

Editorial Screening when Science is Cheap*

Nic Fishman[†]

Gabriel Sekeres[‡]

April 27, 2026

Abstract

A journal screens submissions when authors can privately generate many correlated pieces of evidence and selectively disclose favorable realizations. We study a commitment game in which the journal chooses an acceptance rule, a researcher sequentially searches over evidence at marginal cost γ , and the journal observes only the report the researcher chooses to disclose. The induced operating point is a pair (q_H, q_0) , where q_H is the probability of accepting a high-impact project and q_0 is the probability of accepting a non-high project (in practice, these can be read as a true effect and a false positive). We characterize the attainable screening frontier as testing becomes cheap – the ability of the journal to distinguish between high-quality and low-quality projects.

The results separate two editorial instruments that are often conflated. Tightening standards makes researchers search harder. Requiring disclosure makes the resulting search informative to the editor. First, we prove a mechanism-independent information bound: the Bernoulli Kullback-Leibler divergence between acceptance under high and non-high projects is bounded by the likelihood-ratio information the researcher can generate before optimal stopping. This is the journal’s goal – to attain exponential screening of the submitted projects. Second, we show that capacity-scale search is necessary for exponential purification, but not sufficient. If equilibrium reports contain only $o(1/\gamma)$ disclosed results, false positives cannot fall at the dependence-adjusted exponential rate without collapsing recall. Third, we show that simple robustness-check rules requiring $m(\gamma) = \Theta(1/\gamma)$ disclosed passes in a journal-defined region achieve exponential screening while maintaining nonvanishing recall, and are rate-optimal when information and dependence scales are comparable.

We then construct auditable specification surfaces for 103 empirical economics papers published in AEA journals. The empirical section validates the cheap-testing premise and estimates the type, dependence, and cost primitives that determine counterfactual disclosure requirements. The counterfactual requirements are sobering: to maintain conventional purity at fixed throughput, the journal would need to require disclosure of about 7,000 significant robustness checks. This raises a serious issue for observational work going forward, and we argue for the need to develop methods to interpret sets of many specifications simultaneously, as opposed to current interpretative practice, which focuses on a handful of main specifications and a small set of robustness checks.

Keywords: editorial screening; robustness checks; specification search; selective disclosure; effective sample size; false discovery rate.

JEL Classification: C12; C18; C52; C72; D82; D83.

* We thank Ravi Jagadeesan, Davide Viviano, several seminar audiences, and four anonymous referees.

[†]Harvard Department of Statistics, email: njwfish@gmail.com. Fishman gratefully acknowledges the support of a National Science Foundation Graduate Research Fellowship under grant number XXX.

[‡]Cornell Department of Economics, email: gs754@cornell.edu. Part of this research was conducted when Sekeres was at the Stanford Institute for Economic Policy Research.

1 Introduction

Journals screen claims. Authors can privately generate evidence. When evidence is costly, an editor may treat a reported result as a noisy signal about the underlying project. When evidence is cheap and omissions are hard to verify, the same report has a second meaning: it is also a signal about the author’s search process. A favorable result may reflect a strong claim, or it may reflect the fact that the author searched until a favorable realization appeared and then disclosed only that realization.

This paper studies the screening problem created by cheap private evidence generation. Our leading application is empirical economics, where many claims are evaluated not by collecting new data but by varying controls, samples, functional forms, clustering rules, outcome definitions, or other defensible analytic choices. Recent progress in large language models and automated coding agents make this cost shift concrete. The marginal cost of constructing an analysis pipeline and exploring a broad robustness universe has fallen quickly. This has two mirrored effects: it makes p -hacking easier, but it also makes demanding extensive robustness checks practical.

We model this environment as a screening game. A submission has an unobserved type, either high-impact or non-high. We keep the definition of high- and not-high submissions intentionally vague, but a typical intuition would be that high-impact submissions correctly (at least directionally) estimate some underlying population treatment effect, while not-high-impact submissions are incorrect in the relevant direction. The journal commits to an acceptance rule before evidence is realized. The researcher then observes a correlated stream of private evidence, paying cost γ per draw. After any history, she may stop and submit a report consisting of any finite submultiset of the realized evidence. The journal observes the report but not the omitted evidence. The researcher wants acceptance net of testing costs. The journal’s problem is to choose an acceptance rule knowing that the report is strategically selected.

The object of interest is the screening frontier. For any editorial rule and induced researcher best response, let q_H be the probability of accepting a high-impact project and let q_0 be the probability of accepting a non-high project. These two probabilities determine throughput and the false discovery rate. Accepting nothing trivially eliminates false discoveries, so the relevant question is not whether false discoveries can be made small in isolation, but how small they can be made while sustaining recall or throughput.

