Evolutionarily Stable Benchmarks: Endogenous Information Production, Manipulation, and Focal Points*

Andrew Bird,† Steve Karolyi,‡ and Thomas Ruchti§

This version: May 2026 (v5)

Abstract

Reducing a continuous signal to pass/fail is conventionally understood as destroying information. This paper argues that thresholds concentrate an endogenous information ecosystem at the decision boundary, generating a self-reinforcing loop between evaluator coordination and information investment, and that the concentration produces more decision-relevant information than continuous evaluation would. The stochastically stable benchmark maximizes induced screening quality—the sum of raw screening quality and the average information premium along the adoption path—so a benchmark with mediocre intrinsic properties but a rich ecosystem defeats one with superior properties but a thin ecosystem. Stochastic stability and population-game risk dominance both select this benchmark, and a global-games argument under the noise-vanishing limit of Frankel, Morris, and Pauzner (2003) is sketched for the two-benchmark case (Online Appendix). Goodhart's Law, the proposition that a measure used as a target ceases to be a good measure once agents optimize against it, is moderated rather than overturned, because endogenous information production partially offsets manipulation-induced degradation.

Keywords: Benchmarks, focal points, endogenous information production, evolutionary game theory, manipulation, Goodhart's Law, equilibrium selection, stochastic stability.

JEL Classification: C73, D82, D83, G14, M41.

*We thank Laurence Ales, Ratul Lahkar, Laura Veldkamp, and especially Peyton Young for helpful comments and discussion. All errors are our own.

†Argyros College of Business and Economics, Chapman University, One University Drive, Orange, CA 92866. Email: abird@chapman.edu.

‡School of Business, George Mason University, 4400 University Drive, MSN 5F5, Fairfax, VA 22030. Email: skarolyi@gmu.edu.

§Pamplin College of Business, Virginia Tech, 880 West Campus Drive, Blacksburg, VA 24061. Email: ruchti@vt.edu.

1 Introduction

Benchmarks coarsen rich continuous information into crude binary signals, invite gaming, and evaluators use them anyway. The conventional view treats them as second-best compromises that persist because contracting frictions prevent anything better (Hart and Moore, 1988; Holmstrom and Milgrom, 1991), tolerated only because the alternatives are worse.

This paper proposes a different view. Benchmarks are valued for their crudeness. A binary threshold reorganizes the information environment by giving information producers a single focal question and concentrating decision sensitivity at one point. When evaluators coordinate on a benchmark, producers find it profitable to invest in signals near the threshold because decisions turn on fine distinctions near the cut-off and are insensitive to them far away. The resulting concentration generates a self-reinforcing loop. More evaluators raise demand for benchmark-relative information, producers respond with investment that reduces noise and improves screening quality, and better screening attracts still more evaluators. Coarsening evaluation to a threshold therefore generates more decision-relevant information than continuous evaluation would.

The framework features three populations of actors with a clear order of moves. Evaluators commit to a benchmark first. Information producers then invest in signals about reports near that benchmark, with revenue scaling in the share of evaluators using it. Reporters last decide how much to manipulate their signals given the noise producers' investment delivers, signals are realized, and evaluators apply their thresholds.

The paper formalizes this loop and delivers four results. The information loop generates a population game with strategy-specific externalities, and the benchmark that survives long-run evolutionary selection maximizes induced screening quality, the sum of raw screening quality and the average information premium the benchmark attracts over the path from zero adoption to full coordination. A benchmark with mediocre intrinsic properties but a rich information ecosystem defeats one with superior intrinsic properties but a thin ecosystem.

Building on Lahkar, Mukherjee, and Roy (2024), who establish convergence for general continuum potential games, the selected benchmark on continuous strategy spaces is characterized through a variational first-order condition, a transition barrier

decomposition, and monotone comparative statics that exploit the economic structure of the information loop. Risk dominance also selects the same benchmark.[1] Benchmarks survive because they generate information ecosystems that justify their own existence.

Goodhart's Law (Goodhart, 1975), the proposition that any measure used as a target ceases to be a good measure once agents optimize against it, is moderated rather than overturned in our framework. Asymmetric manipulation costs ensure that an agent who manipulates to pass is on average of higher quality than one who cannot, so manipulation partially preserves rather than destroys information. The self-reinforcing loop then generates information investment that actively offsets the information loss from gaming. Evaluators rationally persist with benchmarks they know are manipulated because the information ecosystem the benchmark sustains partially offsets the gaming it invites, leaving the threshold strictly informative.

Consider the credit score cutoffs at 620 (conforming-loan boundary) and 740 (best-rate boundary) that lenders use to screen mortgage applications. Applicants game the score by paying down balances before reporting dates, disputing items, and timing new accounts, and the cutoffs also shape the information environment. Credit bureaus, scoring vendors, and underwriting consultants concentrate analytical effort between 615 and 625 and between 735 and 745 because that is where lender decisions turn. A refinement that sharpens default prediction at 540 is worth almost nothing because no one is making an approval decision there, while the same refinement at 620 determines whether thousands of loans are written.

Grade point averages operate similarly. A 3.0 cumulative GPA is a common cutoff for entry-level positions, with students as reporters, recruiters as evaluators, and grading professors as information producers. Because grades near the cutoff carry the greatest decision weight, professors apply more care to work hovering between a B minus and a B than to work between an A and an A minus, and the information ecosystem follows the decision boundary rather than the underlying ability distribution.

The same force operates wherever many evaluators screen against a common threshold. Bank capital requirements produce a risk-modeling industry concentrated

[1] A global-games argument in the spirit of Carlsson and van Damme (1993) and Frankel, Morris, and Pauzner (2003) delivers the same selection in the noise-vanishing limit. The construction is sketched in Online Appendix Section OA.2.3.

at the minimum capital ratio, and in financial markets the consensus forecast attracts an analyst ecosystem far larger than any alternative threshold could sustain. The benchmark acts as an active information magnet at the point where information is most valuable, with strategic behavior layered on top rather than corrupting the underlying function.

Three empirical regularities follow from this single mechanism. First, emergence. The self-reinforcing loop selects for information attractiveness, not intrinsic quality, so the 3.0 GPA persists because of the tutoring, advising, and course-selection infrastructure built around it, and the analyst consensus forecast persists despite mediocre raw screening quality because of the information ecosystem it attracts. Second, bunching. Asymmetric manipulation costs create an interval of types that manipulate to exactly meet the benchmark, producing the mass point, the missing bin below, and the undistorted distribution elsewhere documented in earnings distributions (Burgstahler and Dichev, 1997; DeGeorge, Patel, and Zeckhauser, 1999), taxable incomes at EITC kinks (Saez, 2010), and regulated emissions and capital ratios at compliance thresholds. Third, moderation of Goodhart's Law. Asymmetric costs make manipulation partially informative and endogenous information investment offsets the information loss from gaming, so the surviving benchmark is the one where the information response most strongly dominates the manipulation degradation.

The argument draws on several traditions. The empirical starting point is bunching at thresholds in earnings distributions (Burgstahler and Dichev, 1997; DeGeorge, Patel, and Zeckhauser, 1999), at tax kinks (Saez, 2010), and across other settings (Kleven, 2016). These papers establish clustering but leave open why particular thresholds emerge and persist. Schelling (1960) provides focal points, and we endogenize salience through information production. The key building block is Persico (2000), who shows that information value is higher in more risk-sensitive decision problems. Threshold evaluation creates exactly that structure at the decision boundary, and coordination across evaluators amplifies the resulting concentration into a self-reinforcing loop. The contribution is the evolutionary selection mechanism that determines which threshold survives, drawing on stochastic stability (Foster and Young, 1990; Kandori, Mailath, and Rob, 1993; Young, 1993) and continuous-strategy extensions (Oechssler and Riedel, 2001; Lahkar and Riedel, 2015; Lahkar, Mukherjee, and Roy, 2024). Selection coincides with risk dominance (Harsanyi and Selten, 1988). On the manipulation side, Arya, Glover, and Sunder (1998, 2003) show that

managed earnings can be more informative than unmanaged earnings under private information, and we generalize from binary to continuous types and embed the result in a dynamic framework with endogenous information production. The framework reverses the destructive-manipulation conclusions of Stein (1989) and Fischer and Verrecchia (2000) because benchmark-induced information production partially offsets Goodhart degradation. The benchmark-choice problem relates to information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019), but the structure is shaped by decentralized investment rather than a single designer. The decentralization most closely echoes the multi-sender extensions of information design (Mathevet, Perego, and Taneva, 2020), with our information producers playing a role analogous to competing senders, except that here the relevant action of each producer is investment in precision rather than choice of signal structure. The result that coarse evaluation can dominate fine evaluation has an incentive-based counterpart in the optimal-tournament literature (Lazear and Rosen, 1981; Moldovanu and Sela, 2001), while our mechanism is information-based and the two channels are complementary.

2 The Model

The model has three populations. Evaluators choose a benchmark and use it to accept or reject reports. Reporters submit reports and have a private incentive to manipulate upward when they can afford to do so. Information producers invest in monitoring, technology, or expertise that sharpens the signal near the benchmark, earning revenue that scales with the audience using that benchmark. The evaluators are the strategic actors of the evolutionary game, with reporters and information producers playing optimal responses to the prevailing evaluator distribution. The order of moves within a period is direct. Evaluators inherit or choose a benchmark, information producers invest given that benchmark and its share, reporters choose manipulation given the resulting signal environment, and signals realize last with evaluators applying their thresholds. The evaluator population revises its benchmark choice over time through the stochastic dynamics of Section 4, under log-linear response dynamics in a population game with strategy-specific externalities.

Reporting manipulation is a costly real action altering the realized signal (Stein, 1989; Dye, 1988; Fischer and Verrecchia, 2000), not a costless message about type, so the revelation principle does not apply. Signal noise is endogenously determined by

third-party information producers whose investment responds to the evaluator's rule (Persico, 2000), and the information structure is part of what the rule chooses rather than a fixed environment. The threshold is endogenously optimal under this friction because it concentrates decision sensitivity at a single point, attracting the investment that determines precision there. A finer partition spreads investment thinner across the signal space, leaving the evaluator with less precision where it matters.

These populations map to identifiable groups in each application. In credit markets, lenders are evaluators, loan applicants are reporters, and credit bureaus, scoring vendors, and underwriting consultants are information producers. In education, recruiters and admissions committees are evaluators, students are reporters, and professors and the advising apparatus are information producers. In financial markets, investors and the institutions that rely on the benchmark to allocate capital are evaluators, firms are reporters, and sell-side analysts, data providers, and forensic accountants are information producers. The formal structure applies whenever agents produce reports assessed against thresholds by evaluators who benefit from coordination.

2.1 A Two-Benchmark Example

To fix ideas before the formal model, consider a stripped example with just enough structure to show the central mechanism. The example is illustrative and does not yet contain the main findings of the paper. It uses two reporter types, two candidate benchmarks, binary information investment, and no manipulation.

Two reporter types $\theta \in \{L, H\}$ have $\Pr(H) = \pi$. The evaluator chooses between benchmarks b_1, b_2 with intrinsic screening quality $v_1 > v_2$. The benchmarks differ in informational responsiveness. The simple benchmark b_1 requires no information ecosystem, $\Lambda_1 \equiv 0$ by assumption, because b_1 is read directly off the report and information producers cannot profitably specialize on what every evaluator can already interpret unaided. The complex benchmark b_2 has an activation threshold $\bar{x}_2 \equiv c/r \in (0, 1)$ that an information producer must clear to find investment profitable. An information producer at b_2 chooses $I_2 \in \{0, 1\}$ at cost c with revenue rx_2 per unit, so investment is profitable when $x_2 \geq \bar{x}_2$. Investment reduces noise from σ_0^2 to $\sigma_1^2 < \sigma_0^2$ and generates an information premium $\Lambda_2 \equiv U(\sigma_1^2) - U(\sigma_0^2) > 0$ active only when $x_2 \geq \bar{x}_2$. The path-integrated payoffs are therefore $\Phi(\mathbf{e}_1) = v_1$ and

$\Phi(\mathbf{e}_2) = v_2 + \Lambda_2(1 - \bar{x}_2)$, the latter integrating the premium over the share interval $[\bar{x}_2, 1]$ on which it is active. Stochastic stability under log-linear dynamics selects the potential maximizer (Theorem 1), so the intrinsically inferior b_2 survives selection if and only if

$$\Lambda_2(1 - \bar{x}_2) > v_1 - v_2. \tag{1}$$

A larger information premium tilts selection toward b_2, and a higher activation threshold tilts it back toward b_1. The example captures the central economic idea in two parameters and one inequality.

The example deliberately suppresses three forces that the full model puts back. There is no manipulation, so Goodhart's Law moderation cannot be shown. There is no continuum of benchmarks, so threshold placement and the variational characterization of the surviving benchmark cannot be derived. And there is no smooth noise technology, so the information concentration result is hidden behind the binary investment assumption. Sections 2.2 through 4 add these forces and the corresponding results.

2.2 Primitives

The three populations. A continuum of reporters $i \in [0, 1]$ have privately known types $\theta_i \in [0, 1]$ drawn from a distribution F with continuous, strictly positive density f on $(0, 1)$. Each reporter produces a report $s_i = \theta_i + a_i + \varepsilon_i$ by choosing upward manipulation $a_i \geq 0$, at cost $C(a, \theta) = \frac{1}{2}\, m(\theta)\, a^2$, where $m : [0, 1] \to \mathbb{R}_{++}$ is C^1. The noise term ε_i is endogenous in the sense that its variance is pinned down later in the model by the information producers' investment choices, as developed in equation (3) below. When information producers invest more heavily in a benchmark, the noise around reports at that benchmark falls, and this dependence is what makes evaluator coordination feed back into screening quality. In the earnings application, θ_i is unmanaged earnings, a_i is earnings management, and $m(\theta)$ captures how costly it is for a firm with fundamentals θ to inflate its report. The quadratic cost is standard, and the economic content is in how $m(\theta)$ varies with quality.

