Stringency, Capacity, and the Organization of Economic Activity

By Endre J. Reite*

This paper studies enforcement under administrative capacity constraints. When regulated activity is fixed, a higher per-encounter burden shifts activity into fewer, larger encounters. If screening capacity is allocated at the encounter level, the detection probability declines with encounter size because screening intensity per unit falls under bundling. The model yields a non-monotonic relationship between stringency and detection and highlights encounter architecture as a policy lever. The analysis shows that encounter-size caps, frequency requirements, and within-encounter sampling can improve effective detection relative to reforms that only increase per-encounter burdens.

JEL: D82, H26, K42

Keywords: enforcement; monitoring; administrative capacity

I study enforcement regimes in which government agencies interact with regulated agents through discrete encounters under binding time and processing constraints. Examples of such interactions include tax return processing and audits, customs inspections and clearance events, and periodic regulatory filings. When total regulated activity is fixed, enforcement policy affects not the level of activity but how that activity is organized across encounters. This organizational margin matters because enforcement technologies are often capacity-constrained: screening time and attention per encounter are limited. Screening intensity per unit of activity thus falls when activity is bundled. I formalize how agents respond by adjusting the encounter frequency and size, and show that standard levers such as

* Norwegian University of Science and Technology, NTNU Business School (email: endre.j.reite@ntnu.no).

higher penalties or per-encounter 'stringency' can unintentionally reduce effective detection by inducing endogenous bundling. Unlike standard deterrence models that take the timing and structure of encounters as fixed, this paper shows that enforcement can end up reshaping how economic activity is organized rather than how much of it occurs. This shift generates a distinctive set of predictions—sharp consolidation thresholds, non-monotonic detection patterns, and persistence that arises purely from procedural exposure, not deterrence incentives.

Many enforcement environments impose real interaction costs—documentation, reconciliation, follow-up requests, compliance meetings, and managerial attention. When regulated parties can influence encounter frequency, policies that raise per-encounter burden convert a given volume of activity into fewer, larger encounters. The core margin is the organization of activity across encounters—frequency, scope, and timing—rather than the aggregate level of activity over the horizon.

The paper develops a theory of endogenous encounter structure. In the baseline model, an agent allocates a fixed volume of activity across encounters and chooses the number of encounters to minimize total cost. Convex within-encounter processing costs and per-encounter exposure costs jointly pin down an interior optimal encounter count and imply equal splitting conditional on that count. The benchmark isolates a simple tradeoff: spreading activity across more encounters lowers processing costs but increases exposure to repeated scrutiny.

The model is then extended in two directions. First, exposure depends on past interactions, generating persistence in encounter behavior even after enforcement conditions relax. Second, detectability is allowed to decline with encounter size, reflecting screening congestion, informational overload, or complexity effects. Under this structure, increasing the per-encounter stringency can induce consolidation that lowers equilibrium detection. Detection falls discretely at the thresholds where agents switch from n to $n-1$ encounters, producing a non-monotonic relationship between enforcement and detection.

This mechanism yields a *paradox of stringency*: beyond a threshold, tighter

per-encounter enforcement can reduce equilibrium detection by inducing agents to consolidate activity into fewer, larger encounters when screening effectiveness declines with the scope of an encounter. The effect does not rely on behavioral frictions or evasion technologies. It follows from rational organizational adjustments to encounter-level incentives.

The contribution is conceptual and structural. The paper endogenizes encounter frequency in an enforcement model and derives conditions under which enforcement reshapes the organization of a fixed workload across encounters rather than reducing total activity. This yields sharp comparative statics for consolidation thresholds, discontinuous detection responses, and persistence following temporary enforcement shocks. The predictions differ from standard deterrence frameworks in which encounter structure is fixed and enforcement effects are monotone in intensity. I defer administrative design implications until after establishing the mechanism and comparative statics; that discussion focuses on encounter architecture—scope, frequency, and within-encounter allocation—under binding screening capacity.

Section I places the mechanism in the enforcement and organizational adjustment literatures. Section II develops the model and characterizes the optimal encounter structure. Subsections II.H–II.J extend the analysis to persistence, monitoring design, and size-dependent detectability. Section III turns to policy design under binding screening capacity, with an explicit comparison between stringency reforms and encounter-architectural reforms (Subsection III.E). Section IV discusses design implications and boundary conditions. Section V sets out empirical implications and falsification strategies. Section VI concludes.