The frontier is governed by two editorial levers. The first is standards: the editor can make acceptance require more favorable evidence. Conceptually, a tightening of standards would be tantamount to a field- or journal-wide Bonferroni correction. The second is disclosure: the editor can require the author to disclose many favorable checks rather than a single selected realization. These levers have different effects. Tightening standards makes the researcher search longer. Requiring disclosure makes (more of) the search observable. The central message of the paper is that, in the cheap-testing limit, standards and disclosure are not substitutes. Standards alone create a race for favorable evidence, but that race can be won by researchers with bad hypotheses if searching for favorable evidence is sufficiently cheap. Only disclosure, and disclosure increasing in the inverse cost of testing, can restore editorial screening in any reasonable way.

Our first result is a universal information bound. Any acceptance rule induces a binary experiment, acceptance or rejection. The informativeness of this binary experiment, measured by the Bernoulli Kullback-Leibler divergence between acceptance under high and non-high projects, cannot exceed the likelihood-ratio information generated by the researcher’s stopped evidence process. Since the researcher pays cost γ per test, this yields an outer envelope proportional to the total information capacity of private search, summarized by $\kappa_{\text{KL}} = D_{\text{mix}}(\theta)/\gamma$. At any nonvanishing throughput level, no editorial mechanism can make the false discovery rate decay faster than exponentially in this information budget. This is the *desideratum* for the journal – exponential screening in the

information budget.

The second result shows that effort is necessary. If a sequence of editorial rules does not induce high-impact researchers to search on the capacity scale $1/\gamma$, then false positives cannot decay at the capacity-scale exponential rate. In the cheap-testing limit, exponential purification requires a race.

The third result shows that effort is not sufficient. If equilibrium reports contain only $o(1/\gamma)$ disclosed results, then the editor cannot extract dependence-adjusted exponential separation, even if the researcher privately runs many tests. Short reports cannot transmit the information generated by the race. The relevant dependence-adjusted scale is $\kappa_{\text{eff}} = n_{\text{eff}}(\theta)/\gamma$, where $n_{\text{eff}}(\theta)$ is the effective-sample-size rate of the evidence process.

The fourth result is constructive. We study robustness-check rules that accept only when the report contains at least m disclosed passes in a preannounced evidence window. When $m(\gamma) = \Theta(1/\gamma)$, these rules force disclosure at the testing-capacity scale. Under standard concentration and posterior-tail conditions, they achieve $q_0 \leq \exp(-\Omega(\kappa_{\text{eff}}))$ while keeping q_H bounded away from zero. When the information and dependence primitives are comparable, this matches the universal information bound up to constants in the exponent. Robustness checks are therefore not merely a transparency norm. They are an incentive instrument that implements the screening frontier.

We then take the theory to data. To estimate the primitives of the game, we construct paper-specific specification surfaces, conceptually following [Gelman and Loken \(2013\)](#). A specification surface is an *ex ante* commitment object: it fixes the claim under audit, the admissible universe of estimand-preserving variants, and the inference conventions under which those variants are executed. The surface is conservative because it uses only analytical forks revealed by the paper itself. Once the surface is fixed, execution is mechanical and auditable.

We construct a constrained, auditable agentic workflow that constructs an *ex ante* specification surface for each of 103 empirical papers in AEA journals. Comparing our automated runtime to a conservative human benchmark (developed and validated from the mega-replication by [Brodeur et al. \(2026\)](#)), we estimate a roughly 170-fold decline in the marginal cost of running observational specifications. We estimate the other primitives, including type heterogeneity and dependence across similar specifications. These estimates translate directly into counterfactual disclosure requirements: under our baseline calibration, restoring a conventional screening target after the cost shift requires increasing the number of disclosed robustness checks by more than two orders of magnitude (from 50 to roughly 7,000).

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the screening game, operating points, and the two information scales. Section 4 proves the universal information bound, the no-race collapse, and the short-disclosure collapse. Section 5 shows that robustness-check rules with disclosure at the testing-capacity scale attain the optimal exponential screening rate. Section 6 takes these theoretical predictions to the data, constructing specification surfaces and using them to estimate the model primitives and counterfactual disclosure requirements under a cost shift. Section 7 discusses implications for robustness-check practice, editorial capacity, and dependence across specifications, and concludes. The appendices collect proofs, auxiliary results, and empirical implementation details.

2 Related work

This paper studies editorial screening as a commitment problem when researchers can perform cheap, sequential specification search and can omit unfavorable analyses. It contributes to four literatures: publication bias and specification search, multiple testing and post-selection inference, sequential evidence acquisition, and strategic disclosure in scientific communication.

A large literature documents selective reporting and publication bias and develops methods to detect or correct it (Ioannidis, 2005; Franco et al., 2014; Andrews and Kasy, 2019; Elliott et al., 2022). In economics and the social sciences, this problem is amplified by researcher degrees of freedom: authors make many defensible choices about samples, controls, outcomes, transformations, and inference (Leamer, 1983; Simmons et al., 2011; Gelman and Loken, 2013, 2014). Multiverse and specification-curve methods make this dependence visible by reporting distributions of estimates across alternative analyses (Steege et al., 2016; Simonsohn et al., 2020). Our specification surface is in this tradition, but the theory gives it a different role. The surface is not only a diagnostic object. It is the commitment object that defines admissible evidence before outcomes are observed.