A continuum of evaluators $j \in [0, 1]$ each apply a benchmark $b_j \in \mathbb{R}$, accepting if $s \geq b_j$ and rejecting otherwise. An evaluator who accepts type θ earns $\theta - \kappa$, where $\kappa \in (0, 1)$ is the breakeven type, with rejection yielding 0. The reporter's reward is a step function, $R(s, b) = u_H \cdot \mathbb{1}\{s \geq b\} + u_L \cdot \mathbb{1}\{s < b\}$ with $\Delta u \equiv u_H - u_L > 0$, so the

reporter's payoff is $\pi_R = R(s,b) - C(a,\theta)$. This discrete jump at the threshold is the defining feature of benchmarks (Bartov, Givoly, and Hayn, 2002).

The third population, comprising sell-side analysts, financial media, data providers, and forensic accountants, invests in public signals whose precision depends on evaluator coordination. A continuum of information producers $l \in [0,1]$ each invest $i_l \geq 0$, earning profit $\pi_I(i_l, x_b) = r \cdot x_b \cdot i_l - \frac{1}{2}\kappa_I\, i_l^2$, where $r > 0$ is revenue per unit of information, x_b is the share of evaluators using benchmark b, and $\kappa_I > 0$ is the cost parameter. Revenue scales with x_b, because more evaluators on a benchmark means more demand for benchmark-relative information. With aggregate investment I, report noise is $\varepsilon_i \sim N(0, \sigma^2(I))$ where $\sigma^2(I) = \sigma_0^2/(1+I)$, following the Grossman and Stiglitz (1980) tradition of diminishing returns to information investment. The specific functional forms are used for closed-form results. The main theorems require only that σ^2 be strictly decreasing and strictly convex in I and that manipulation costs be strictly convex in a with $C_{a\theta} < 0$.[2]

Key assumptions. The model rests on four assumptions, which we introduce here so the reader has the complete parameter structure before the equilibrium analysis.

The most consequential is that higher-type reporters face lower manipulation costs, an ordering we maintain throughout.

Assumption 1 ((A1) Asymmetric Manipulation Costs)**.** $m'(\theta) < 0$ for all $\theta \in (0,1)$.

The logic is compelling across applications. Firms in better financial health, healthier borrowers, and higher-quality issuers all have more operational slack to adjust reported numbers cheaply. (A1) generalizes the binary-type condition $m_B > m_G$ of the earnings-management literature (Arya, Glover, and Sunder, 2003) to a continuum and is the critical condition that makes manipulation informationally beneficial. When it fails, Goodhart's Law holds in full.

For the model to have interior solutions, the manipulation incentive must be bounded relative to costs at both ends of the type space.

Assumption 2 ((A2) Interior Solutions)**.** $\frac{1}{2}\, m(0) > \frac{\Delta u}{\kappa^2}$ and $\frac{1}{2}\, m(\kappa)(1-\kappa)^2 > \Delta u$.

[2]We restrict manipulation to $a \geq 0$ (upward only). If reporters could also manipulate downward, the equilibrium for types below the benchmark would be unchanged. Types in $[\tilde{\theta}(b), b)$ still manipulate upward, and types below $\tilde{\theta}(b)$ still report truthfully. Under the asymmetric-cost condition (A1), high types face the lowest manipulation costs and the incentive near b points upward, so the restriction is without loss for the types that drive all screening and selection results. A full analysis of bidirectional manipulation is a possible extension.

The first inequality says that the lowest-quality reporters cannot afford to manipulate their way to the benchmark, since the distance is too great and the cost too high, guaranteeing that the marginal type $\tilde{\theta}(b)$ is strictly positive. The second ensures that the evaluator's screening problem has an interior solution (Proposition 1). Both are mild. They hold jointly whenever κ is bounded away from 0 and 1 and Δu is not too large relative to manipulation costs, the empirically relevant parameter region.

Information producers are competitive and identical in cost structure.

Assumption 3 ((A3) Competitive Information Producers)**.** Information producers maximize profits, taking prices as given.

This is deliberately spare. It yields a clean information supply function (Lemma 1) and isolates demand as the economic driver, since what matters is that demand for benchmark-relative information scales with evaluator coordination, not that any individual producer has market power. The assumption also has empirical support, since many analysts cover the same firm and compete on forecast accuracy. The Online Appendix shows the results extend to strategic (Cournot) producers.

Finally, information costs are benchmark-specific.

Assumption 4 ((A4) Benchmark-Specific Information Costs)**.** $\kappa_I(b) = \kappa_0 + \gamma(b) \cdot \kappa_1$,

where $\kappa_0 > 0$ is a base cost, $\gamma(b) \geq 0$ measures the *complexity* of benchmark b, and $\kappa_1 \geq 0$ scales the cost of producing information about complex benchmarks. A simple benchmark (zero earnings) has $\gamma \approx 0$, and a complex one (the analyst consensus forecast) has $\gamma > 0$. This assumption drives the benchmark-transition results in Section 5.2.

One last piece of notation before we turn to the equilibrium analysis. The *information premium* at benchmark b and evaluator share x_b is $\Lambda(b, x_b) \equiv U_E^{\text{info}}(b, x_b) - U_E^{\text{base}}(b)$, where $U_E^{\text{base}}(b) \equiv U_E(b; 0)$ is the screening payoff with no information investment. It captures the screening improvement attributable to the information ecosystem the benchmark attracts.

2.3 Manipulation Equilibrium

Given a benchmark and evaluator share, what happens? Information producers invest first, pinning down signal quality. Reporters then choose manipulation strategies.

Information investment. Because information producers are competitive and identical (A3), aggregation is immediate.

Lemma 1 (Information Production Equilibrium)**.** *Each information producer's optimal investment is* $i_l^* = r \cdot x_b/\kappa_I$*, and aggregate information investment is:*

$$I^*(b, x_b) = \frac{r \cdot x_b}{\kappa_I}. \tag{2}$$

Aggregate information investment is linear in the evaluator share x_b*.*

Proof. First-order condition of the information producer's profit. See Appendix A. □

The linearity is worth pausing on. Each additional evaluator creates the same incremental demand for information, so signal noise declines smoothly with adoption,

$$\sigma^2(x_b) = \frac{\sigma_0^2}{1 + r \cdot x_b/\kappa_I}. \tag{3}$$

At $x_b = 0$, noise sits at its base level σ_0^2, and as evaluator coordination grows, the information ecosystem sharpens the signal.

Why information concentrates at the benchmark. The evaluator's binary decision at b makes benchmark-relative information maximally valuable, a principle formalized by Persico (2000) in auctions and extended here to benchmark coordination. Because the evaluator's action changes only when s is near b, the marginal value of precision peaks at the decision boundary. Competitive producers therefore invest exclusively in benchmark-relative signals. The concentration is a market outcome, not a modeling assumption.

To be precise, under the threshold rule $d(s) = \mathbb{1}\{s \geq b\}$, the marginal value of benchmark-relative information is strictly positive ($\text{MV}_B > 0$), while the marginal value of general quality information that does not reduce report noise is zero ($\text{MV}_G = 0$). All information investment goes to the benchmark-relative topic, with $I_B^* = r \cdot x_b/\kappa_I$ and $I_G^* = 0$. This is a special case of the stronger concentration result in Proposition 4, where thresholds not only direct investment to benchmark-relative topics but concentrate it at the decision boundary in a way that strictly dominates any dispersed allocation.

Manipulation in the deterministic limit. The clearest view of the model's economics comes from the deterministic limit ($\sigma^2 \to 0$, equivalently $I \to \infty$), the case of a benchmark with maximal information investment. We build intuition here, then state the general noisy case.

Lemma 2 (Optimal Manipulation)**.** *In the deterministic limit, given benchmark $b \in (0,1)$, the reporter's optimal manipulation strategy is:*

$$a^*(\theta, b) = \begin{cases} 0 & \text{if } \theta \geq b, \\ b - \theta & \text{if } \theta \in [\tilde{\theta}(b),\, b), \\ 0 & \text{if } \theta < \tilde{\theta}(b), \end{cases} \tag{4}$$

where $\tilde{\theta}(b)$ is the unique marginal type *defined by the indifference condition:*

$$\frac{1}{2}\, m\big(\tilde{\theta}(b)\big)\big(b - \tilde{\theta}(b)\big)^2 = \Delta u. \tag{5}$$

Proof. See Appendix A. □

Three regions emerge. Types above the benchmark already pass and have no reason to manipulate. Types in an interval just below it ($\theta \in [\tilde{\theta}(b), b)$) manipulate to exactly meet it, because the benefit Δu of passing outweighs the cost $\frac{1}{2}m(\theta)(b-\theta)^2$. And types far below the benchmark cannot afford to reach it, since the distance is too great and the cost too high, so they report truthfully. The marginal type $\tilde{\theta}(b)$ is exactly indifferent. Its uniqueness follows from (A1), because higher types face lower manipulation costs, the net cost of reaching b is strictly decreasing in θ on $[0, b)$, guaranteeing a single crossing of the indifference threshold.

What does this do to the observed distribution of reports?

Lemma 3 (Bunching at the Benchmark)**.** *Under the optimal manipulation strategy of Lemma 2, the distribution of reported signals has the following structure:*

(a) A mass point *at $s = b$ with mass $F(b) - F(\tilde{\theta}(b)) > 0$.*

(b) A continuous distribution *on $[0, \tilde{\theta}(b))$ following the type distribution F.*

(c) A continuous distribution *on $(b, 1]$ following the type distribution F.*

(d) A gap*: no reported signals in the interval $(\tilde{\theta}(b), b)$.*

Proof. See Appendix A. Types in $[\tilde{\theta}(b), b)$ all report $s = b$, producing the mass point; types below $\tilde{\theta}(b)$ and above b report truthfully, producing the continuous tails; no type has an optimal report in $(\tilde{\theta}(b), b)$. □

This is the model's micro-foundation for empirical bunching. The mass point at the benchmark, the gap just below it, and the undistorted tails replicate the key features of the earnings distribution documented by Burgstahler and Dichev (1997), namely a "surplus bin" of firms reporting small profits and a "missing bin" of firms reporting small losses. The gap $(\tilde{\theta}(b), b)$ is precisely the firms that would have reported small losses absent manipulation but have instead inflated their reports to exactly zero.[3] With endogenous noise, the mass point dissolves into a smooth density with a single positive peak near b, matching the empirical pattern, but the economic forces survive intact.

The smooth manipulation equilibrium. The deterministic case is a useful limiting benchmark, but the argument is more subtle when noise is positive. With $\sigma^2(I) > 0$, the reporter with type θ facing benchmark b chooses $a \geq 0$ to maximize the expected payoff

$$\mathbb{E}[\pi_R \mid \theta, a] = \Delta u \cdot \Phi\left(\frac{\theta + a - b}{\sigma(I)}\right) + u_L - \frac{1}{2}\, m(\theta)\, a^2, \tag{6}$$

where Φ is the standard normal CDF, and the reporter passes with probability $\Phi((\theta + a - b)/\sigma(I))$.

Lemma 4 (Smooth Manipulation Equilibrium)**.** *For $\sigma(I) > 0$, the reporter's optimal manipulation $a^*(\theta, b, I)$ is characterized by:*

$$\frac{\Delta u}{\sigma(I)} \cdot \phi\left(\frac{\theta + a - b}{\sigma(I)}\right) = m(\theta) \cdot a, \tag{7}$$

where ϕ is the standard normal density. The solution has the following properties:

(a) All types manipulate to some degree: $a^* > 0$ for all $\theta \in [0, 1]$, since even types above b face a positive marginal benefit from further increasing their passing probability. The manipulation is economically significant only for types near or below b; for types well above b, a^* is negligible (vanishing as $\sigma \to 0$).

[3] See Figure 1 for a graphical illustration.

(b) The total reported signal $\theta + a^*$ is strictly increasing in θ: higher types produce higher reports. The manipulation amount a^* converges to the deterministic pattern $a^* = b - \theta$ on $[\tilde{\theta}(b), b)$ as $\sigma \to 0$, which is strictly decreasing in θ. For $\sigma > 0$, a^* vanishes far below and far above b and attains a positive maximum in between, but a sharp characterization of the maximizer requires further parametric assumptions and is not needed for any downstream result.

(c) The effect of noise on manipulation is type-dependent. For types within one standard deviation of b (i.e., $|z^*| < 1$), a^* is decreasing in $\sigma(I)$: lower noise sharpens the pass/fail boundary, increasing *the marginal return to manipulation for types very near b. For types farther below b (i.e., $|z^*| > 1$), a^* is increasing in $\sigma(I)$: lower noise reduces their already-small passing probability,* reducing *the return to manipulation. Despite the mixed individual-level effects, the net effect on the* evaluator's screening quality *is unambiguously positive (Lemma 5).*

(d) The smooth solution converges to the bang-bang solution of Lemma 2 as $\sigma(I) \to 0$.

Proof. See Appendix A. □

The information loop offsets Goodhart's Law not because manipulation uniformly falls, but because lower noise sharpens the evaluator's ability to classify manipulators versus non-manipulators (Lemma 5). Under positive noise, the deterministic mass point at b dissolves into the smooth density with a single positive peak near b documented by Burgstahler and Dichev (1997), while the underlying economic forces survive intact.

2.4 The Evaluator's Problem

What benchmark should an evaluator choose? And what happens when many evaluators face the same question simultaneously?

The evaluator's screening payoff given benchmark b and information investment I takes the form

$$U_E(b; I) = \mathbb{E}\big[\theta - \kappa \;\big|\; s \geq b,\, a^*(\cdot, b, I)\big] \cdot \Pr\big(s \geq b \;\big|\; a^*(\cdot, b, I)\big). \tag{8}$$

In the deterministic limit, this simplifies to the integral

$$U_E(b) = \int_{\tilde{\theta}(b)}^{1} (\theta - \kappa)\, f(\theta)\, d\theta, \tag{9}$$

where the evaluator accepts all types $\theta \geq \tilde{\theta}(b)$, those who either genuinely exceed the benchmark or manipulate to reach it. What emerges from this structure is a clean characterization of the optimal threshold.

Proposition 1 (Optimal Benchmark)**.** *The evaluator's optimal benchmark b^* satisfies:*

$$\tilde{\theta}(b^*) = \kappa. \tag{10}$$

That is, the optimal benchmark is set so that the marginal manipulating type equals the evaluator's breakeven type κ.