I. 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.

A. Batching and lumpy adjustment (background)

The benchmark mechanism—partitioning a fixed total Y across a chosen number of encounters—connects to classic batching and lumpy-adjustment logic. In inventory or cash-management and adjustment-cost settings, agents trade off fixed or convex costs against the benefits of smoothing, generating discrete changes in frequency and batch size. The present model uses this logic to interpret administrative interactions as an endogenous *encounter process* rather than following 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 from the typical approach in its emphasis: it focuses on the equilibrium reorganization of encounters and the implications of this for what regulators observe.

B. History-dependent Enforcement and Persistence

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.

C. Administrative Burden and Modern Enforcement Delivery

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.

D. 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.

E. Processing Constraints and Capacity Dilution

The capacity-dilution mechanism in Subsection II.J 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).

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.

F. Contribution relative to the literature

Taken together, the framework endogenizes encounter frequency and clarifies how enforcement can change what regulators observe by reshaping the encounter process itself. With history-dependent scrutiny and size-sensitive detectability, the model generates persistence in interaction patterns and identifies conditions under which stricter per-encounter requirements can lower equilibrium detection by inducing consolidation into fewer, larger encounters.

Appendix B provides an extended mapping and additional references.

II. Model

A. 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.

B. 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 III, I treat encounter design as a policy lever and describe how capacity constraints shape optimal administrative rules.

C. Cost Structure and Cumulative Exposure

Total cost has two components: an internal (batching) cost and an enforcement (exposure) cost. Interaction timing is assumed to be flexible. 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. 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.

D. 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.

E. 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).

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

G. 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:

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

H. 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* $a_1 = a_2$.

I. 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 II.H.

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): *Monitoring rules condition on encounter size such that the incentive to consolidate is attenuated or reversed relative to encounter-only monitoring.*

Sketch. Conditioning scrutiny on interaction size raises the marginal cost of bundling activity, offsetting the benefit of reducing the encounter count. When monitoring intensity increases sufficiently with y_t, the agent trades off size against frequency rather than collapsing interactions.

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 C.

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.

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

J. 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 $\pi''(\cdot) < 0$.

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 stringency): *Suppose Assumption 4 holds and the agent chooses an encounter frequency optimally, as in Section II. Then there exist values of the exposure schedule (stringency) at which an increase in per-encounter stringency reduces the regulator's equilibrium probability of detection.*

Table 3 reports a numerical calibration of the model. As per-encounter exposure increases from 400 to 625, the optimal number of encounters falls from 5 to 4, the average encounter size rises, and total detection declines 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.

III. 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.

A. 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 those that increase statutory stringency (higher penalties or 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.

B. 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 $y > 0$. Fix total activity Y and consider any feasible allocation $\{y_i\}_{i=1}^n$ with $\sum_i y_i = Y$. 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 that each satisfy $\tilde{y}_j \leq \bar{y}$. Then*

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

If π is strictly concave and at least one encounter is split, the inequality is strict under standard regulatory conditions.

PROOF SKETCH:

Under Assumption 4, $p(y)$ increases as y falls because screening intensity per unit rises. Since $D(\cdot)$ is increasing in each $p(y_i)$, splitting a large encounter into smaller ones weakly increases $D(\cdot)$. Strictness obtains when the concavity of π implies diminishing returns to concentrating bounded screening intensity within a single large encounter.

C. 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.

D. 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 IV, and empirical signatures and falsification strategies in Section V.

E. 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.

A standard stringency reform. Consider a reform that raises per-encounter procedural burden or scrutiny intensity, without scaling encounter-level screening resources. In the model, this is naturally represented as an upward shift in the exposure schedule $C(n)$. Proposition 2 implies weak consolidation: the agent reduces encounter frequency and increases average encounter size. Under capacity dilution, larger encounters receive lower screening intensity per unit, $e(y) = B/y$, which reduces per-encounter detectability $p(y) = \pi(B/y)$. At consolidation thresholds, this channel can dominate the mechanical increase in per-encounter burden, so equilibrium detection falls as stringency rises (Theorem 1). Table 3 provides a numerical illustration, and Figure 3 maps the parameter region where the detection loss occurs.