A complementary statistical literature treats specification search as a multiple-testing or post-selection-inference problem. Classical false-discovery-rate control begins with Benjamini and Hochberg (1995), with important extensions under dependence (Benjamini and Yekutieli, 2001; Storey, 2002). Selective-inference methods characterize valid inference conditional on a selection event (Berk et al., 2013; Fithian et al., 2014; Lee et al., 2016). Econometric work on model selection, specification search, and data snooping has developed related corrections and robustness procedures (White, 2000; Hansen, 2005; Viviano et al., 2026). Our focus is different: rather than repairing inference after endogenous selection, we ask whether editorial acceptance rules can be designed to make selection and omission informationally costly, and characterize them. We additionally emphasize the role of the quality-throughput trade-off faced by editors under capacity constraints.

The model also relates to sequential testing and optimal stopping. A researcher in our game observes evidence sequentially, pays a flow cost, and chooses when to stop (Wald, 1950; Snell, 1952; Shiryaev, 1978). Our universal information bound is an information-theoretic analogue of the idea that sequential procedures can only extract bounded discrimination per unit time, and our constructions exploit large-deviation screening once disclosure scales with testing capacity.

The strategic-disclosure literature studies how informed agents reveal verifiable evidence when unfavorable evidence can be concealed (Grossman, 1981; Milgrom, 1981). Bayesian persuasion studies related commitment problems in information design (Kamenica and Gentzkow, 2011). More closely related are models of hidden testing, endogenous information acquisition, and selective disclosure (Henry, 2009; Di Tillio et al., 2017; Henry and Ottaviani, 2019; Herresthal, 2022). Our contribution is to characterize how disclosure must scale when evidence generation becomes cheap and evidence is correlated. The key distinction is between inducing private effort and extracting public information from that effort.

Recent work models scientific publication as a strategic institution with incentives, congestion, and screening (Tetenov, 2016; Andrews and Shapiro, 2021; Frankel and Kasy, 2022; Jagadeesan and Viviano, 2025; McCloskey and Michailat, 2024; Carnehl and Schneider, 2025; Spiess, 2025). We complement this literature by isolating the effect of cheap private testing. The editor does not simply choose a statistical cutoff. She designs an acceptance rule in a game where authors can search and selectively disclose. This makes robustness checks an implementation device rather than a purely statistical convention.

Finally, our empirical analysis connects to recent work on AI-assisted scientific reproduction and agentic data analysis. Bertran et al. (2026) use autonomous AI analysts to generate many analyses of the same dataset and hypothesis, showing that cheap automated analysis can reveal large analytic dispersion but also creates a selective-reporting risk. Kohler et al. (2026) study whether agents can reproduce social-science results from papers and data without original code. These papers are closely related to the measurement environment in our empirical section. Our theory abstracts from the identity of the technology and treats these developments as a fall in the marginal cost γ of private evidence generation.

3 The screening game

We study a publication game in which an editor screens a claim using evidence disclosed by a strategic researcher. The researcher can generate additional evidence at cost, observes the results privately, and may omit unfavorable realizations. The editor commits to an acceptance rule, but cannot verify what the researcher chose not to disclose.

The model has three primitives. First, submissions differ in latent quality. Second, evidence arrives through a correlated evidence process. Third, reporting is selective: after any private history, the researcher may disclose any finite submultiset of realized evidence. The central comparative static is the cheap-testing limit $\gamma \rightarrow 0$.

3.1 The Gaussian benchmark

It is useful to begin with the Gaussian case. There are two types, 0 and H , with H denoting a high-quality submission. Conditional on type t , the researcher observes a stationary AR(1) sequence of evidence scores:

$$Z_{n+1} = \phi Z_n + (1 - \phi)\mu_t + \varepsilon_{n+1}$$

where $\phi \in [0, 1)$ and the innovations are Gaussian with variance $1 - \phi^2$. The observed evidence can be written as the one-sided p -value $P_n = 1 - \Phi(Z_n)$, with $\mu_0 = 0 < \mu_H$.

A larger value of Z_n is more favorable evidence. When $\phi = 0$, attempted specifications are independent. When ϕ is close to one, nearby specifications are persistent and a new check contains little new information. In this benchmark, the effective number of independent opportunities among n attempted specifications is proportional to $(1 - \phi)n$.

This example also clarifies the two editorial instruments. A standard is a rule for how favorable an individual realization must be. For instance, the editor might require Z_n above a high threshold. Disclosure is a rule for how many qualifying realizations must be shown. For instance, the editor might require many disclosed p -values in a preannounced window. The first instrument changes how long the researcher searches. The second changes what the editor learns from the search.