Proof. See Appendix A. The evaluator's payoff is maximized when the lowest type admitted (the marginal type $\tilde{\theta}(b)$) exactly equals the breakeven type κ, because admitting types above κ is profitable and admitting types below κ is costly. Uniqueness follows from the strict monotonicity of $\tilde{\theta}(b)$ in b. □

Since $\tilde{\theta}(b) < b$, the optimal benchmark lies above the breakeven type, so $b^* > \kappa$. The evaluator deliberately sets the threshold higher than she would absent manipulation, compensating for the upward inflation by marginal types. Comparative statics (b^* increasing in κ and Δu, decreasing in $m(\kappa)$) are in Appendix A. For positive noise, an analogous FOC applies with $b^*(\sigma^2) \to b^*$ as $\sigma^2 \to 0$, and b^* hereafter refers to the individually optimal benchmark absent coordination, the maximizer of $U_E^{\text{base}}(b) \equiv U_E(b; 0)$.

Multiplicity. The information channel becomes central once evaluator coordination is allowed. When more evaluators adopt the same benchmark, more information producers invest in signals around it (Lemma 1), reducing noise and improving screening. The evaluator's payoff from using b when a share x_b already uses it is

$$U_E^{\text{info}}(b, x_b) \equiv U_E\left(b;\, \frac{r \cdot x_b}{\kappa_I}\right), \tag{11}$$

strictly increasing in x_b (Online Appendix Figure 1). The self-reinforcing loop is derived from the information market rather than assumed, with the formal derivation

in Section 3, and it generates multiplicity.

Proposition 2 (Multiple Equilibria)**.** *Any benchmark* $\hat{b} \in [\underline{b}, \bar{b}]$ *can be sustained as a* monomorphic equilibrium *(all evaluators using* $\hat{b}$*), where*

$$\underline{b} \approx b^* - \sqrt{\frac{2\Lambda(b^*, 1)}{|U_E''(b^*)|}}, \qquad \bar{b} \approx b^* + \sqrt{\frac{2\Lambda(b^*, 1)}{|U_E''(b^*)|}}, \tag{12}$$

with b^* *the optimal benchmark from Proposition 1,* $U_E''(b^*) < 0$*, and* $\Lambda(b, x_b) \equiv U_E^{\text{info}}(b, x_b) - U_E^{\text{base}}(b)$ *the* information premium *—the screening improvement attributable to information investment attracted by evaluator coordination.*

Proof. See Appendix A. At a monomorphic equilibrium with benchmark $\hat{b}$, a deviator loses the information premium $\Lambda(\hat{b}, 1)$ (since no one else uses the deviator's benchmark, it attracts no information investment). The equilibrium holds as long as this information premium exceeds the screening-quality loss from using $\hat{b}$ instead of the evaluator's individually optimal b^*. □

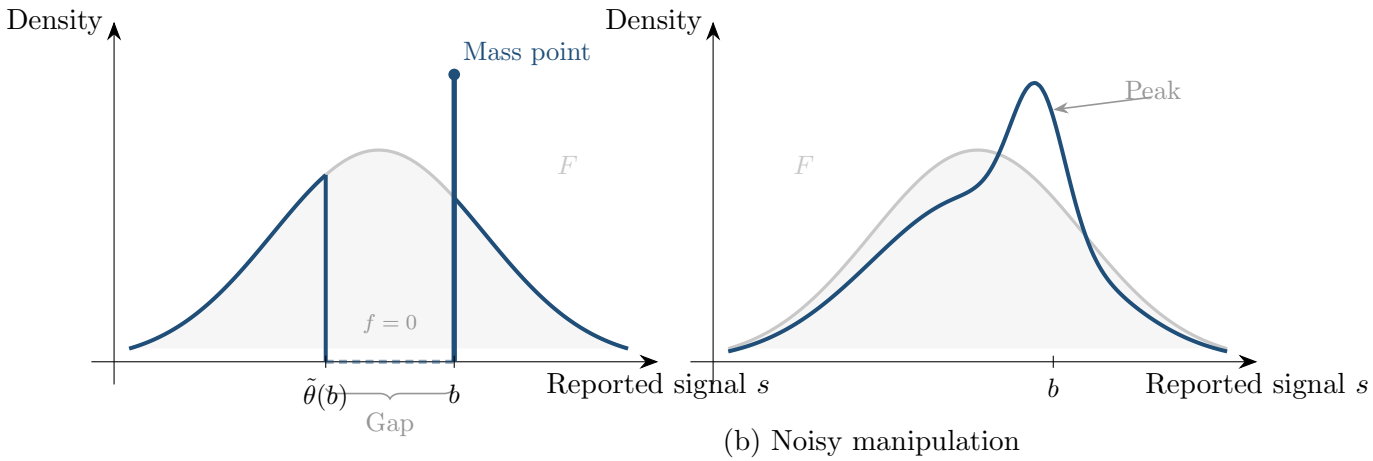
The upshot is a continuum of equilibria, and even a suboptimal benchmark persists if it attracts enough information investment to compensate for its screening deficiency. The mechanism is the same in every case, since the benchmark generates an information ecosystem that justifies its own existence. But which benchmark actually wins? Section 4 uses evolutionary dynamics to answer that question.

3 The Information Feedback Loop

Section 2.4 stated the coordination externality as a property, namely that U_E^{info} is increasing in x_b. This section derives it from first principles and shows that the loop is self-reinforcing for economic reasons rather than by assumption. We first address a prior question, why evaluators use thresholds at all.

3.1 Why Evaluators Use Thresholds

A threshold rule is endogenously optimal by the Karlin-Rubin monotone-decision argument applied to a binary action space, with no role for the information cost function. The evaluator's payoff is strictly increasing in θ and the posterior mean is



(a) Deterministic manipulation

(b) Noisy manipulation

Figure 1: Distribution of reported signals under the manipulation equilibrium. Panel (a): deterministic manipulation creates a mass point at b, a gap on $(\tilde{\theta}(b), b)$, and the undistorted distribution elsewhere. Panel (b): noisy manipulation dissolves the mass point into a smooth density with a single positive peak near b, consistent with the earnings discontinuity documented by Burgstahler and Dichev (1997).

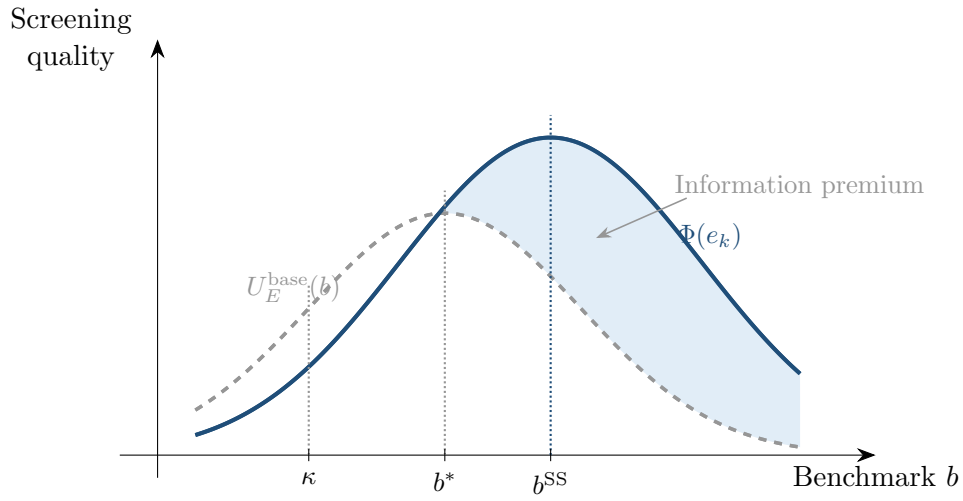


Figure 2: The potential landscape. The dashed curve shows raw screening quality $U_E^{\text{base}}(b)$, which peaks at b^*. The solid curve shows induced screening quality (the potential evaluated at monomorphic states), which accounts for information investment attracted by the benchmark. The shaded region represents the information premium. The stochastically stable benchmark b_{SS} maximizes induced, not raw, screening quality (Theorem 1).

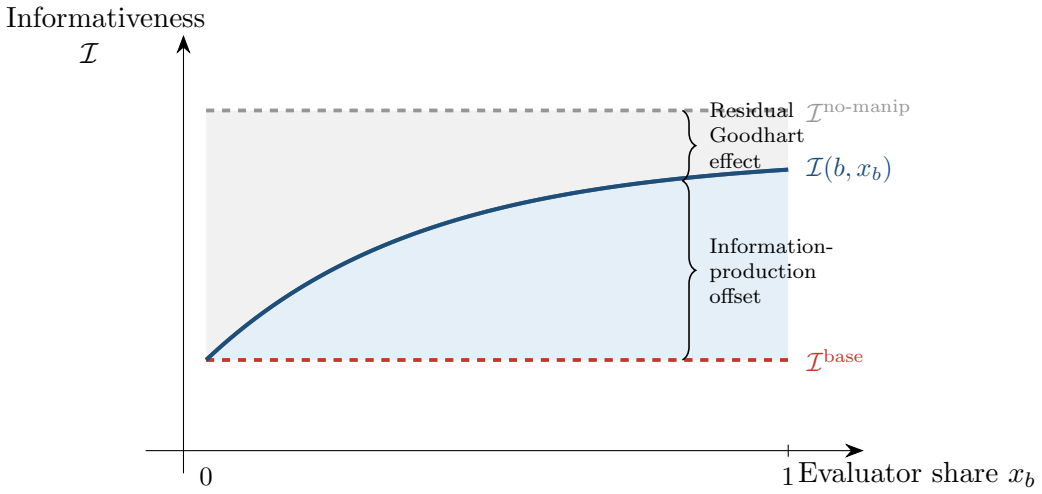


Figure 3: Goodhart's Law moderation. The dashed gray line shows informativeness without manipulation. The dashed red line shows informativeness with manipulation but without endogenous information production—Goodhart's Law in full force. The solid blue line shows informativeness with both manipulation and endogenous information production: as evaluator coordination x_b increases, information investment partially offsets the manipulation-induced degradation. The gap between the gray and blue lines is the residual Goodhart effect; the gap between the red and blue lines is the information-production offset (Proposition 6).

monotone in any affiliated signal, so the optimal binary decision is a threshold in the signal.

Proposition 3 (Optimality of Threshold Evaluation)**.** *Suppose the evaluator chooses a decision rule $d : \mathbb{R} \to \{0, 1\}$ to maximize $\mathbb{E}[(\theta - \kappa) \cdot d(s)]$, where s is any signal affiliated with θ. Then the optimal rule is a threshold: $d^*(s) = \mathbb{1}\{s \geq b\}$ for a unique b satisfying $\mathbb{E}[\theta \mid s = b] = \kappa$.*

Proof. $\mathbb{E}[\theta - \kappa \mid s]$ is strictly increasing in s by affiliation. Accept iff $\mathbb{E}[\theta \mid s] \geq \kappa$, with b unique by strict monotonicity. □

The threshold creates the information concentration that the next subsection derives, and the resulting concentration justifies the threshold. Threshold evaluation and concentrated information are a single self-consistent equilibrium phenomenon amplified by the coordination externality of Section 3.3.

3.2 Why Thresholds Create Information

Standard information theory says that coarsening destroys information, because a binary pass/fail partition discards everything a continuous signal contains about where the reporter lies relative to the evaluator's ideal. This assumes information is exogenous. When information production is endogenous, the opposite can hold. A threshold rule concentrates decision sensitivity at b, and information producers invest there because that is where their product changes outcomes (Section 2.3), while under a continuous rule $p(s)$ decision sensitivity spreads across the signal space and each point receives thin investment. Persico (2000) formalized this principle in auctions, and our setting extends it, since each additional evaluator using the same threshold adds demand for benchmark-relative information, amplifying the concentration into a collective phenomenon.

The key question is whether concentrating a fixed information budget at one point produces a more informative signal than spreading it. The answer is yes by Jensen's inequality, since the noise technology $\sigma^2(I) = \sigma_0^2/(1 + I)$ is strictly convex.[4] Concentrating all investment $\bar{I}$ at the decision boundary yields noise $\sigma^2(\bar{I})$, while spreading it yields average noise strictly above.

[4]Strict convexity follows from $\sigma^{2\prime\prime}(I) = 2\sigma_0^2/(1 + I)^3 > 0$, and the result extends to $\sigma^2(I) = \sigma_0^2/(1 + I)^\alpha$ for $\alpha \in (0, 1]$ (Grossman and Stiglitz, 1980; Veldkamp, 2006).

Proposition 4 (Information Concentration)**.** *Under the noise technology* $\sigma^2(I) = \sigma_0^2/(1+I)$ *and competitive information production (A3), consider two evaluation rules that attract the same total information investment* $\bar{I} > 0$*:*

(i) *A* threshold rule $d(s) = \mathbb{1}\{s \geq b\}$*, under which all investment concentrates at* b*: signal noise at the decision boundary is* $\sigma^2(\bar{I})$*.*

(ii) *A* continuous rule $p : \mathbb{R} \to [0,1]$ *with* p *Lipschitz and non-degenerate* ($p'(s) > 0$ *on a set of positive measure), under which investment is dispersed:* $\eta(s) \propto |p'(s)| \cdot g(s)$ *with* $\int \eta(s)\, ds = \bar{I}$ *and* η *non-constant.*

Then:

(a) ***Noise comparison.*** *The threshold rule achieves strictly lower noise than the decision-weighted average noise under the continuous rule:*

$$\sigma^2(\bar{I}) < \frac{\int \sigma^2(\eta(s))\, |p'(s)|\, g(s)\, ds}{\int |p'(s')|\, g(s')\, ds'}.$$

The weight $|p'(s)|g(s)/\int |p'|g$ *is the natural decision-relevance weight, which assigns mass in proportion to where the continuous rule's outcomes actually depend on the signal.*

(b) ***Screening payoff dominance.*** *The evaluator's screening payoff at the optimally chosen threshold* b *is strictly higher than the screening payoff under any continuous rule with the same total investment:*

$$U_E^{\text{thresh}}(b;\, \bar{I}) > U_E^{\text{cont}}(p;\, \eta).$$

Proof sketch. Part (a) is Jensen's inequality applied to the strictly convex $\sigma^2(\cdot)$ with non-constant η. Part (b) combines Part (a) with the single-crossing structure of the screening payoff, since U_E is strictly decreasing in noise at the decision boundary and the continuous rule's intermediate acceptance probabilities create an additional screening loss. Full proof in Online Appendix Section OA.3.1. □

Proposition 4 holds the total information budget fixed, so the threshold wins by allocating investment better rather than attracting more. Strict convexity of the noise technology is necessary and sufficient, with concave noise technologies reversing the dominance. The comparison is decision-relevant rather than Blackwell-dominant.