An architectural reform. Now consider reforms that restrict bundling directly rather than raising per-encounter burden. Two examples are encounter size caps, $y_i \leq \bar{y}$, and minimum frequency requirements, $n \geq n_{\min}$. These instruments operate on the encounter process itself and therefore act on the channel that generates capacity dilution. Proposition 6 shows that capping encounter size weakly in-

creases detection when screening capacity per encounter is fixed, with strict gains under standard concavity conditions. A closely related design is within-encounter sampling: for large encounters, the regulator inspects a subset of components so that effective screening capacity scales with volume, as in Subsection III.C. In the model, this corresponds to making B increasing in y, which attenuates the dilution channel and reduces the scope for the paradox.

Implication. Per-encounter stringency works through exposure and therefore strengthens incentives to consolidate along the encounter margin. Architectural reforms act on that margin directly by limiting bundling or preserving screening intensity per unit. When processing constraints bind, pairing higher per-encounter burdens with an architecture that constrains consolidation avoids a mechanical loss of observability.

IV. 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.

V. Empirical Implications and Measurement

Empirical scope. Although the paper is theoretical, its main predictions concern objects routinely observed in administrative data: encounter counts, encounter sizes, and encounter timing. The cleanest tests exploit enforcement changes that shift per-encounter burden or targeting intensity while permitting

longer-horizon totals to be tracked. Accordingly, filings and inspections should be interpreted as equilibrium outcomes of encounter design rather than direct measures of underlying activity. Many of the predictions are most relevant in environments with encounter-level screening and binding processing constraints (see Section II.B).

These predictions can be examined using standard empirical designs. Under pure deterrence, reductions in the frequency of interactions must be accompanied by reductions in total activity. Under exposure–driven consolidation, interaction frequency falls while total activity remains stable, and per-interaction size rises. If interaction counts and per-interaction size both fall, the consolidation mechanism is inconsistent with the model.

Event studies of increases in monitoring intensity, documentation requirements, or review procedures can trace how interaction counts, sizes, and timing adjust relative to totals. A difference-in-differences design that exploits differential exposure to enforcement changes—such as risk-based targeting or procedural thresholds—can be used to test whether treated units exhibit consolidation and clustering. Policy changes that alter review or follow-up practices provide particularly clean settings, as they plausibly affect the encounter burden without directly changing statutory incentives (Choudhary and Gupta, 2025; Zhou et al., 2023).

The result concerning information collapse also has implications for measurement. When the probability of detecting a violation within an interaction declines with interaction size, consolidation affects both exposure and observability. In that case, organizational responses induced by enforcement can reduce the total information generated by regulatory encounters. Beyond a threshold level of per-encounter stringency, further increases in enforcement can, therefore, lower overall detection probabilities, even as scrutiny intensifies.

Empirically, this implies a nonmonotonic relationship between enforcement intensity and detection outcomes. Detection per–encounter may rise as encounters are more heavily scrutinized, while detection per unit of underlying activity de-

clines as activity is bundled into fewer, larger episodes that are harder to screen. This pattern distinguishes information collapse from standard deterrence or capacity constraints, which predict monotonic improvements in detection as enforcement increases (Maćkowiak et al., 2023).

A final implication concerns interpretation. Short aggregation windows can mechanically exaggerate declines in totals when activity is re-timed into fewer encounters, while declines in filing counts are often misread as reductions in underlying activity. My analysis clarifies why such patterns need not reflect weak enforcement or transitory responses. Instead, they may be the equilibrium outcome of rational agents reorganizing activity to economize on exposure and, when detectability is size-dependent, to reduce the informational value of each encounter. Empirical analyses that track encounter counts, sizes, timing, and detection outcomes jointly are therefore essential for assessing enforcement effects in environments with persistent scrutiny.

VI. 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.

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

Period	Scrutiny level (s_t)	Optimal encounters (n_t^*)	Average size (Y/n_t^*)
1	High (s_H)	5	20
2	Low (s_L)	6	16.67

The table illustrates the two-period extension used to formalize persistence. Total activity is normalized to $Y = 100$. High scrutiny in period 1 induces consolidation ($n_1^* = 5$). If there is residual scrutiny that carries over into period 2, the period-2 optimum level can remain below the baseline even when s returns to s_L.