The general model below abstracts from this Gaussian structure. The results are stated for finite type spaces and general dependent evidence processes, but the Gaussian benchmark is the intended mental model throughout. The empirical calibration in Section 6 returns to this structure by estimating a folded-Gaussian type mixture and an AR(1) dependence parameter.

3.2 Types and evidence

A submission has latent type T in a finite set \mathcal{T} , with common prior $\mathbb{P}(T = t) = \pi_t$. Types are partitioned into high-impact and non-high types

$$\mathcal{H} \subseteq \mathcal{T} \quad \mathcal{T}_0 \equiv \mathcal{T} \setminus \mathcal{H} \quad \pi_H \equiv \mathbb{P}(T \in \mathcal{H}) \quad \pi_0 \equiv 1 - \pi_H$$

The researcher and the journal share the prior. The type is not directly observed by the journal. The researcher learns about it through the evidence process. Conditional on $T = t$, the researcher observes a stochastic process $(P_n)_{n \geq 1}$ taking values in $(0, 1)$. In the empirical application, P_n is a p -value or a monotone transformation of a test statistic produced by a specification. The process may be dependent, so repeated tests need not create independent evidence.

Let \mathcal{F}_n denote the sigma-field generated by the first n evidence draws. A history after n tests is denoted by $h_n = (P_1, \dots, P_n)$. The researcher pays cost $\gamma > 0$ per draw and can stop at any finite stopping time τ adapted to (\mathcal{F}_n) .

The sequential notation is not meant to rule out parallel computation. If an automated system runs a batch of b specifications in parallel, one can order the completed specifications arbitrarily and treat them as b draws with total cost $b\gamma$. Equivalently, one can define the primitive draw as a batch vector with cost proportional to its size. The results use the cost of privately generated evidence and the size of the disclosed report, not the literal real-time order in which code is executed.

3.3 Reports and selective disclosure

After stopping, the researcher sends a report R . A report is a finite multiset of realized evidence values drawn from the stopped history. We write $R \preceq h_\tau$ when R is a submultiset of h_τ , and $|R|$ for the number of disclosed values, counting multiplicity.

The journal observes R but not the full history h_τ . Omissions are unverifiable. Hence, after any realized history, the researcher can choose the report that maximizes the probability of acceptance among all subreports of that history.

A journal policy is a measurable acceptance rule

$$\delta : \mathcal{R} \rightarrow [0, 1]$$

where \mathcal{R} is the set of finite reports. Given R , the journal accepts with probability $\delta(R)$. Because omissions are free, any policy is behaviorally equivalent to its disclosure envelope

$$\bar{\delta}(h_n) = \sup_{R \preceq h_n} \delta(R)$$

The envelope is the continuation value from stopping after history h_n . We therefore analyze policies through the stopping problem induced by their envelopes. This reduction is without loss for equilibrium outcomes and is formalized in Appendix A.2.

The policy is chosen by the journal. Thus, in a robustness-check interpretation, the relevant evidence window, specification surface, and qualification rule are editor-defined or referee-audited objects. The researcher chooses which realized elements to disclose, but not the rule that maps reports into acceptance. Referee requests can be incorporated by enlarging the journal-defined surface or by requiring local checks around disclosed specifications. If requests are unpredictable and nonlocal, they are an additional audit technology outside the baseline model. If the plausible request set is predictable and local, it is part of the search universe and the same scaling logic applies. We discuss the differences between these cases in more detail in Section 7.

3.4 Timing and payoffs

The timing is as follows.

- (1) Nature draws T from the common prior
- (2) The journal commits to an acceptance rule δ
- (3) The researcher observes evidence and pays γ per draw
- (4) After each private history, the researcher chooses whether to continue or stop
- (5) After stopping at τ , the researcher chooses a report $R \preceq h_\tau$
- (6) The journal accepts with probability $\delta(R)$

Let $A \in \{0, 1\}$ be the acceptance indicator. The researcher's payoff is

$$u_R(A, \tau) = A - \gamma\tau$$

The researcher chooses a stopping rule and a reporting rule to maximize expected payoff. At the reporting stage, she chooses an envelope-attaining report whenever one exists. At the stopping stage, she solves the optimal stopping problem induced by the envelope. When there are multiple best responses, we use the earliest Snell-envelope optimizer as a selected best response. The lower bounds below apply to any best response unless otherwise stated. The achievability results use the selected best response, equivalently the existence of an equilibrium with the stated operating point.

The editor's primitive preference is to accept high submissions and reject non-high submissions. We do not need to choose a particular cardinal loss function. Any editor whose payoff is increasing in q_H and decreasing in q_0 , possibly subject to a throughput or capacity constraint, chooses a point on the same screening frontier. Characterizing that frontier is therefore more informative than solving a single weighted objective.