3.3 The Self-Reinforcing Loop

With the concentration result in hand, the central claim follows.

Lemma 5 (Information Feedback Loop)**.** *There exists $x_* \in [0,1)$ such that the evaluator's induced screening payoff $U_E^{\text{info}}(b, x_b)$ satisfies*

$$\frac{\partial U_E^{\text{info}}}{\partial x_b}(b, x_b) > 0 \qquad \text{for } x_b \in [x_*, 1]. \tag{13}$$

Moreover, $U_E^{\text{info}}(b, x_b) \geq U_E^{\text{base}}(b)$ for every $x_b \in [0,1]$, so the path-integrated payoff $\int_0^1 U_E^{\text{info}}(b, \rho)\, d\rho > U_E^{\text{base}}(b)$ whenever $x_* < 1$.

Proof. See Appendix B.3. The chain rule splits $\partial U_E^{\text{info}}/\partial x_b$ into a direct precision channel and an indirect manipulation-response channel. The direct channel is positive because lower noise raises passing probability for types above b and lowers it for types below κ. The indirect channel is bounded by the curvature of the manipulation cost (equation (28), satisfied by standard parametric families). On the high-share region $[x_*, 1]$ the direct channel dominates and the partial derivative is strictly positive. On the low-share region the path-integral inequality $U_E^{\text{info}}(b, x_b) \geq U_E^{\text{base}}(b)$ follows directly from noise being weakly decreasing in I. □

Lemma 5 establishes a genuine, micro-founded coordination externality, the self-reinforcing loop depicted in Online Appendix Figure 1. Unlike ad hoc coordination specifications (e.g., $\Lambda = \lambda x_b$), the information channel (i) varies across benchmarks, (ii) enters through the screening payoff rather than as an additive bonus, and (iii) exhibits diminishing returns because $\sigma^2(I)$ is convex in I.

The information premium

$$\Lambda(b, x_b) = U_E^{\text{info}}(b, x_b) - U_E^{\text{base}}(b) = U_E\left(b;\, \frac{r \cdot x_b}{\kappa_I}\right) - U_E(b; 0) \tag{14}$$

is increasing in x_b with $\Lambda(b, 0) = 0$ by Lemma 5, confirming the properties posited in Section 2.2. Its benchmark-specificity is what makes the information channel matter for selection. A generic linear coordination payoff λx_b with λ constant across benchmarks would drop out of the potential comparison entirely, leaving raw screening quality as the only criterion. Standard focal-point models cannot generate this specificity.

4 Equilibrium Selection

4.1 The Selection Problem

Proposition 2 leaves us with a continuum of sustainable benchmark equilibria. Any threshold in a wide interval can survive if it attracts a sufficiently rich information ecosystem. The evolutionary approach asks which monomorphic state evaluators keep returning to as the noise in their choices vanishes. We discretize the benchmark space as $\mathcal{B} = \{b_1, \ldots, b_n\}$ with population state $\mathbf{x}(\tau) \in \Delta^{n-1}$, with the formal machinery in Appendix C.

4.2 The Potential Game Structure

The selection argument rests on a structural property of the coordination game that is not assumed but derived from the information production technology.

Proposition 5 (Potential Game Structure)**.** *The evaluator coordination game with payoffs* $U_E^{\text{info}}(b_k, x_k)$ *is an exact population potential game* (Sandholm, 2010) *with potential function:*

$$\Phi(\mathbf{x}) = \sum_{k=1}^{n} \int_0^{x_k} U_E^{\text{info}}(b_k, \rho)\, d\rho. \tag{15}$$

Proof. See Appendix C.1. The key is that $U_E^{\text{info}}(b_k, x_k)$ depends only on own share x_k, making the payoff Jacobian diagonal and hence symmetric—the Monderer–Shapley condition for an exact potential (Monderer and Shapley, 1996). □

The result is the paper's principal game-theoretic contribution. The economic content is the own-share property, that $U_E^{\text{info}}(b_k, x_k)$ depends only on x_k because information producers invest in each benchmark independently. This is not an assumption but an endogenous consequence of the information production technology, and decentralized information markets, by concentrating investment at the benchmark that evaluators actually use, generate the potential game structure that enables sharp equilibrium selection. Robustness to cross-benchmark spillovers and to strategic information producers is in the Online Appendix.

4.3 Main Result: The Stochastically Stable Benchmark

With the potential structure in hand, the selection is immediate.

Theorem 1 (The Stochastically Stable Benchmark)**.** *Generically (i.e., for all parameter values outside a set of Lebesgue measure zero), the unique stochastically stable benchmark maximizes the potential function at monomorphic states:*

$$b_{\text{SS}} = \arg\max_{b_k \in \mathcal{B}} \ \Phi(\mathbf{e}_k) = \arg\max_{b_k \in \mathcal{B}} \int_0^1 U_E^{\text{info}}(b_k, \rho)\, d\rho. \tag{16}$$

Equivalently:

$$b_{\text{SS}} = \arg\max_{b_k \in \mathcal{B}} \left[\underbrace{U_E^{\text{base}}(b_k)}_{\text{raw screening quality}} + \underbrace{\int_0^1 \Lambda(b_k, \rho)\, d\rho}_{\text{average information premium}} \right], \tag{17}$$

where $\Lambda(b_k, \rho) = U_E^{\text{info}}(b_k, \rho) - U_E^{\text{base}}(b_k)$ *is the information premium at evaluator share* ρ*.*

The decomposition (17) is the paper's main selection result. The stochastically stable benchmark maximizes the sum of raw screening quality $U_E^{\text{base}}(b_k)$ and the average information premium $\int_0^1 \Lambda(b_k, \rho)\, d\rho$, so a benchmark with mediocre raw quality but high average information premium can dominate one with superior raw quality but a thin ecosystem.

Proof. See Appendix C.2. The potential game structure (Proposition 5) yields the integral representation at monomorphic states; the stochastic stability bridge (Section 4.3) selects the potential maximizer; the decomposition into raw screening quality and information premium follows by linearity. Uniqueness holds generically (ties are codimension-one). □

The integral averages screening payoff over the entire adoption path, not just at full adoption. Stochastic stability asks which benchmark can recover after a shock, and recovery requires attracting information even when the evaluator base is thin. A benchmark with a rich ecosystem only at scale is fragile, while one that attracts investment even from a small base of adopters is resilient. The path integral $\int_0^1 \Lambda(b_k, \rho)\, d\rho$ captures this bootstrapping capacity (Figure 2).

The result generalizes beyond benchmarks to any coordination game with strategy-specific externalities (Online Appendix Section OA.4).

Risk dominance and global games (Online Appendix). In the two-benchmark case, the potential maximizer coincides with the population-game risk-dominant benchmark, defined as the benchmark with the larger path-integrated payoff (Sandholm, 2010, Chapter 11), the population-game analog of Harsanyi and Selten (1988) basin-size risk dominance. The model's strictly concave own-share payoff U_E^{info} rules out the textbook basin-size equivalence, and the path-integral criterion is the one that aligns with stochastic stability in our setting. A complementary global-games argument in the spirit of Carlsson and van Damme (1993) and Frankel, Morris, and Pauzner (2003) delivers the same selection in the noise-vanishing two-benchmark case, sketched in Online Appendix Section OA.2.3.

Strategic producers and cross-benchmark spillovers (Online Appendix). Replacing (A3) with $N \geq 2$ symmetric Cournot information producers under a linear-inverse-demand cost parameter κ_I^C scales aggregate investment to $N/(N+1) \cdot rx_b/\kappa_I^C$ but preserves the own-share property, so the potential game structure and Theorem 1 are unchanged for any $N \geq 2$. The own-share property could fail if information at one benchmark partially reduced noise at another. The Online Appendix bounds such cross-benchmark effects by a spillover parameter δ and shows that the stochastically stable benchmark is exactly preserved provided δ is small relative to the potential gap Δ, a mild condition given empirical analyst specialization across benchmark types.

Example 1 (Analyst Forecast vs. Zero Earnings)**.** Zero earnings (b_1) has high raw screening quality but a low information premium—there is little that information producers can add to a simple profit/loss test. The analyst forecast (b_2) has lower raw screening quality (the forecast is noisy and moves over time) but a large information premium: analyst coverage, whisper numbers, and revision tracking all improve screening substantially. The analyst forecast displaces zero earnings when its information advantage outweighs zero earnings' raw screening advantage: $\int_0^1 \Lambda(b_2, \rho)\, d\rho - \int_0^1 \Lambda(b_1, \rho)\, d\rho > U_E^{\text{base}}(b_1) - U_E^{\text{base}}(b_2)$. This is precisely the historical pattern as information markets developed through the 1980s–1990s (see Figure 2).

The proof bridges population potential games and stochastic stability via log-linear response dynamics rather than the uniform mutations of Kandori, Mailath, and Rob (1993) and Young (1993), which do not in general select the potential maximizer here. Under log-linear (logit) response, an evaluator chooses b_k with probability

proportional to $\exp(\beta\, U_E^{\text{info}}(b_k, x_k))$, and as $\beta \to \infty$ the unique stochastically stable state in a finite potential game is the global potential maximizer (Blume, 1993; Young, 1998). The grid-refinement extension to a continuum of benchmarks follows Sandholm (2010) and is in Appendix C. Convergence is exponential at a rate determined by the payoff gap, and transitions triggered by structural changes in information costs exhibit delay, sudden switching amplified by the loop, and hysteresis (Remarks 2–3, Appendices C.3–C.3).

4.4 Continuous Benchmark Spaces

The body discretizes the benchmark space to a finite grid. In practice the benchmark space is a continuum, and Oechssler and Riedel (2001) show that evolutionary dynamics on continuous spaces can behave pathologically. Lahkar, Mukherjee, and Roy (2024) establish that for general continuum potential games the stochastically stable outcome converges to the potential maximizer as the discretization refines. Our contribution adds that the potential game structure arises endogenously from the information production technology, and we characterize the selected benchmark through a variational first-order condition, a transition barrier decomposition, and monotone comparative statics that use the economic structure of the model rather than just the abstract potential.

Theorem 2 (Continuous-Space Stochastic Stability)**.** *Consider the benchmark coordination game on the compact strategy space $\mathcal{B} = [\underline{b}, \bar{b}]$, with evaluator payoff $U_E^{\text{info}}(b, x_b)$ continuous on $\mathcal{B} \times [0,1]$ and strictly increasing in x_b, and monomorphic states the unique absorbing states of the unperturbed best-response dynamics on every grid $\mathcal{B}_n$. Define the* potential functional

$$\mathcal{V}(b) \;\equiv\; \int_0^1 U_E^{\text{info}}(b,\rho)\, d\rho \;=\; U_E^{\text{base}}(b) + \int_0^1 \Lambda(b,\rho)\, d\rho. \tag{18}$$

Then:

(a) ***Convergence of discrete selection.*** *For any sequence of grids $\mathcal{B}_n = \{b_1^n, \ldots, b_n^n\}$ with mesh $\|\mathcal{B}_n\| \to 0$, the stochastically stable benchmark b_{SS}^n of the discretized game converges to $b^* \equiv \arg\max_{b\in\mathcal{B}} \mathcal{V}(b)$, with the limit independent of the discretization sequence.*

(b) ***Variational characterization.*** *If U_E^{info} is C^1 in b and $b^* \in \text{int}(\mathcal{B})$, then b^* satisfies*

$$\frac{dU_E^{\text{base}}}{db}(b^*) + \int_0^1 \frac{\partial \Lambda}{\partial b}(b^*, \rho)\, d\rho = 0. \tag{19}$$

(c) ***Transition barrier.*** *Let b' denote the challenger benchmark (the alternative with the highest potential after b^*). The expected time to exit the basin of attraction of b^* satisfies $\log \mathbb{E}[\tau(b^* \to b')] \sim (N/\varepsilon)[\mathcal{V}(b^*) - \mathcal{V}(b')]$ as $N/\varepsilon \to \infty$, with the barrier decomposing as*

$$\mathcal{V}(b^*) - \mathcal{V}(b') = \left[U_E^{\text{base}}(b^*) - U_E^{\text{base}}(b')\right] + \int_0^1 \left[\Lambda(b^*, \rho) - \Lambda(b', \rho)\right] d\rho. \tag{20}$$

A benchmark with a thin information ecosystem faces a low barrier to displacement even if its raw screening quality is high.

Remark 1 (Monotone Comparative Statics)**.** Suppose $U_E^{info}(b, x_b; \alpha)$ depends on a parameter α (information cost, market development) and $\partial^2 \mathcal{V}/\partial b\, \partial \alpha > 0$. Then $b^*(\alpha) = \arg\max_b \mathcal{V}(b; \alpha)$ is increasing in α by Topkis's theorem. Applied to benchmark complexity, as information costs fall (κ_1 decreases) the information premium of complex benchmarks grows faster than that of simple benchmarks (A4), generating the increasing-differences condition and providing a monotone foundation for Proposition 7 of Section 5.2.

The proof is in Online Appendix Section OA.3.11.

5 Implications

Theorem 1 does not just select a benchmark, it reshapes how we think about three phenomena that the existing literature treats as independent puzzles.

5.1 Goodhart's Law Moderation

Goodhart's Law says that once a measure becomes a target, it ceases to be a good measure. The standard logic stops on the demand side. Reporters below the benchmark manipulate to appear above it, and this gaming degrades screening quality. The omitted force is the supply side. Adoption attracts information investment that

reduces noise and sharpens the evaluator's ability to distinguish high types from manipulators. Neither force fully wins.

Define the informativeness of benchmark b at evaluator share x_b as the gap in average quality between accepted and rejected reporters,

$$\mathcal{I}(b, x_b) \equiv \mathbb{E}[\theta \mid s \geq b,\, a^*(\cdot, b, I^*)] - \mathbb{E}[\theta \mid s < b,\, a^*(\cdot, b, I^*)], \tag{21}$$

with $I^* = I^*(b, x_b) = r \cdot x_b / \kappa_I$.