Table 2—Size-contingent monitoring: encounter-only versus size-contingent inspections (illustration).

Encounters (n)	Average size (Y/n)	Encounter-only detection (%)	Size-contingent detection (%)
3	33.33	5.32	7.88
4	25.00	6.45	7.97
5	20.00	7.12	8.05

The table compares total detection under two monitoring designs. Under encounter-only monitoring, each encounter is inspected with the same intensity. Where monitoring is contingent on size, scrutiny increases with encounter size, limiting the advantage of consolidation. Numbers are illustrative and correspond to the parameter values used in the numerical example in Section III.

Table 3—The paradox of stringency: tightening enforcement can reduce detection (illustration).

Regime	Exposure cost (per encounter)	Optimal encounters (n^*)	Average size (Y/n^*)	Detectability $p(Y/n^*)$	Total detection $D(n^*)$
A (moderate)	0.40	5	20.00	0.0100	0.0490
B (stricter)	0.50	4	25.00	0.0055	0.0218

The example shows how a stricter exposure regime can reduce detection when the optimal number of encounters falls. Detectability $p(y)$ declines in encounter size y under capacity dilution, so the drop in n raises Y/n and lowers $p(Y/n)$. Even if regime B is stricter in the exposure schedule, the induced consolidation can dominate and reduce total detection.