The relevant equilibrium concept is commitment equilibrium. Commitment is an institutional assumption: the journal announces how reports will be evaluated before evidence is realized. This is the natural benchmark for editorial policies, robustness requirements, and preannounced specification-surface protocols. The model does not claim that commitment is automatic. Rather, it asks what screening is possible when such rules can be made credible.

3.5 Operating points and the screening frontier

For a policy δ and induced equilibrium behavior, define the two acceptance probabilities

$$\begin{aligned} q_H(\delta) &\equiv \mathbb{P}(A = 1 \mid T \in \mathcal{H}) \\ q_0(\delta) &\equiv \mathbb{P}(A = 1 \mid T \in \mathcal{T}_0) \end{aligned}$$

We interpret q_H as recall and q_0 as the false-positive probability. These determine throughput and the false discovery rate:

$$\begin{aligned} \rho(\delta) &\equiv \mathbb{P}(A = 1) \\ &= \pi_H q_H(\delta) + \pi_0 q_0(\delta) \\ \text{FDR}(\delta) &\equiv \mathbb{P}(T \in \mathcal{T}_0 \mid A = 1) \\ &= \frac{\pi_0 q_0(\delta)}{\rho(\delta)} \end{aligned}$$

The screening frontier at testing cost γ is the set of operating points induced by feasible journal policies and researcher best responses:

$$\{(q_H(\delta), q_0(\delta)) : \delta \in \Delta_\gamma\}$$

A fixed-capacity journal chooses a point on the frontier subject to a throughput requirement. A journal targeting purity chooses a point subject to an FDR requirement. Proposition 4.2 shows how these two slices are linked by the same information budget.

Figure 1 provides a schematic view of the screening frontier and of the standard fixed-capacity interpretation in (q_H, q_0) space.

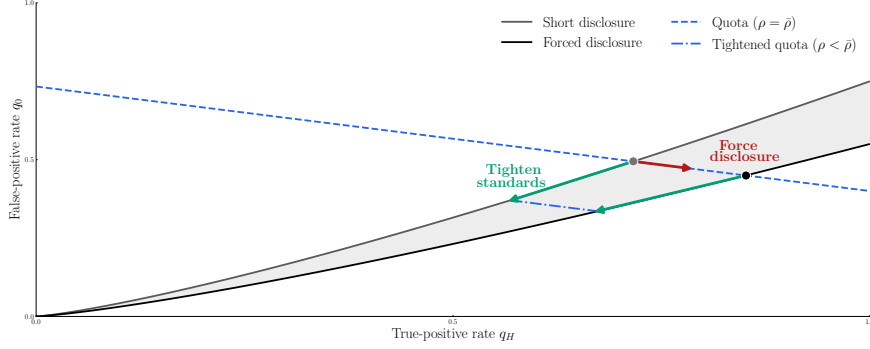


Figure 1: Screening frontier in (q_H, q_0) space (schematic; primitives held fixed). Each curve plots attainable operating points (q_H, q_0) under a disclosure regime as standards vary: tightening standards moves down-left along a frontier. With prior (π_H, π_0) , throughput is $\rho = \pi_H q_H + \pi_0 q_0$, so a fixed-capacity (fixed-throughput) constraint $\rho = \bar{\rho}$ is a straight line (dashed). Forcing disclosure shifts the frontier downward, allowing lower q_0 and hence lower FDR at a given throughput.

3.6 Standards and disclosure

The journal has two distinct instruments. The first is standards. The editor can require more favorable evidence for acceptance. Under a one-test rule, for example, the editor might accept only if the report contains a p -value below a stringent cutoff. Tightening the cutoff makes a non-high researcher search longer before finding a favorable realization.

The second instrument is disclosure. The editor can require the report to contain many favorable pieces of evidence, or to satisfy a structured robustness requirement. Disclosure changes what the editor observes. It makes acceptance depend not only on the best realization found during search, but on the number and pattern of realized checks the researcher is willing to show.

The distinction matters because cheap testing creates a wedge between private effort and public information. Standards can induce effort, but effort remains hidden if reports are short. Disclosure converts the private search history into evidence the editor can use.

3.7 Information and dependence scales

Two scales govern screening in the cheap-testing limit.

The first is an information scale. Under the standing regularity condition in Appendix A.1.1, log-likelihood information accumulates at most linearly in the number of tests. Let $D_{\text{mix}}(\theta)$ be the relevant per-test likelihood-ratio bound and let C_ν be the finite offset in Assumption A.1. Define

$$\kappa_{\text{KL}} \equiv \frac{D_{\text{mix}}(\theta)}{\gamma}$$

This is the total information capacity available per unit testing cost. It governs the universal outer bound: no mechanism can make acceptance more informative than the stopped evidence process from which acceptance is generated.