Proposition 6 (Goodhart's Law Moderation)**.** *At the stochastically stable benchmark b_{SS} with full evaluator adoption ($x_{b_{\text{SS}}} = 1$):*

(a) ***Goodhart's Law holds in principle:*** *Compared to a hypothetical benchmark that attracts the same information investment but induces no manipulation, the stochastically stable benchmark is less informative:*

$$\mathcal{I}(b_{\text{SS}}, 1) < \mathcal{I}^{\text{no-manip}}(b_{\text{SS}}, 1). \tag{22}$$

(b) ***Information production moderates:*** *The benchmark's informativeness under full adoption exceeds what it would be without the information ecosystem:*

$$\mathcal{I}(b_{\text{SS}}, 1) > \mathcal{I}(b_{\text{SS}}, 0) = \mathcal{I}^{\text{base}}(b_{\text{SS}}). \tag{23}$$

The information premium in informativeness, $\Delta\mathcal{I} = \mathcal{I}(b_{\text{SS}}, 1) - \mathcal{I}(b_{\text{SS}}, 0) > 0$, captures the degree to which the endogenous information ecosystem offsets the manipulation-induced degradation.

(c) ***Informativeness remains strictly positive:***

$$\mathcal{I}(b_{\text{SS}}, 1) > 0. \tag{24}$$

Asymmetric manipulation costs (A1) ensure positive selection—reporters who manipulate to pass are, on average, higher quality—and endogenous information investment sharpens the signal further.

Proof. See Appendix D.1. Part (a) follows from the pool-composition argument in Lemma 3 (manipulators in $[\tilde{\theta}(b_{\text{SS}}), b_{\text{SS}})$ pollute the accepted pool with types below the no-manipulation conditional mean). Parts (b)–(c) follow from Lemma 5 and (A1). □

Under $F \sim U[0,1]$, $m(\theta) = m_0(1-\theta)$, $\sigma_0 = 0.3$, and $r/\kappa_I = 2$, numerical evaluation yields $\mathcal{I}^{\text{no-manip}} \approx 0.52$, $\mathcal{I}^{\text{base}} \approx 0.31$, and $\mathcal{I}(b_{\text{SS}}, 1) \approx 0.44$, an offset fraction of about 0.62 that ranges from 0.25 to 0.80 as r/κ_I varies over $[0.5, 5]$ (see Figure 3). The ranking $\mathcal{I}^{\text{base}} < \mathcal{I}(b_{\text{SS}}, 1) < \mathcal{I}^{\text{no-manip}}$ is general.

5.2 Benchmark Transitions

Benchmarks do not last forever. As information-processing costs decline, through technological progress, the expansion of the analyst profession, the development of data infrastructure, the balance between raw screening quality and information premium shifts, and the stochastically stable benchmark can change abruptly.

Recall from (A4) that benchmark complexity $\gamma(b) \geq 0$ enters the information cost as $\kappa_I(b) = \kappa_0 + \gamma(b) \cdot \kappa_1$, where κ_1 captures the cost of producing information about complex benchmarks. The equilibrium information investment is $I^*(b, x_b) = r \cdot x_b / [\kappa_0 + \gamma(b) \cdot \kappa_1]$.

Proposition 7 (Benchmark Evolution)**.** *Consider two benchmarks b_s (simple, $\gamma(b_s) \approx 0$) and b_c (complex, $\gamma(b_c) > 0$), and suppose the simple benchmark has superior raw screening quality: $U_E^{\text{base}}(b_s) > U_E^{\text{base}}(b_c)$, while the complex benchmark is more informationally responsive (higher marginal value of information at equal investment levels). As information processing costs decrease (i.e., as κ_1 decreases):*

(a) ***Simple to complex.*** *The stochastically stable benchmark shifts toward more complex benchmarks. Formally, there exist thresholds κ_1^* such that for $\kappa_1 > \kappa_1^*$, b_{SS} is a simple benchmark, while for $\kappa_1 < \kappa_1^*$, b_{SS} is a complex benchmark.*

(b) ***The transition is sudden.*** *There is no gradual "blending" of benchmarks across the threshold κ_1^*. The transition is a discrete jump in the identity of b_{SS}, reflecting the self-reinforcing nature of the information loop: once a complex benchmark becomes viable, it attracts information investment that further enhances its screening quality, tipping the selection abruptly.*

The stochastically stable benchmark maximizes the potential $\Phi(\mathbf{e}_k) = \int_0^1 U_E^{info}(b_k, \rho)\, d\rho$, which decomposes as:

$$\Phi(\mathbf{e}_k) = \underbrace{U_E^{base}(b_k)}_{\textit{raw screening quality}} + \underbrace{\int_0^1 \Lambda(b_k, \rho)\, d\rho}_{\textit{average information premium}} . \tag{25}$$

For a simple benchmark b_s with $\gamma(b_s) \approx 0$, the information premium is relatively insensitive to κ_1. For a complex benchmark b_c with $\gamma(b_c) > 0$, the information premium grows rapidly as κ_1 falls. The transition occurs when the complex benchmark's growing information premium overcomes the simple benchmark's raw screening advantage.

Proof. See Appendix D.2. Part (a) follows from the intermediate value theorem applied to the potential gap $\Phi(\mathbf{e}_c) - \Phi(\mathbf{e}_s)$ as κ_1 varies. Part (b) follows from transversal crossing of the potential functions. □

At the tipping point $\kappa_1 = \kappa_1^*$, a small further decrease triggers a cascade in which information producers shift investment, screening quality tips, and evaluators follow (Online Appendix Figure 2). Section 6.2 maps this to the historical shift from zero earnings to analyst forecasts. The same simple-to-complex pattern appears in debt covenant design moving from balance-sheet to income-statement covenants (Demerjian, 2011), in Basel capital regulation layering buffers atop a single ratio, in credit scoring expanding from FICO to multidimensional frameworks, and in ESG screens consolidating into integrated ratings, each transition driven by falling information costs and the reorganization of the ecosystem around a more complex benchmark. The mechanism is symmetric. Information cost increases (Reg FD, MiFID II) can revert complex to simple (Agrawal, Chadha, and Chen, 2006), and the empirical coexistence of zero earnings and analyst forecasts is consistent with a transition still in progress, segmented evaluator types, or a small potential gap that lets both persist as metastable states.

Welfare under decentralized selection. The stochastically stable benchmark maximizes the evaluator's path-integrated screening payoff but not in general social welfare. Let $b^W \equiv \arg\max_b W(b, 1)$ denote the social planner's optimal benchmark. Evaluators do not internalize manipulation costs, and information producers capture only private revenue, so the evolutionary criterion partially internalizes the social objective without matching it. Under the parametric illustration of Section 5.1, numerical computation yields $W(b_{\text{SS}}, 1)/W(b^W, 1) \approx 0.89$, ranging from 0.72 to 0.96 across the grid in Online Appendix Table 1, with the gap rising in r/κ_I and $m(\kappa)$ and falling in Δu. Preliminary numerical evidence suggests that subsidies for information production close a substantial fraction of the gap, while mandating the welfare-maximizing benchmark requires parameters the planner is unlikely to observe. The

formal welfare proposition is in Appendix D.3.

6 Applications

The formal structure is abstract by design. This section puts institutional flesh on it by mapping reporters, evaluators, and information producers to observable populations in two domains where the theory's predictions are testable, credit markets and financial markets. The mapping is the same in each case. A reporter with private type θ faces a benchmark b, manipulation costs satisfy (A1), and information producers invest in signals near the threshold because that is where decisions turn. What differs is who plays each role and what the information ecosystem looks like. Two further domains, education and regulatory compliance, work in parallel and are developed in Online Appendix Section OA.6.

6.1 Credit Markets

Mapping. Loan applicants are reporters with type $\theta \in [0, 1]$ the true creditworthiness. The reported credit score is the manipulable signal. Lender cutoffs at 620 (conforming) and 740 (best rate) are the benchmarks. Credit bureaus, scoring vendors (Fair Isaac, VantageScore), underwriting consultants, and credit-monitoring services are information producers, investing in scoring model development, dispute-resolution infrastructure, data acquisition, and underwriting analytics.

(A1) holds. Stronger applicants can shift the score on the margin by paying down balances, closing inactive lines, or timing new accounts. Applicants with impaired credit have fewer levers and the levers that exist demand liquid resources they do not have, so the marginal cost of pushing the score up is decreasing in true creditworthiness.

Information ecosystem and predictions. Scoring vendors update their models more frequently in the ranges where lender decisions turn, and underwriting expertise concentrates at the 620 and 740 boundaries, exactly as Lemma 5 predicts. The model predicts (i) bunching at 620 with thinning of density just below, (ii) denser scoring-model revisions and dispute volumes near 620 and 740 than elsewhere, and (iii) persistence of these specific cutoffs rather than nearby alternatives reflecting the

depth of the installed ecosystem at historical focal numbers, not a feature of the underlying credit risk distribution.

6.2 Financial Markets

Earnings benchmarks. The most extensively documented benchmark-beating phenomenon is the discontinuity in reported earnings, with Burgstahler and Dichev (1997) documenting bunching at zero and DeGeorge, Patel, and Zeckhauser (1999) at the analyst consensus. The model maps directly. Unmanaged earnings are the type θ, the benchmark is zero earnings (simple) or the analyst forecast (complex), and firms with higher unmanaged earnings face lower manipulation costs through greater operational flexibility, satisfying (A1). Sell-side analysts, data providers (I/B/E/S, Bloomberg), financial media, and forensic accountants are information producers.

The historical transition from zero earnings to the analyst forecast illustrates Proposition 7. Before I/B/E/S launched in 1976, no standardized forecast data existed and zero earnings dominated because it required minimal information infrastructure. The subsequent expansion of analyst coverage built a deep information ecosystem around the consensus forecast, and by the late 1990s firms meeting or beating it earned a significant quarterly return premium over those that narrowly missed (Bartov, Givoly, and Hayn, 2002). The transition was sudden, consistent with the tipping-point prediction.

Other financial benchmarks. The model applies to covenant thresholds in syndicated lending (current ratio of 2.0, debt-to-equity of 0.5) where many lenders coordinate on common cutoffs, predicting more standardized covenants in syndicated than bilateral loans and the migration toward income-statement covenants documented by Demerjian (2011). At the investment-grade boundary (BBB−/BB+), CDS liquidity, analyst coverage, and media attention all peak, and institutional investor charter restrictions create a large mass coordinated on BBB−. However, (A1) is less compelling for credit ratings since the AAA-versus-BBB cost ordering is not obvious, making this application more speculative.

Identification. The sharpest test exploits exogenous shocks to the information ecosystem. Brokerage closures reduce analyst coverage for affected firms and allow a difference-in-differences test that the analyst-forecast benchmark weakens once the ecosystem thins, predicting weaker bunching, smaller meet-or-beat premia, and

lower screening informativeness afterward. The EPA's differential mandate of continuous emissions monitoring across pollutants generates a parallel test that monitored thresholds should exhibit sharper bunching and better regulatory screening than unmonitored ones (Burgstahler and Dichev, 1997; Saez, 2010).

7 Conclusion

The conventional understanding of thresholds as information-destroying devices is incomplete. When information production is endogenous, a threshold concentrates the information ecosystem at the decision boundary, and the resulting self-reinforcing loop, in which more evaluators bring more information investment, better screening, and still more evaluators, explains why specific benchmarks emerge, why agents bunch at them, and why evaluators persist with benchmarks they know are manipulated. Goodhart's Law follows as a corollary rather than a contradiction. The gaming a benchmark invites is sustained by the very information infrastructure that the coarsening creates, and the benchmark that survives evolutionary selection is the one whose coarsening generates the richest information response, making the gaming visible, interpretable, and ultimately informative.

The practical implication is that the design of thresholds should account not only for what they measure but for what information ecosystems they attract. Any institution that coarsens evaluation to a threshold, from educational testing to regulatory standards to financial reporting, may be better understood as an information-creation device than as a measurement compromise. The mechanism applies wherever coordination generates strategy-specific externalities, from platform adoption to standard-setting to institutional design. The surviving institution is not the one with the best intrinsic properties but the one that generates the richest ecosystem along the path to adoption. And as information costs fall over time, the same selection dynamics push surviving benchmarks toward greater complexity, a pattern visible in the multiplication of covenant dimensions, the layering of Basel capital ratios, the integration of alternative-data inputs into credit risk, and the consolidation of single-issue ESG screens into multidimensional ratings.

A standard critique of stochastic stability is that the predicted long-run distribution can take a very long time to assert itself (Ellison, 2000). The framework here is partially insulated. The relevant dynamic is log-linear response rather than rare

uniform mutation, with convergence to the Gibbs measure governed by the precision parameter and the payoff gap rather than the population size alone. The information feedback loop amplifies transitions, since a benchmark gaining share attracts investment that sharpens the signal and attracts further adoption, concentrating movements between basins into short bursts. The historical displacement of zero earnings by the analyst consensus forecast unfolded over roughly two decades from 1976 through the late 1990s, an interval consistent with the model's transition dynamics rather than the slow-convergence critique.

The most important limitation is the assumption that evaluators are homogeneous in the breakeven type κ. If evaluators differ (risk-averse lenders versus risk-tolerant venture capitalists), the model predicts segmentation rather than universal convergence, with the information loop operating within each segment and selection applying segment by segment. Formalizing this as a multi-population potential game is the natural next step.

Data and Code Availability

This paper is purely theoretical. Numerical illustrations in Section 5.1 and Online Appendix Table 1 are computed from the model's closed-form expressions; the calculations are reproducible from the parametric assumptions stated in the relevant sections. No external data were used and no proprietary code was developed.

