n; s) = ng(Y/n) + C(n; s)$, where $C(n; s)$ is increasing in s and satisfies increasing differences in (n, s). By Proposition 7, higher s weakly reduces the optimal n. Because n is discrete, $n^\star(s)$ is a step function in s with switching thresholds.

Let per-encounter detectability be $p(y) = \pi(B/y)$ as in Assumption 4. For a given n, total detection over the horizon is

$$D(n) \equiv 1 - \big(1 - p(Y/n)\big)^n.$$

When n falls from n to $n-1$, encounter size rises from Y/n to $Y/(n-1)$ and, under capacity dilution, per-encounter detectability weakly falls: $p(Y/(n-1)) \leq p(Y/n)$. Whether total detection falls depends on the tradeoff between fewer independent draws and lower per-draw detectability. A drop at a switching threshold occurs when

$$D(n-1) < D(n),$$

equivalently,

$$\big(1 - p(Y/(n-1))\big)^{n-1} > \big(1 - p(Y/n)\big)^n.$$

Thus there exist values of s at which a tightening that induces a discrete reduction in n reduces equilibrium detection, yielding the paradox-of-tightening pattern. □

Remark. (Boundary condition.) If screening resources scale proportionally with encounter size so that $p(y)$ does not decline with y, the dilution channel disappears and detection is monotone in s.