The second is a dependence scale. Repeated specifications are correlated, so the number of attempted tests is not the number of effectively independent opportunities. Let $n_{\text{eff}}(\theta) \in (0, 1]$ denote the effective-sample-size rate. Define

$$\kappa_{\text{eff}} \equiv \frac{n_{\text{eff}}(\theta)}{\gamma}$$

This is the exponent that appears in the constructive robustness-check rules. It measures how many effectively independent chances are generated at the testing-capacity scale.

In the Gaussian AR(1) benchmark, dependence is summarized by ϕ . When ϕ rises, nearby specifications become more redundant, and the effective-sample-size scale falls. Thus the theory does not say that thousands of highly correlated checks are equivalent to thousands of independent checks. The relevant object is the number of effective checks.

3.8 Assumptions

The main text uses the following economic content of the formal assumptions in Appendix A.1.1. These assumptions largely ensure that the journal’s problem is not trivial in the primitives. The more technical assumptions in the appendix ensure existence for the more general case, but what we need here is that the journal’s screening problem is tractable but that they cannot perfectly differentiate types, as that would remove all of the interesting economic content.

First, evidence is not arbitrarily informative in one draw. Per-test likelihood-ratio information is bounded by $D_{\text{mix}}(\theta)$ up to the finite offset C_ν .

Second, dependence admits an effective-sample-size rate $n_{\text{eff}}(\theta)$. This gives concentration inequalities for counts of evidence in preannounced windows.

Third, likelihood ratios are well-defined across types. This absolute-continuity condition rules out degenerate evidence that perfectly reveals type in one observation.

The short-disclosure result requires two additional restrictions. A no-smoking-guns condition bounds the information content of any small subvector of disclosed evidence. A selection-control condition bounds the extra information conveyed by the stopping and reporting event beyond the likelihood ratio of the disclosed coordinates. These assumptions isolate the economically relevant case in which screening is neither trivial nor impossible. If a single test can reveal type, the editor does not need robustness checks. If even the researcher cannot learn type from long histories, no disclosure rule can generate sharp screening.

The robustness-check achievability result adds a witness-window condition. There must exist a preannounced evidence region B_0 that high types hit more often than non-high types in a likelihood-ratio sense. In the two-type Gaussian location example, sufficiently favorable p -value regions provide such windows. In richer Gaussian type structures, the window may be bounded rather than one-sided, for example when extremely large statistics are more consistent with a distinct non-credible type than with the high type.

4 Frontier bounds

This section develops the limits on editorial screening. The editor ultimately makes a binary decision, accept or reject. Thus every policy induces a binary experiment: the distribution of acceptance under high submissions and the distribution of acceptance under non-high submissions. The natural measure of how informative that binary experiment is equals the Bernoulli Kullback–Leibler divergence between the two acceptance probabilities.

For an acceptance policy δ , let (q_H, q_0) denote the induced operating point and recall that

$$\rho(\delta) = \pi_H q_H(\delta) + \pi_0 q_0(\delta) \quad \text{FDR}(\delta) = \frac{\pi_0 q_0(\delta)}{\rho(\delta)}$$

Define

$$K(\delta) \equiv \text{KL}(\text{Bern}(q_H(\delta)) \parallel \text{Bern}(q_0(\delta)))$$

The organizing idea is simple. Once we upper-bound $K(\delta)$, exponential limits on screening follow. We first translate a bound on $K(\delta)$ into operational statements about throughput and FDR. We then bound $K(\delta)$ itself, first by the researcher’s total information capacity and then by the amount of information that a short report can transmit.

4.1 A Bernoulli–KL budget implies exponential slice laws

We begin with a simple inequality for the Bernoulli KL, which converts bounds on the KL divergence into quantitative rates for throughput and FDR control.

Lemma 4.1. *For any $(q_H, q_0) \in (0, 1] \times (0, 1)$,*

$$\text{KL}(\text{Bern}(q_H) \parallel \text{Bern}(q_0)) \geq q_H \log \frac{q_H}{q_0} - \frac{1}{e} \quad (1)$$

Proof. This is the standard inequality $\text{KL}(q_H \parallel q_0) \geq q_H \log(q_H/q_0) - C(q_H)$ with $C(q_H) \equiv -(1 - q_H) \log(1 - q_H) \leq 1/e$ (Lemma A.31 in Appendix A.4.1). \square

Lemma 4.1 connects the KL budget to the two editorial slices of the frontier: a fixed-throughput slice and a fixed-FDR slice.

Proposition 4.2. *Fix a policy δ and write $K \equiv K(\delta)$.*

(i) *Fix $\bar{\rho} \in (0, \pi_H]$ and suppose $\rho(\delta) \geq \bar{\rho}$. Then there exist constants $c, C > 0$ depending only on $(\bar{\rho}, \pi_H, \pi_0)$ such that*

$$\text{FDR}(\delta) \geq c \exp(-C \cdot K) \quad (2)$$

In particular, at any nonvanishing throughput floor, the fastest possible purification scale is exponential in the Bernoulli–KL budget.