A Proofs for Section 2

A.1 Proof of Lemma 2 (Optimal Manipulation)

Sketch. The reporter's payoff $\pi_R(\theta, a, b) = R(\theta + a, b) - \frac{1}{2}m(\theta)a^2$ has a discontinuous reward at $s = b$, so the optimum is bang-bang. Any $a \in (0, b - \theta)$ pays cost without crossing the threshold; any $a > b - \theta$ pays excess cost above the minimum needed to reach b. The two candidate strategies are $a = 0$ and $a = b - \theta$, and the latter is preferred whenever the net manipulation cost $\phi(\theta) \equiv \frac{1}{2}m(\theta)(b - \theta)^2$ does not exceed the reward spread Δu. Differentiating shows $\phi'(\theta) = (b-\theta)\left[\frac{1}{2}m'(\theta)(b-\theta) - m(\theta)\right] < 0$ for $\theta \in (0, b)$ under (A1), so ϕ is strictly decreasing on $[0, b)$ and crosses Δu at most once. Combined with (A2), which ensures $\phi(0) > \Delta u$, this yields a unique interior marginal type $\tilde{\theta}(b) \in (0, b)$ with optimal manipulation $a^*(\theta, b) = (b - \theta) \cdot \mathbb{1}\{\theta \in [\tilde{\theta}(b), b)\}$. The full proof is in Online Appendix Section OA.3.3. □

A.2 Proof of Lemma 6 (Comparative Statics of the Marginal Type)

Lemma 6 (Comparative Statics of the Marginal Type)**.** *The marginal type $\tilde{\theta}(b)$ defined by $\frac{1}{2}\, m(\tilde{\theta})(b - \tilde{\theta})^2 = \Delta u$ satisfies:*

(a) $\tilde{\theta}(b)$ is continuously differentiable in b.

(b) $\tilde{\theta}'(b) > 0$: higher benchmarks require higher marginal types.

(c) $\tilde{\theta}$ is increasing in overall manipulation cost and decreasing in Δu.

Proof. The marginal type is defined implicitly by $G(\tilde{\theta}, b) \equiv \frac{1}{2}\, m(\tilde{\theta})(b - \tilde{\theta})^2 - \Delta u = 0$.

Part (a): Continuous differentiability. By the Implicit Function Theorem, $\tilde{\theta}(b)$ is continuously differentiable wherever $\partial G/\partial \tilde{\theta} \neq 0$. We compute:

$$\frac{\partial G}{\partial \tilde{\theta}} = \frac{1}{2}\, m'(\tilde{\theta})(b - \tilde{\theta})^2 + m(\tilde{\theta}) \cdot (-(b - \tilde{\theta})) = (b - \tilde{\theta}) \left[\frac{1}{2}\, m'(\tilde{\theta})(b - \tilde{\theta}) - m(\tilde{\theta})\right].$$

Since $\tilde{\theta} < b$ (so $b - \tilde{\theta} > 0$), $m'(\tilde{\theta}) < 0$ (by A1), and $m(\tilde{\theta}) > 0$, every term in the bracket is strictly negative. Thus $\partial G/\partial \tilde{\theta} < 0$, and the IFT applies.

Part (b): $\tilde{\theta}'(b) > 0$. We have:

$$\frac{\partial G}{\partial b} = m(\tilde{\theta})(b - \tilde{\theta}) > 0.$$

By the IFT:

$$\tilde{\theta}'(b) = -\frac{\partial G/\partial b}{\partial G/\partial \tilde{\theta}} = -\frac{m(\tilde{\theta})(b - \tilde{\theta})}{(b - \tilde{\theta})\left[\frac{1}{2}\, m'(\tilde{\theta})(b - \tilde{\theta}) - m(\tilde{\theta})\right]} = \frac{m(\tilde{\theta})}{m(\tilde{\theta}) - \frac{1}{2}\, m'(\tilde{\theta})(b - \tilde{\theta})}.$$

The numerator $m(\tilde{\theta}) > 0$. The denominator is $m(\tilde{\theta}) + \frac{1}{2}|m'(\tilde{\theta})|(b - \tilde{\theta}) > 0$ (since $m' < 0$ implies $-\frac{1}{2}m'(\tilde{\theta})(b - \tilde{\theta}) = \frac{1}{2}|m'(\tilde{\theta})|(b - \tilde{\theta}) > 0$). Hence $\tilde{\theta}'(b) > 0$.

Part (c): Comparative statics with respect to $m(\cdot)$ and Δu. Suppose $m(\cdot)$ is replaced by $\lambda\, m(\cdot)$ for $\lambda > 1$ (scaling up all manipulation costs). The indifference condition becomes $\frac{1}{2}\lambda\, m(\tilde{\theta})(b - \tilde{\theta})^2 = \Delta u$. At the original $\tilde{\theta}$, the left side exceeds Δu, so the new marginal type must increase to restore equality (since ϕ is decreasing). By the IFT applied to $G(\tilde{\theta}, \lambda) = \frac{1}{2}\lambda\, m(\tilde{\theta})(b - \tilde{\theta})^2 - \Delta u$:

$$\frac{\partial \tilde{\theta}}{\partial \lambda} = -\frac{\frac{1}{2}m(\tilde{\theta})(b - \tilde{\theta})^2}{\partial G / \partial \tilde{\theta}} > 0,$$

since $\frac{1}{2}m(\tilde{\theta})(b - \tilde{\theta})^2 = \Delta u / \lambda > 0$ and $\partial G / \partial \tilde{\theta} < 0$.

For Δu: the condition $G(\tilde{\theta}, \Delta u) = \frac{1}{2}m(\tilde{\theta})(b - \tilde{\theta})^2 - \Delta u = 0$ gives

$$\frac{\partial \tilde{\theta}}{\partial (\Delta u)} = -\frac{-1}{\partial G / \partial \tilde{\theta}} = \frac{1}{\partial G / \partial \tilde{\theta}} < 0,$$

since $\partial G / \partial \tilde{\theta} < 0$. A larger reward spread lowers the marginal type, expanding the set of manipulators. □

A.3 Proof of Lemma 3 (Bunching Distribution)

Sketch. Lemma 2 partitions types into three regions: $\theta \in [0, \tilde{\theta}(b))$ report truthfully, $\theta \in [\tilde{\theta}(b), b)$ manipulate to exactly b, and $\theta \geq b$ report truthfully. The induced CDF H of reported signals therefore equals F on $[0, \tilde{\theta}(b))$, is flat on $[\tilde{\theta}(b), b)$ (no type produces a signal in this interval), jumps at $s = b$ by the mass $F(b) - F(\tilde{\theta}(b)) > 0$ contributed by manipulators, and equals F on $(b, 1]$. The four parts of the lemma read directly from this decomposition, with a mass point at b, a gap on $(\tilde{\theta}(b), b)$, and undistorted tails. The full proof is in Online Appendix Section OA.3.4. □

A.4 Proof of Proposition 1 (Optimal Benchmark)

Proof. The evaluator's expected payoff from benchmark b is

$$U_E(b) = \int_{\tilde{\theta}(b)}^{1} (\theta - \kappa)\, f(\theta)\, d\theta.$$

The evaluator accepts all reporters with $s \geq b$, which under the manipulation equilibrium of Lemma 2 means all types $\theta \geq \tilde{\theta}(b)$ (types in $[\tilde{\theta}(b), b)$ manipulate to $s = b$, and types $\theta \geq b$ report truthfully). Each accepted type yields net payoff $\theta - \kappa$.

First-order condition. By the Leibniz integral rule:

$$\frac{dU_E}{db} = -\big(\tilde{\theta}(b) - \kappa\big) \cdot f\big(\tilde{\theta}(b)\big) \cdot \tilde{\theta}'(b).$$

Since $f(\tilde{\theta}(b)) > 0$ (density is positive on the interior) and $\tilde{\theta}'(b) > 0$ (Lemma 6(b)), the FOC $dU_E/db = 0$ requires $\tilde{\theta}(b) = \kappa$.

Second-order condition. Differentiating again:

$$\frac{d^2U_E}{db^2} = -[\tilde{\theta}'(b)]^2 \, f(\tilde{\theta}(b)) - (\tilde{\theta}(b) - \kappa) \, \frac{d}{db}\big[f(\tilde{\theta}(b)) \, \tilde{\theta}'(b)\big].$$

At the optimum $\tilde{\theta}(b^*) = \kappa$, the second term vanishes:

$$U_E''(b^*) = -[\tilde{\theta}'(b^*)]^2 \, f(\kappa) < 0,$$

confirming a strict local maximum.

Uniqueness. Since $\tilde{\theta}(\cdot)$ is strictly increasing and continuous (Lemma 6), $\tilde{\theta}(b) = \kappa$ has at most one solution. For existence: as $b \downarrow 0$, $\tilde{\theta}(b) \to 0$; as $b \uparrow 1$, $\tilde{\theta}(b) \to \tilde{\theta}(1)$. The second inequality of (A2) gives $\frac{1}{2} \, m(\kappa)(1-\kappa)^2 > \Delta u$, which implies $\phi(\kappa)\big|_{b=1} > \Delta u$, and since ϕ is strictly decreasing, $\tilde{\theta}(1) > \kappa$. Hence $\kappa \in (0, \tilde{\theta}(1))$, and by the IVT there exists a unique b^* with $\tilde{\theta}(b^*) = \kappa$.

Global optimality. As $b \to 0^+$, $\tilde{\theta}(b) \to 0 < \kappa$, so $dU_E/db = -(\tilde{\theta}(b) - \kappa) f(\tilde{\theta}(b)) \tilde{\theta}'(b) > 0$ (the payoff is increasing). As $b \to 1$, $\tilde{\theta}(b) > \kappa$, so $dU_E/db < 0$ (the payoff is decreasing). Since there is a unique interior critical point with a strictly negative second derivative, b^* is the global maximum. □

Comparative statics of b^*. The optimal benchmark condition $\tilde{\theta}(b^*) = \kappa$, combined with Lemma 6, yields the following comparative statics via the Implicit Function Theorem.

(i) *b^* is increasing in κ:* Differentiating $\tilde{\theta}(b^*(\kappa)) = \kappa$ with respect to κ: $\tilde{\theta}'(b^*) \, \frac{db^*}{d\kappa} = 1$, so $\frac{db^*}{d\kappa} = 1/\tilde{\theta}'(b^*) > 0$.

(ii) *b^* is increasing in Δu:* From $\tilde{\theta}(b^*, \Delta u) = \kappa$: $\tilde{\theta}'(b^*) \, \frac{\partial b^*}{\partial(\Delta u)} + \frac{\partial \tilde{\theta}}{\partial(\Delta u)} = 0$. By Lemma 6(c), $\partial\tilde{\theta}/\partial(\Delta u) < 0$, so $\frac{\partial b^*}{\partial(\Delta u)} > 0$. Higher manipulation incentives require a higher threshold to maintain screening quality.

(iii) *b^* is decreasing in the cost level $m(\kappa)$:* A uniform scaling $m \mapsto \lambda m$ raises $\tilde{\theta}$ at every b (Lemma 6(c)), requiring b^* to decrease to restore $\tilde{\theta}(b^*) = \kappa$. Formally, $\frac{\partial b^*}{\partial \lambda} = -\frac{\partial\tilde{\theta}/\partial\lambda}{\tilde{\theta}'(b^*)} < 0$ since $\partial\tilde{\theta}/\partial\lambda > 0$.

A.5 Proof of Proposition 2 (Multiple Equilibria)

Sketch. At a monomorphic equilibrium using $\hat{b}$, an evaluator using $\hat{b}$ earns $U_E^{\text{base}}(\hat{b}) + \Lambda(\hat{b}, 1)$, while a unilateral deviator to b' earns $U_E^{\text{base}}(b')$ because the deviator is the sole user of b' and so attracts no information investment (i.e., $\Lambda(b', 0) = 0$). The no-deviation condition reduces to $\Lambda(\hat{b}, 1) \geq U_E^{\text{base}}(b^*) - U_E^{\text{base}}(\hat{b}) \equiv D(\hat{b})$. Because $\Lambda(b^*, 1) > 0 = D(b^*)$ and both are continuous, the sustainable set $\mathcal{S} = \{b : \Lambda(b, 1) \geq D(b)\}$ is an interval containing b^*. A second-order expansion $D(\hat{b}) \approx \frac{1}{2}|U_E''(b^*)|(\hat{b} - b^*)^2$ combined with $\Lambda(\hat{b}, 1) \approx \Lambda(b^*, 1)$ yields the explicit bounds in (12). The full proof is in Online Appendix Section OA.3.7. □

B Proofs for the Information Feedback Loop

The proof of Proposition 4 is in Online Appendix Section OA.3.1. The remaining proofs are below.

B.1 Proof of Lemma 1 (Equilibrium Information Investment)

Sketch. The information producer's profit $\pi_I(i_l, x_b) = r\, x_b\, i_l - \frac{1}{2}\kappa_I\, i_l^2$ is strictly concave in i_l, so the unique first-order condition $r\, x_b = \kappa_I\, i_l$ delivers $i_l^* = r\, x_b/\kappa_I$. With a unit mass of identical producers under (A3), aggregating across producers gives $I^*(b, x_b) = r\, x_b/\kappa_I$. Linearity in x_b inherits directly from the linear-quadratic structure of revenue and cost. The full proof is in Online Appendix Section OA.3.2. □

B.2 Proof of Lemma 4 (Manipulation Under Endogenous Noise)

Sketch. With noise $\sigma(I) > 0$, the reporter's payoff $V(\theta, a) = \Delta u\, \Phi(z(a)) + u_L - \frac{1}{2}m(\theta)a^2$ is smooth in a, where $z(a) = (\theta + a - b)/\sigma$. The interior first-order condition (7) balances the marginal benefit of pushing the passing probability up against the quadratic marginal cost; the second-order condition holds under (A2). Part (a) follows because $\phi > 0$ everywhere, so $\partial V/\partial a > 0$ at $a = 0$ for every type. Part (b)'s monotonicity of the *total report* $\theta + a^*$ uses the IFT on the FOC and balances the proximity-to-threshold and cost-monotonicity effects, with the bound $|da^*/d\theta| < 1$ following from $m(\theta) > |m'(\theta)| \cdot a$ for types sufficiently close to b. Part (c) computes $\partial H/\partial\sigma = (\Delta u/\sigma^2)\phi(z)(z^2 - 1)$, which changes sign at $|z^*| = 1$ and delivers the type-dependent comparative statics. Part (d) follows because $\Phi(z/\sigma) \to \mathbb{1}\{z \geq 0\}$ as $\sigma \to 0$, reducing the smooth FOC to the bang-bang solution of Lemma 2. The full proof is in Online Appendix Section OA.3.5. □

B.3 Proof of Lemma 5 (Self-Reinforcing Information Loop)

Proof. By the chain rule,

$$\frac{\partial U_E^{\text{info}}}{\partial x_b} = \left.\frac{dU_E}{dI}\right|_{I=I^*} \cdot \frac{\partial I^*}{\partial x_b}, \tag{26}$$

where $U_E(b;I) = \int_0^1 (\theta-\kappa)\Pr(s \geq b \mid \theta, I) f(\theta)\, d\theta$ with $\Pr(s \geq b \mid \theta, I) = \Phi((\theta + a^* - b)/\sigma)$. The second factor is $r/\kappa_I > 0$ by Lemma 1, so it suffices to sign dU_E/dI.