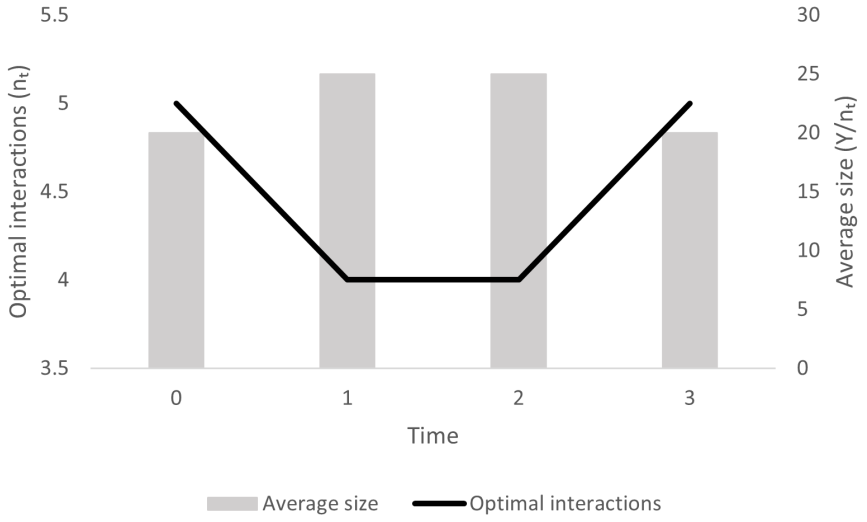


Figure 1. Optimal interaction frequency and average interaction size over time

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

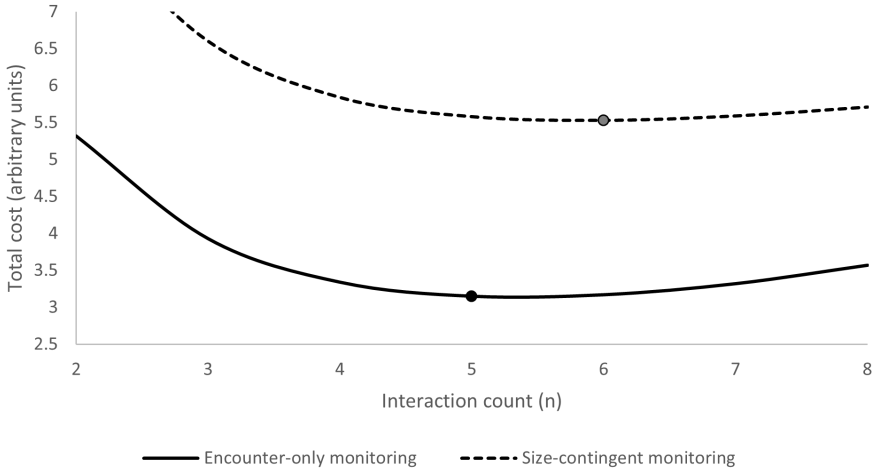


FIGURE 2. TOTAL COST UNDER ENCOUNTER-ONLY AND SIZE-CONTINGENT MONITORING

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

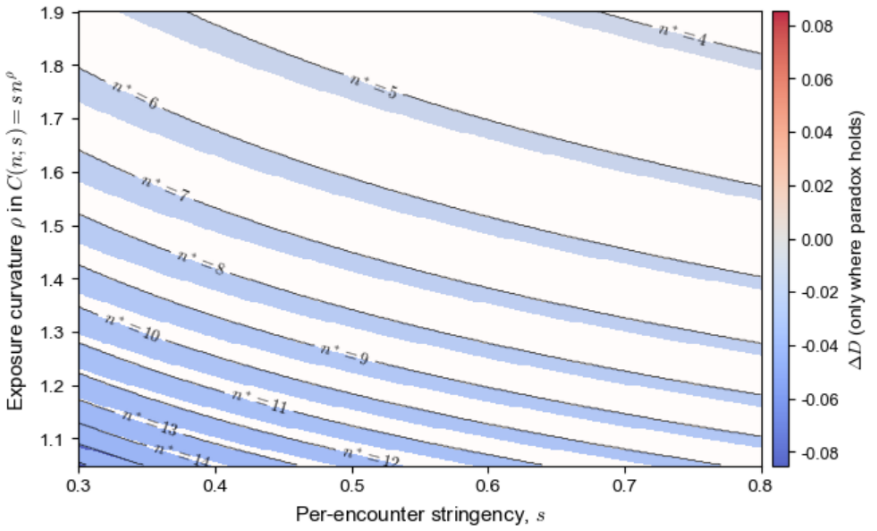


FIGURE 3. PARADOX-OF-STRINGENCY REGION AND CONSOLIDATION THRESHOLDS.

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

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.

*

Appendix

Supplementary Material

This appendix collects supplementary material that supports the main text. It provides extended references and illustrative examples that complement the lit-

erature discussion in Section I, and it records notation and minor derivations omitted from the main analysis for readability.

Extended Framework Mapping and Empirical Guidance

B1. Additional example: targeted capacity under resource constraints

I treat exposure costs in reduced form to isolate how procedural frictions shape the organization of encounters rather than relying on institution-specific details. This approach is consistent with modern enforcement systems that allocate scarce scrutiny using risk information and targeted capacity. For example, in environmental regulation, inspection allocation can be framed as a resource-constrained problem, and data-driven methods can be used to target high-risk facilities for inspection (Zhou et al., 2023).

B2. Extended mapping: discrete regimes and reorganization around thresholds

A closely related implication of the paper is that monitoring design can induce reorganization around discrete regimes, altering the mapping from observed reports to underlying activity. When targeting relies on risk scores or cutoffs constructed from those reports, reorganization can also change 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 timing.

B3. Robustness to Alternative Exposure-cost Schedules

The consolidation comparative static does not depend on a narrow functional form for the exposure schedule. To make this explicit, consider five stylized forms of the marginal exposure cost $c(k)$ (so $C(n) = \sum_{k=1}^{n} c(k)$):

Case 1: Constant marginal cost. $c(k) = \bar{c} > 0$ (so $C(n) = \bar{c}\, n$). Case 2: Linear escalation. Later encounters are more burdensome; higher slope shifts the optimum toward fewer interactions, $n^\star$, and larger average batches, $Y/n^\star$.
Case 3: Convex escalation. Larger increases in the incremental burden produce sharper (discrete) reductions in $n^\star$.
Case 4: Concave escalation or saturation. $c(k) = c_0 + b(1 - e^{-\lambda(k-1)})$, $\lambda > 0$. Consolidation persists as long as $c(k)$is nondecreasing over the relevant range.
Case 5: Threshold escalation (trigger rules). $c(k) = c_L$for $k \leq K$and $c(k) = c_H$for $k > K$, with $c_H > c_L$. This captures institutional triggers (repeat amendments, follow-ups, and escalation rules) and implies discontinuous shifts in $n^\star$ and, potentially, strong bunching.