(ii) *Fix $\varepsilon \in (0, 1)$ and define*

$$\eta(\varepsilon) \equiv \frac{\pi_H}{\pi_0} \cdot \frac{\varepsilon}{1 - \varepsilon} \quad L(\varepsilon) \equiv \log \frac{1}{\eta(\varepsilon)} = \log \left(\frac{\pi_0(1 - \varepsilon)}{\pi_H \varepsilon} \right)$$

If $\text{FDR}(\delta) \leq \varepsilon$, then

$$\rho(\delta) \leq \frac{\pi_H}{1 - \varepsilon} \cdot \frac{K + 1/e}{L(\varepsilon)} \quad (3)$$

Equivalently, sustaining nonvanishing throughput while driving $\varepsilon \downarrow 0$ requires K to grow at least on the order of $L(\varepsilon) \asymp \log(1/\varepsilon)$.

Proof. For (ii), $\text{FDR} \leq \varepsilon$ is equivalent to $q_0 \leq \eta(\varepsilon)q_H$, hence $\log(q_H/q_0) \geq L(\varepsilon)$. Applying Lemma 4.1 yields $K \geq q_H L(\varepsilon) - 1/e$, so $q_H \leq (K + 1/e)/L(\varepsilon)$. On the slice $\text{FDR} \leq \varepsilon$ we also have $\pi_H q_H \geq (1 - \varepsilon)\rho$, hence $\rho \leq \pi_H q_H / (1 - \varepsilon)$, giving (3).

For (i), combine Lemma 4.1 with the identities $\rho = \pi_H q_H + \pi_0 q_0$ and $\text{FDR} = \pi_0 q_0 / \rho$. On the event $\text{FDR} \leq 1/2$, the constraint $\rho \geq \bar{\rho}$ forces q_H to be bounded below by a constant depending only on $(\bar{\rho}, \pi_H)$, and Lemma 4.1 then yields an exponential lower bound on q_0 and hence on FDR. On the complementary event $\text{FDR} > 1/2$ we have a constant lower bound. Appendix A.4 records the algebra and constants. \square

Proposition 4.2 gives the rate language used throughout the paper. At any nonvanishing throughput level, the FDR can improve at most exponentially in the Bernoulli–KL budget $K(\delta)$. Conversely, sustaining throughput while targeting a small FDR requires a KL budget of order $\log(1/\varepsilon)$. The remaining question is how large $K(\delta)$ can be in the strategic disclosure game.

4.2 Universal information budget: a mechanism-independent benchmark

The first bound is mechanism-independent. Acceptance is generated from a selectively disclosed report, and the disclosed report is generated from the researcher's stopped evidence history. Therefore acceptance cannot be more informative than the evidence history from which it is produced. The following defines the journal's *desideratum*, an information-theoretic benchmark bound on how much information the editor could possibly extract from the researcher.

Theorem 4.3. *There exists a constant $C_\nu < \infty$ from Assumption A.1 such that for every policy δ with selected best-response stopping time $\tau^*(\delta)$,*

$$K(\delta) = \text{KL}(\text{Bern}(q_H(\delta)) \parallel \text{Bern}(q_0(\delta))) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^*(\delta)] \leq C_\nu + \frac{\kappa_{\text{KL}}}{\pi_H}$$

Combined with Proposition 4.2(i), Theorem 4.3 implies an information-theoretic limit on purification at any throughput floor: if $\rho(\delta) \geq \bar{\rho} > 0$, then necessarily $\text{FDR}(\delta) \geq \exp(-O(\kappa_{\text{KL}}))$, with constants depending on $(\bar{\rho}, \pi_H, \pi_0)$. This is an outer envelope. It describes what would be possible if the editor could extract essentially all of the information the researcher can generate before stopping.

The intuition is data processing. In the Gaussian benchmark, each additional observation contributes only bounded likelihood-ratio information about whether the mean is μ_H or μ_0 . A stopped history with τ observations therefore contains information proportional to τ . The acceptance decision is only a randomized function of a disclosed report, which is itself a function of that stopped history. Thus the binary accept/reject experiment cannot contain more KL information than the stopped evidence process. Since the researcher pays γ per draw, optimal search has expected length at most proportional to $1/\gamma$, giving the information scale κ_{KL} .

This universal benchmark yields the first substantive implication of cheap testing: if a policy does not induce capacity-scale search, then it cannot generate capacity-scale separation.