Decompose into direct (precision) and indirect (manipulation response) channels:

$$\frac{dU_E}{dI} = \underbrace{\int_0^1 (\theta-\kappa)\frac{\partial \Pr}{\partial \sigma^2}\frac{d\sigma^2}{dI} f\, d\theta}_{\text{Direct}} + \underbrace{\int_0^1 (\theta-\kappa)\frac{\partial \Pr}{\partial a}\frac{\partial a^*}{\partial I} f\, d\theta}_{\text{Indirect}}. \tag{27}$$

Direct effect, region by region. Let $\psi(\theta, I) \equiv \partial \Pr / \partial \sigma^2$ with $\Pr(s \geq b \mid \theta, I) = \Phi((\theta + a^* - b)/\sigma(I))$. Holding a^* fixed at equilibrium, $\psi = -(\theta + a^* - b)\phi(z^*)/(2\sigma^3)$, so $\psi > 0$ when $\theta + a^* < b$ and $\psi < 0$ when $\theta + a^* > b$. With $d\sigma^2/dI < 0$, the direct integrand $D(\theta) \equiv (\theta - \kappa)\,\psi\,(d\sigma^2/dI)\, f(\theta)$ partitions the type space into three regions. On $R_1 = \{\theta + a^* > b,\ \theta > \kappa\}$, $D > 0$. On $R_3 = \{\theta + a^* < b,\ \theta < \kappa\}$, $D > 0$. On $R_2 = \{\theta \in [\kappa, b),\ \theta + a^* < b\}$, $D < 0$. The integral is therefore positive iff R_1, R_3 contributions dominate the R_2 contribution.

Dominance on the high-share region. Let x_* be the smallest x_b at which $\sigma^2(I^*(b, x_b)) = \sigma_0^2/(1 + r x_b/\kappa_I)$ is small enough that $|R_2| \cdot \sup_{R_2} |\psi| < (|R_1 \cup R_3|/2) \cdot \inf_{R_1 \cup R_3} |\psi|$ at the evaluator's optimum b^*. At b^*, the marginal type $\tilde{\theta}(b^*) = \kappa$ (Proposition 1), so manipulating types in $[\kappa, b)$ satisfy $\theta + a^*(\theta) \to b$ as $\sigma \to 0$ by Lemma 4(d), driving $|\psi| \to 0$ uniformly on R_2. On R_1, $\theta + a^* > b$ by a margin that does not vanish with σ, so $|\psi|$ stays bounded away from zero. On R_3, $|\psi|$ decays exponentially in $|z^*|$ but the R_3 integrand magnitude is bounded above and dominated by the R_1 contribution for σ below the threshold defining x_*. Thus the net direct effect is strictly positive on $[x_*, 1]$.

Indirect effect. At the reporter's FOC, $\phi(z^*)/\sigma = m(\theta)a^*/\Delta u$, so the indirect integrand factors as $(\theta - \kappa)(m(\theta)a^*/\Delta u)(\partial a^*/\partial I)$. By IFT, $|\partial a^*/\partial I| \leq C\,|d\sigma^2/dI|$ for a finite C whenever

$$\sup_{\theta \in [0, \bar{b}]} |m''(\theta)/m'(\theta)| < \infty \tag{28}$$

on the compact $[0, \bar{b}]$ with $\bar{b} < 1$. On $[x_*, 1]$ the indirect channel is of the same order as the direct channel and under (28) its absolute value is bounded by half the direct channel's lower bound on R_1, so equation (26) yields $\partial U_E^{\text{info}}/\partial x_b > 0$ on $[x_*, 1]$. Online Appendix Section OA.3.6 sub-partitions $R_1 \cup R_3$ to resolve the indirect-channel sign flip at $|z^*| = 1$ (Lemma 4(c)) directly rather than through the magnitude bound.

Path-integral monotonicity at low shares. For $x_b \in [0, x_*)$, σ is close to σ_0 and the sign of $\partial U_E^{\text{info}}/\partial x_b$ is not pinned down by the dominance argument. The lemma's

weaker claim, $U_E^{\text{info}}(b, x_b) \geq U_E^{\text{base}}(b)$ for all x_b, follows from σ^2 being weakly decreasing in I (hence in x_b) combined with U_E being weakly decreasing in σ^2. The path integral $\int_0^1 U_E^{\text{info}}(b, \rho)\, d\rho$ used by Theorem 1 therefore strictly exceeds $U_E^{\text{base}}(b)$ because monotonicity holds on the positive-measure subinterval $[x_*, 1]$. □

C Proofs for Equilibrium Selection

C.1 Proof of Proposition 5 (Potential Game Structure)

Proof. We verify the Monderer–Shapley condition for population potential games, following Sandholm (2010). A population game with payoff functions $\{F_k(\mathbf{x})\}_{k=1}^n$ is an exact population potential game if there exists a C^1 function $\Phi : \Delta^{n-1} \to \mathbb{R}$ such that

$$\frac{\partial \Phi}{\partial x_k}(\mathbf{x}) = F_k(\mathbf{x}) \qquad \text{for all } k \text{ and all } \mathbf{x} \in \Delta^{n-1}. \tag{29}$$

In our game, $F_k(\mathbf{x}) = U_E^{\text{info}}(b_k, x_k)$. Define

$$\Phi(\mathbf{x}) = \sum_{k=1}^{n} \int_0^{x_k} U_E^{\text{info}}(b_k, \rho)\, d\rho.$$

By the Fundamental Theorem of Calculus:

$$\frac{\partial \Phi}{\partial x_k} = \frac{\partial}{\partial x_k} \int_0^{x_k} U_E^{\text{info}}(b_k, \rho)\, d\rho = U_E^{\text{info}}(b_k, x_k).$$

For $j \neq k$:

$$\frac{\partial^2 \Phi}{\partial x_j \partial x_k} = \frac{\partial}{\partial x_j} U_E^{\text{info}}(b_k, x_k) = 0,$$

since $U_E^{\text{info}}(b_k, x_k)$ depends only on x_k, not on x_j. This confirms the Monderer–Shapley integrability condition (the payoff Jacobian is symmetric—in fact, diagonal—as required for an exact potential).

The potential Φ is C^1 since $U_E^{\text{info}}(b_k, \cdot)$ is continuous (by the continuity of the screening payoff in $\sigma^2(I)$ and the smoothness of the manipulation equilibrium), and the integral of a continuous function is C^1. □

C.2 Proof of Theorem 1 (The Stochastically Stable Benchmark)

This proof bridges from finite-player stochastic stability results to the population game setting through a sequence of approximations.

Proof. The proof proceeds in six steps: establishing the finite approximation, verifying its potential structure, applying the stochastic stability theorem, taking the large-population limit, refining the benchmark grid, and establishing uniqueness.

Order of limits. The construction involves three limits: logit precision $\beta \to \infty$ (Gibbs concentration), population $N \to \infty$ (Riemann sum to integral), and grid mesh $\|\mathcal{B}_n\| \to 0$ (refinement). We take them in the order $\beta \to \infty$, $N \to \infty$, mesh $\to 0$, the standard order in stochastic-stability arguments (Blume, 1993; Young, 1998). By Heine-Cantor, U_E^{info} is uniformly continuous on the compact $\mathcal{B} \times [0,1]$, so the Riemann-sum potential converges uniformly in b_k and the maximum over a refining grid converges to the maximum over $\mathcal{B}$. Details are in the proof of Theorem 2 (Online Appendix Section OA.3.11).

Step 1: Finite-population approximation. Consider a finite population of N evaluators, each choosing a benchmark from the finite set $\mathcal{B} = \{b_1, \ldots, b_n\}$. The state is a vector $\mathbf{n} = (n_1, \ldots, n_n) \in \{0, 1, \ldots, N\}^n$ with $\sum_k n_k = N$, where n_k is the number of evaluators using benchmark b_k. The population share is $\mathbf{x}^N = \mathbf{n}/N$.

Each evaluator's payoff from benchmark b_k when n_k evaluators (including herself) use it is

$$\pi_k^N(\mathbf{n}) = U_E^{\text{info}}\left(b_k,\, \frac{n_k}{N}\right).$$

We specify the learning rule as log-linear response, sometimes called logit response. In each period one evaluator is selected uniformly at random to revise her strategy. She selects benchmark b_k with probability

$$\Pr(b_k \mid \mathbf{n}) = \frac{\exp\big(\beta\, U_E^{\text{info}}(b_k,\, n_k/N)\big)}{\sum_{l=1}^{n} \exp\big(\beta\, U_E^{\text{info}}(b_l,\, n_l/N)\big)},$$

where $\beta > 0$ is a precision parameter that scales how sharply the evaluator favors higher-payoff benchmarks. As $\beta \to \infty$ the rule converges to deterministic best response, and as $\beta \to 0$ the rule converges to uniform random choice. Uniform mutations of the kind used by Young (1993) can be recovered as a limit of this rule, but the limit selects the potential maximizer only when the perturbation has the log-linear structure used here.

This defines an ergodic Markov chain $\{X_t^N\}_{t \geq 0}$ on the finite state space $S_N = \{\mathbf{n} \in \mathbb{N}^n : \sum_k n_k = N\}$. The chain is irreducible (every state is reachable because every action carries strictly positive probability for finite β) and aperiodic. Hence there exists a unique stationary distribution μ_β^N.

Step 2: Potential structure of the finite game. Define the finite-population potential:

$$\Phi_N(\mathbf{n}) = \sum_{k=1}^{n} \sum_{j=1}^{n_k} U_E^{\text{info}}\left(b_k,\, \frac{j}{N}\right) \cdot \frac{1}{N}. \tag{30}$$

This is a Riemann-sum approximation of the continuous potential. We verify that

it serves as a potential for the finite game. When a single evaluator switches from b_k to b_l (so n_k decreases by 1 and n_l increases by 1):

$$\begin{aligned}\Phi_N(\mathbf{n}') - \Phi_N(\mathbf{n}) &= \frac{1}{N}\left[U_E^{\text{info}}\left(b_l,\ \frac{n_l+1}{N}\right) - U_E^{\text{info}}\left(b_k,\ \frac{n_k}{N}\right)\right] \\ &= \frac{1}{N}\left[\pi_l^N(\mathbf{n}') - \pi_k^N(\mathbf{n})\right],\end{aligned}$$

so the change in the potential equals (up to the positive constant $1/N$) the payoff difference between the new and old strategies. This confirms the Monderer and Shapley potential game property for the finite game: the potential tracks the incentive to switch strategies.

Step 3: Stochastic stability in the finite game. By Blume (1993), in a finite-player potential game under log-linear response with precision β, the unique stationary distribution is the Gibbs measure

$$\mu_\beta^N(\mathbf{n}) \ \propto\ \exp\bigl(\beta\, N\, \Phi_N(\mathbf{n})\bigr).$$

As $\beta \to \infty$ the mass of this distribution concentrates on the global maximizer of Φ_N over the state space. The textbook treatment in Young (1998, Chapter 6) connects log-linear response to potential games in the population setting used here.

Restrict the grid $\mathcal{B}$ to lie within the sustainable interval of Proposition 2. (Stochastically stable selection picks the global potential maximizer, which generically lies in the interior of the sustainable interval, so this restriction is without loss for the selected benchmark.) The absorbing states of the unperturbed (deterministic best-response) dynamics on $\mathcal{B}$ are then the monomorphic states $\mathbf{e}_k^N = (0, \ldots, N, \ldots, 0)$, since at any monomorphic state $\mathbf{e}_k^N$ the population share on b_k is one and $U_E^{\text{info}}(b_k, 1) > U_E^{\text{base}}(b_j) = U_E^{\text{info}}(b_j, 0)$ for all $j \neq k$ in $\mathcal{B}$ by the sustainability condition. Each evaluator's best response is to continue using the common benchmark.

The potential at monomorphic state $\mathbf{e}_k^N$ is

$$\Phi_N(\mathbf{e}_k^N) = \sum_{j=1}^{N} U_E^{\text{info}}\left(b_k,\ \frac{j}{N}\right) \cdot \frac{1}{N}. \tag{31}$$

By the Blume–Young selection result, the unique stochastically stable state in the limit $\beta \to \infty$ is

$$\mathbf{e}_{k^*(N)}^N \quad \text{where} \quad k^*(N) = \underset{k \in \{1,\ldots,n\}}{\arg\max}\ \Phi_N(\mathbf{e}_k^N).$$

Three conditions support this conclusion. The game is a finite-player potential game (verified in Step 2). The learning rule is log-linear response (specified in Step 1), under which the Gibbs measure is the unique stationary distribution. The state space is finite (guaranteed by finite N and finite $\mathcal{B}$). The log-linear protocol is the

load-bearing assumption. Uniform mutations of the form used in Young (1993) require a separate resistance-tree argument and do not in general identify the potential maximizer in this setting.

Step 4: Large-population limit. We show that $\Phi_N(\mathbf{e}_k^N) \to \Phi(\mathbf{e}_k)$ as $N \to \infty$. Note that $\Phi_N(\mathbf{e}_k^N)$ is the right-endpoint Riemann sum of the continuous function $\rho \mapsto U_E^{\text{info}}(b_k, \rho)$ on $[0,1]$ with partition $\{0, 1/N, 2/N, \ldots, 1\}$:

$$\Phi_N(\mathbf{e}_k^N) = \sum_{j=1}^{N} U_E^{\text{info}}\left(b_k,\, \frac{j}{N}\right) \cdot \frac{1}{N} \quad \xrightarrow{N\to\infty} \quad \int_0^1 U_E^{\text{info}}(b_k, \rho)\, d\rho = \Phi(\mathbf{e}_k).$$

The convergence holds because $U_E^{\text{info}}(b_k, \cdot)$ is continuous (and hence Riemann integrable) on $[0,1]$, which follows from the continuity of the evaluator's screening payoff in the noise level $\sigma^2(I)$ and the smoothness of the manipulation equilibrium (Lemma 4).