Key robustness point: the magnitude of the consolidation response depends on the extent to which $c(k)$ increases around the pre-change optimum.

Mathematical Appendix

C1. Benchmark: Equal-split Allocation and Existence of an Optimal Interaction Count

Benchmark optimization and monotone comparative statics

This appendix records the benchmark allocation problem and the monotone comparative-statics result invoked in the main text. The core equal-split argument and existence claim are identical to the proofs given in the main text; to avoid duplication, I state them here and refer to the corresponding main-text steps for proofs.

A1. Problem

An agent must complete total activity $Y > 0$ over the horizon. The agent chooses an integer number of encounters $n \in \mathbb{N}$ and encounter sizes $\{y_i\}_{i=1}^{n}$ with

$y_i \geq 0$ and $\sum_{i=1}^{n} y_i = Y$. Total cost is

$$\sum_{i=1}^{n} g(y_i) + C(n), \tag{A1}$$

where $g : \mathbb{R}_+ \to \mathbb{R}_+$ captures within-encounter processing costs and $C : \mathbb{N} \to \mathbb{R}_+$ captures exposure (or other per-encounter) costs.

ASSUMPTION 5 (Strict convexity): *g is strictly convex on $\mathbb{R}_+$.*

ASSUMPTION 6 (Finite optimum): *$C(n) \to \infty$ as $n \to \infty$.*

The agent solves

$$\min_{n \in \mathbb{N}} \min_{\{y_i\}_{i=1}^{n}} \sum_{i=1}^{n} g(y_i) + C(n) \quad \text{s.t.} \quad \sum_{i=1}^{n} y_i = Y, \quad y_i \geq 0. \tag{A2}$$

Define the minimized inner value

$$\Phi(n) \equiv \min_{\{y_i\}_{i=1}^{n}} \sum_{i=1}^{n} g(y_i) \quad \text{s.t.} \quad \sum_{i=1}^{n} y_i = Y, \quad y_i \geq 0. \tag{A3}$$

A2. Equal-split and reduction to a one-dimensional problem

LEMMA 2 (Equal split conditional on n): *Fix $n \geq 1$. Under Assumption 5, the unique solution to* (A3) *satisfies $y_i^\star = Y/n$ for all $i = 1, \ldots, n$, and thus*

$$\Phi(n) = n\, g(Y/n). \tag{A4}$$

Proof. See the Jensen-based argument in the main text (steps establishing Lemma 1 / equal-split allocation). □

Lemma 2 implies the discrete problem

$$\min_{n \in \mathbb{N}} V(n) \quad \text{where} \quad V(n) \equiv n\, g(Y/n) + C(n). \tag{A5}$$

PROPOSITION 7 (Existence of an optimal finite interaction count): *Under Assumptions 5–6, there exists a finite $n^\star \in \mathbb{N}$ such that $V(n^\star) \leq V(n)$ for all $n \in \mathbb{N}$.*

Proof. Identical to the main-text existence argument: since $V(n) \geq C(n)$ and $C(n) \to \infty$, the minimizer lies in a finite set. □

A3. Consolidation under higher marginal exposure

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

$$\min_{n \in \mathbb{N}} V(n; s) \quad \text{where} \quad V(n; s) \equiv n\, g(Y/n) + C(n; s). \tag{A6}$$

Assume:

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

(ii) (*Increasing differences / single-crossing*) For all $n \in \mathbb{N}$ and $s_2 > s_1$,

$$\Delta(n; s) \equiv V(n+1; s) - V(n; s) \quad \text{satisfies} \quad \Delta(n; s_2) \geq \Delta(n; s_1), \tag{A7}$$

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

PROPOSITION 8 (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 (larger s) weakly reduces the optimal number of encounters.*

Proof. Condition (ii) is increasing differences on the lattice $\mathbb{N} \times \mathcal{S}$, stated in discrete-increment form (A7). By Topkis' monotonicity theorem for minimization problems on lattices, the smallest minimizer is weakly decreasing in s (Topkis, 1978). □

A4. Benchmark Where Encounter Frequency is Irrelevant

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

A5. Parametric Robustness (Explicit Role of Increment Size)

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

Then $\Phi(n) = n \cdot g\left(\frac{Y}{n}\right) + C(n) = \frac{Y^2}{n} + \ \alpha \cdot n^{\rho}$.