Step 5: Convergence of maximizers. Since there are finitely many benchmarks $\{b_1, \ldots, b_n\}$, the maximizer of $\Phi_N(\mathbf{e}_k^N)$ over k converges to the maximizer of $\Phi(\mathbf{e}_k)$ over k, provided the maximizer is unique. Formally, for each k:

$$\Phi_N(\mathbf{e}_k^N) \to \Phi(\mathbf{e}_k) \quad \text{as } N \to \infty.$$

If $k^* = \arg\max_k \Phi(\mathbf{e}_k)$ is unique, then for all sufficiently large N, $k^*(N) = k^*$. This is because for each $k \neq k^*$, the gap $\Phi(\mathbf{e}_{k^*}) - \Phi(\mathbf{e}_k) > 0$ is bounded away from zero, and $|\Phi_N(\mathbf{e}_j^N) - \Phi(\mathbf{e}_j)| \to 0$ uniformly in j.

For the grid refinement as the mesh of $\mathcal{B}$ shrinks (i.e., $n \to \infty$ with $\max_k |b_{k+1} - b_k| \to 0$), the continuity of $\Phi(\mathbf{e}_k) = \int_0^1 U_E^{\text{info}}(b_k, \rho)\, d\rho$ in b_k—which follows from the smoothness of the screening payoff $U_E(b; I)$ in b—ensures that the discrete maximizer converges to the continuous maximizer $b_{\text{SS}} = \arg\max_{b \in [\underline{b}, \bar{b}]} \int_0^1 U_E^{\text{info}}(b, \rho)\, d\rho$.

Step 6: Uniqueness (genericity condition). The stochastically stable benchmark is unique provided the potential has a strict maximizer:

$$\Phi(\mathbf{e}_{k^*}) > \Phi(\mathbf{e}_k) \quad \text{for all } k \neq k^*.$$

This is a generic condition: the set of parameters $(\kappa, m(\cdot), \Delta u, F, r, \kappa_I, \sigma_0^2)$ for which $\Phi(\mathbf{e}_j) = \Phi(\mathbf{e}_k)$ for some $j \neq k$ has measure zero in parameter space. This follows from the Transversality Theorem (see, e.g., Guillemin and Pollack 1974): the function $\Phi(\mathbf{e}_k) - \Phi(\mathbf{e}_j) = \int_0^1 [U_E^{\text{info}}(b_k, \rho) - U_E^{\text{info}}(b_j, \rho)]\, d\rho$ is a smooth function of the parameters, and ties occur on a codimension-one submanifold.

When the genericity condition holds, the argument in Steps 1–5 yields the unique stochastically stable benchmark:

$$b_{\text{SS}} = b_{k^*} = \arg\max_{b_k \in \mathcal{B}} \int_0^1 U_E^{\text{info}}(b_k, \rho)\, d\rho. \qquad \square$$

C.3 Convergence and Transition Dynamics

Remark 2 (Convergence Rate)**.** Starting from any interior initial condition in the basin of attraction of b_{SS}, the replicator dynamics converge exponentially: $\|\mathbf{x}(t) - \mathbf{e}_{k^*}\| \leq C \cdot e^{-\gamma t}$, where $\gamma > 0$ depends on the payoff gap.

Proof. The replicator dynamics are $\dot{x}_k = x_k[U_E^{\text{info}}(b_k, x_k) - \bar{U}_E(\mathbf{x})]$. At $\mathbf{e}_{k^*}$, the Jacobian eigenvalues are $\lambda_j = U_E^{\text{base}}(b_j) - U_E^{\text{info}}(b_{k^*}, 1) < 0$ for $j \neq k^*$. By Hartman–Grobman, local convergence is exponential at rate $\gamma = \min_{j \neq k^*} |\lambda_j|$. Global asymptotic stability follows from the Lyapunov function $V(\mathbf{x}) = \Phi(\mathbf{e}_{k^*}) - \Phi(\mathbf{x}) \geq 0$, which satisfies $\dot{V} < 0$ along replicator trajectories in potential games (Sandholm, 2010, Theorem 7.2.1). □

Remark 3 (Benchmark Transitions)**.** Transitions exhibit delay (waiting time $\sim \exp(N \cdot \psi)$), sudden switching (amplified by the information loop), and hysteresis (small parameter changes preserve the incumbent).

Proof. Part (a): By Freidlin and Wentzell (1998), the waiting time to escape $\mathbf{e}_k^N$ scales as $e^{N \cdot \psi_k}$, where ψ_k is the potential barrier (Blume, 1993). Part (b): Once the tipping point is exceeded, Remark 2 gives exponential convergence, amplified by the information loop reallocating investment. Part (c): ψ_k is continuous in parameters, so small changes preserve the ranking. □

D Proofs for Implications

D.1 Proof of Proposition 6 (Goodhart's Law Moderation)

Sketch. For Part (a), introducing manipulation expands the accepted set with types $\theta \in [\tilde{\theta}(b_{\text{SS}}), b_{\text{SS}})$ whose true qualities sit below the threshold and below the no-manipulation accepted conditional mean. Adding low-θ types to the accepted pool and removing them from the rejected pool widens both conditional means in the wrong direction relative to the no-manipulation benchmark, so $\mathcal{I}^{\text{no-manip}}(b_{\text{SS}}, 1) > \mathcal{I}(b_{\text{SS}}, 1)$. For Part (b), Lemma 5 delivers $\partial U_E^{\text{info}}/\partial x_b > 0$; the same single-crossing argument that drives screening quality also drives informativeness, so $\partial \mathcal{I}/\partial I > 0$ and full adoption ($I^* = r/\kappa_I > 0$) yields $\mathcal{I}(b_{\text{SS}}, 1) > \mathcal{I}(b_{\text{SS}}, 0)$. For Part (c), even at $I = 0$, asymmetric costs (A1) ensure positive selection. Types with lower manipulation costs (higher θ) are more likely to reach the threshold, so $\mathbb{E}[\theta \mid s \geq b, I = 0] > \mathbb{E}[\theta]$, and combined with Part (b) this gives $\mathcal{I}(b_{\text{SS}}, 1) > \mathcal{I}(b_{\text{SS}}, 0) > 0$. The full proof is in Online Appendix Section OA.3.8. □

D.2 Proof of Proposition 7 (Benchmark Transitions)

Sketch. Define the potential gap $\Delta\Phi(\kappa_1) \equiv \Phi(\mathbf{e}_c; \kappa_1) - \Phi(\mathbf{e}_s; \kappa_1)$ as a function of the information-cost parameter. As $\kappa_1 \to \infty$, information investment vanishes and

both potentials collapse to base screening payoffs, with $\Delta\Phi < 0$ because the simple benchmark has the superior raw screen. As $\kappa_1 \to 0$, both benchmarks attract heavy information investment, and the informational-responsiveness condition gives $\Delta\Phi > 0$. Continuity of $\Delta\Phi$ in κ_1 plus the Intermediate Value Theorem deliver the threshold κ_1^* with $\Delta\Phi(\kappa_1^*) = 0$, which is Part (a). Part (b) follows because the stochastically stable state is always monomorphic (Theorem 1), so the transition is a discrete jump at κ_1^*, and generic transversality of the crossing follows from $\gamma(b_c) > \gamma(b_s)$. Part (c) reads off the metastable regime when potential values are close, with multiple benchmarks coexisting under long residence times even though only one is stochastically stable. The full proof is in Online Appendix Section OA.3.9. □

D.3 Welfare

Definition 1 (Social Welfare)**.** Social welfare at benchmark b with evaluator share x_b sums the surplus of all three populations,

$$W(b, x_b) = U_E^{\text{info}}(b, x_b) + W_R^{\text{gross}}(b, x_b) - TC(b, x_b) - IC(b, x_b),$$

where U_E^{info} is the evaluator's screening payoff, W_R^{gross} is reporters' gross expected reward, TC is aggregate manipulation cost, and $IC = \frac{1}{2}\kappa_I (I^*)^2$ is information production cost. Producer revenue $r\, x_b\, I^*$ is a transfer from evaluators to producers and cancels in the social aggregate, so only the real resource cost IC appears.

Proposition 8 (ESS vs. Social Optimum)**.** *The stochastically stable benchmark b_{SS} does not generally maximize $W(b, 1)$, due to manipulation and information externalities.*

Sketch. The stochastically stable benchmark maximizes the evaluator's path integral $\int_0^1 U_E^{\text{info}}(b, \rho)\, d\rho$ (Theorem 1), whereas social welfare $W = U_E^{\text{info}} + W_R^{\text{gross}} - TC - IC$ additionally credits reporter surplus and debits manipulation and information costs. These extra terms vary with b through the manipulation and information equilibria, so generically their maximizers differ. Two externalities drive the wedge. The *manipulation externality* biases b_{SS} upward relative to b^W: the evaluator gains from positive selection of marginal manipulators but does not internalize their costs. The *information externality* pushes the other way: information producers capture only private revenue, but the evolutionary criterion partly internalizes the screening improvement, leaving a smaller residual gap than a naive evaluator-only objective would suggest. Parts (c) and (d) follow by definition, with $W(b^W, 1) \geq W(b_{\text{SS}}, 1)$ trivially, and the benchmark dominating the no-benchmark counterfactual when the screening benefit $\int_0^\kappa (\kappa - \theta) f(\theta)\, d\theta$ exceeds aggregate manipulation cost plus information cost less foregone reporter rewards. The full proof is in Online Appendix Section OA.3.10. □

References

Agrawal, A., S. Chadha, and M. A. Chen (2006). Who is afraid of Reg FD? The behavior and performance of sell-side analysts following the SEC's Fair Disclosure rules. *Journal of Business* 79(6), 2811–2834.

Arya, A., J. C. Glover, and S. Sunder (1998). Earnings management and the revelation principle. *Review of Accounting Studies* 3, 7–34.

Arya, A., J. C. Glover, and S. Sunder (2003). Are unmanaged earnings always better for shareholders? *Accounting Horizons* 17, 111–116.

Bartov, E., D. Givoly, and C. Hayn (2002). The rewards to meeting or beating earnings expectations. *Journal of Accounting and Economics* 33, 173–204.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature* 57, 44–95.

Blume, L. E. (1993). The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5, 387–424.

Burgstahler, D. and I. Dichev (1997). Earnings management to avoid earnings decreases and losses. *Journal of Accounting and Economics* 24, 99–126.

Carlsson, H. and E. van Damme (1993). Global games and equilibrium selection. *Econometrica* 61, 989–1018.

DeGeorge, F., J. Patel, and R. Zeckhauser (1999). Earnings management to exceed thresholds. *Journal of Business* 72, 1–33.

Demerjian, P. R. (2011). Accounting standards and debt covenants: Has the "balance sheet approach" led to a decline in the use of balance sheet covenants? *Journal of Accounting and Economics* 52, 178–202.

Dye, R. A. (1988). Earnings management in an overlapping generations model. *Journal of Accounting Research* 26, 195–235.

Ellison, G. (2000). Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *Review of Economic Studies* 67, 17–45.

Fischer, P. E. and R. E. Verrecchia (2000). Reporting bias. *The Accounting Review* 75, 229–245.

Foster, D. and H. P. Young (1990). Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38, 219–232.

Frankel, D. M., S. Morris, and A. Pauzner (2003). Equilibrium selection in global games with strategic complementarities. *Journal of Economic Theory* 108, 1–44.

Freidlin, M. I. and A. D. Wentzell (1998). *Random Perturbations of Dynamical Systems* (2nd ed.). New York: Springer.

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics, 1975*, pp. 1–20. Sydney: Reserve Bank of Australia.

Grossman, S. J. and J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *American Economic Review* 70, 393–408.

Guillemin, V. and A. Pollack (1974). *Differential Topology*. Englewood Cliffs, NJ: Prentice-Hall.

Harsanyi, J. C. and R. Selten (1988). *A General Theory of Equilibrium Selection in*

Games. Cambridge, MA: MIT Press.

Hart, O. and J. Moore (1988). Incomplete contracts and renegotiation. *Econometrica* 56, 755–785.

Holmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7, 24–52.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review* 101, 2590–2615.

Kandori, M., G. J. Mailath, and R. Rob (1993). Learning, mutation, and long run equilibria in games. *Econometrica* 61, 29–56.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics* 8, 435–464.

Lahkar, R., S. Mukherjee, and S. Roy (2024). Equilibrium selection via stochastic evolution in continuum potential games. Working paper, SSRN 4771829.

Lahkar, R. and F. Riedel (2015). The logit dynamic for games with continuous strategy sets. *Games and Economic Behavior* 91, 268–282.

Lazear, E. P. and S. Rosen (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89(5), 841–864.

Mathevet, L., J. Perego, and I. Taneva (2020). On information design in games. *Journal of Political Economy* 128(4), 1370–1404.

Moldovanu, B. and A. Sela (2001). The optimal allocation of prizes in contests. *American Economic Review* 91(3), 542–558.

Monderer, D. and L. S. Shapley (1996). Potential games. *Games and Economic Behavior* 14, 124–143.

Oechssler, J. and F. Riedel (2001). Evolutionary dynamics on infinite strategy spaces. *Economic Theory* 17, 141–162.

Persico, N. (2000). Information acquisition in auctions. *Econometrica* 68, 135–148.

Saez, E. (2010). Do taxpayers bunch at kink points? *American Economic Journal: Economic Policy* 2, 180–212.

Sandholm, W. H. (2010). *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press.

Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Stein, J. C. (1989). Efficient capital markets, inefficient firms: A model of myopic corporate behavior. *Quarterly Journal of Economics* 104, 655–669.

Veldkamp, L. L. (2006). Media frenzies in markets for financial information. *American Economic Review* 96, 577–601.

Young, H. P. (1993). The evolution of conventions. *Econometrica* 61, 57–84.

Young, H. P. (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.