Treating n as continuous for approximation, the first-order condition gives $n* \approx \left(\frac{Y^2}{\alpha \cdot \rho}\right)^{\left\{\frac{1}{\rho+1}\right\}}$.

This expression makes the comparative static transparent: higher α (or higher ρ near the relevant range) lowers n* and increases average batch size Y/n*. In the discrete problem, the size of the shift depends on how much marginal exposure increments change around the pre-change optimum.

A6. Learning/State-dependent Scrutiny microfoundation

A simple microfoundation rationalizes escalating exposure costs. Let $s(t)$ denote a scrutiny state that summarizes accumulated information and attention about the agent. Between interactions, scrutiny decays according to

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

Each interaction at time t_j increases scrutiny via information accumulation or escalation triggers:

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

Let the per-interaction 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.

Interpretation. ('learning > 0 implies underestimation'). When η>0, observed declines in the interaction frequency after enforcement changes can understate underlying organizational responses: agents optimally reduce how often they appear even when total activity is unchanged.

A7. Unequal Batch Sizes When Exposure Depends on Order and Size

In the baseline, exposure costs depend only on the number of encounters n, so equal splitting follows. If exposure depends on the sequence position or the size of the encounter, equal splitting can fail.

Example: let the exposure cost of encounter i be $c(i) + \kappa \cdot y_i$, where $c(i)$ is increasing in i and $\kappa \geq 0$. Then, the agent may optimally front-load activity:

(A8) $$y^1 \geq\ y^2 \geq\ \cdots\ \geq\ y_n,$$

because shifting mass from later encounters (high fixed exposure) to earlier encounters (lower fixed exposure) reduces the total cost. This generates an additional empirical prediction: under escalating scrutiny, activity can shift to earlier in the reporting cycle ('front–loading') even when total activity is unchanged.

A8. 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 has increasing differences in (n, s); therefore, higher s weakly reduces the optimal n. Let the per-encounter detection probability be $p(y) = \pi(B/y)$, decreasing in y under Assumption 4.

$$D(n) \equiv 1 - \big(1 - p(Y/n)\big)^n, \tag{A9}$$

Fix s and compare $D(n)$ across adjacent n. Under Assumption 4, $p(y) = \pi(B/y)$ is decreasing in y, so consolidation from n to $n-1$ increases the encounter size and lowers the per-encounter detectability: $p(Y/(n-1)) \leq p(Y/n)$. The effect on $D(n)$ is ambiguous because the regulator receives fewer independent draws.

Step 1 (existence of consolidation thresholds). By Proposition B.2 (monotone comparative statics), if the per-encounter exposure term is strengthened by s, the optimal interaction count $n^\star(s)$is weakly decreasing in s. Because n is an integer, $n^\star(s)$ is a step function: there exist values s_n at which the optimum switches from $n$to $n-1$.

Step 2 (detection drop at a threshold). Fix a threshold s_n at which the agent switches from n to $n-1$ encounters as s increases. A drop in equilibrium detection at the threshold occurs if

$$D(n-1) < D(n), \tag{A10}$$

equivalently,

$$\big(1 - p(Y/(n-1))\big)^{n-1} > \big(1 - p(Y/n)\big)^n. \tag{A11}$$

Step 3 (non-monotonicity in s). For fixed n, $D(n)$ is determined by the encounter size Y/n under Assumption 4. As stringency increases through the exposure schedule $C(\cdot; s)$, the agent's optimal encounter count $n^*(s)$ falls at discrete thresholds. These downward steps raise encounter size and reduce $p(Y/n)$, which can produce downward jumps in equilibrium detection $D(n^*(s))$.

As Theorem 1 shows, the regulator's detection probability can fall when stringency tightens. Figure 3 maps the (s, ρ) combinations for which this paradox arises and highlights the consolidation thresholds that give rise to the non-monotonic detection pattern.

Remark (boundary condition). If the regulator can scale screening resources proportionally with encounter size (so that $p(y)$ does not decline with y), then (B.6.1) fails and downward jumps disappear; detection becomes monotone in s.