Theorem 4.4. *Fix θ and consider a sequence $\gamma \downarrow 0$. Let δ^γ be any sequence of policies with selected best-response stopping times τ^γ and induced operating points (q_H^γ, q_0^γ) . Assume $\liminf_{\gamma \downarrow 0} q_H^\gamma \geq c_H > 0$ (nonvanishing recall).*

If the researcher does not run a capacity-scale search under high types,

$$\mathbb{E}_H[\tau^\gamma] = o(1/\gamma) \tag{4}$$

then the induced Bernoulli–KL budget is subcapacity: $K(\delta^\gamma) = o(\kappa_{\text{KL}})$. Consequently, false positives cannot decay at the capacity-scale exponential rate:

$$-\log q_0^\gamma = o(\kappa_{\text{KL}}) \quad q_0^\gamma = \exp(-o(\kappa_{\text{KL}}))$$

In particular, on any throughput-floor slice $\rho(\delta^\gamma) \geq \bar{\rho} > 0$,

$$\text{FDR}(\delta^\gamma) \geq \exp(-o(\kappa_{\text{KL}}))$$

If instead $\mathbb{E}_H[\tau^\gamma] = O(1)$, then there exists $\underline{q}_0 > 0$ such that $q_0^\gamma \geq \underline{q}_0$ for all sufficiently small γ (and hence $\text{FDR}(\delta^\gamma) \geq \underline{\text{FDR}} > 0$).

Proof sketch. By the universal KL budget (Theorem 4.3),

$$K(\delta^\gamma) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^\gamma]$$

By the Bernoulli–KL lower bound (Lemma 4.1),

$$K(\delta^\gamma) \geq q_H^\gamma \log \frac{q_H^\gamma}{q_0^\gamma} - \frac{1}{e}$$

Rearranging and using $q_H^\gamma \geq c_H$ yields

$$\log \frac{q_H^\gamma}{q_0^\gamma} \leq \frac{K(\delta^\gamma) + 1/e}{q_H^\gamma}$$

so $-\log q_0^\gamma \leq O(1) + O(K(\delta^\gamma))$ uniformly for small γ .

Under (4), Theorem 4.3 gives $K(\delta^\gamma) \leq C_\nu + o(\kappa_{\text{KL}}) = o(\kappa_{\text{KL}})$, so $-\log q_0^\gamma = o(\kappa_{\text{KL}})$. If instead $\mathbb{E}_H[\tau^\gamma] = O(1)$, then $K(\delta^\gamma) = O(1)$ and hence q_0^γ is bounded away from 0.

Finally, on any throughput floor $\rho(\delta^\gamma) \geq \bar{\rho} > 0$ we have $\text{FDR} = \pi_0 q_0 / \rho \geq (\pi_0 / \bar{\rho}) q_0$, giving the displayed slice implication. \square

Theorem 4.4 says that exponential screening requires a race. But it does not say that any race is useful to the editor. A policy can induce extensive private search while leaving almost all of that search hidden. The next result shows that this is exactly what goes wrong under short disclosure.

4.3 Short disclosure: the frontier collapses below the universal benchmark

The universal bound is stated in terms of the researcher’s total information capacity. The editor, however, observes only the disclosed report R . To compare standards with disclosure, we therefore need a bound that accounts for the length of what is disclosed.

The Gaussian benchmark gives the intuition. Suppose a non-high researcher can draw many correlated z-statistics and report only the best one. A very large reported z-statistic is not just evidence about the type. It is also evidence that the researcher searched. If the editor observes only one selected realization, the report cannot reveal how many near misses were omitted. Tightening standards can make the researcher search longer, but a short report cannot transmit the resulting information at the effective-sample-size scale.

One subtlety is that the report can be informative not only through the likelihood ratio of the disclosed p -values, but also through the researcher’s stopping and disclosure decisions. For example, if a researcher stops only when the unreported evidence history looks unusually favorable, then the event that the report arrived can shift the editor’s posterior even when the report itself is short. Appendix A.5 formalizes this additional contribution as an extra information from selection term and imposes Assumption A.38, which requires it to be negligible on the κ_{eff} scale along the equilibrium sequence.

Theorem 4.5. *Fix θ with $n_{\text{eff}}(\theta) > 0$ and consider $\gamma \downarrow 0$. Let δ^γ be any sequence of policies and let (q_H^γ, q_0^γ) be the induced operating points under the selected best response.*

Assume short disclosure: under the selected best response,

$$|R| \leq m(\gamma) \quad \text{almost surely} \quad m(\gamma) = o(1/\gamma)$$

Maintain Assumption A.6 for some $\alpha > 1$, and assume the extra information from selection is controlled as in Assumption A.38 (Appendix A.5).

Then, exponential FDR control at the effective-sample-size scale is impossible at nonvanishing recall: for every $c_H > 0$, if $\liminf_{\gamma \downarrow 0} q_H^\gamma \geq c_H$ then

$$-\log q_0^\gamma = o(\kappa_{\text{eff}}) \quad q_0^\gamma = \exp(-o(\kappa_{\text{eff}}))$$

This result is the formal distinction between tightening standards and forcing disclosure. Standards can induce private effort. Disclosure determines how much of that effort becomes public information. If equilibrium reports are sublinear in testing capacity, false positives cannot fall at any fixed exponential rate $\exp(-c \cdot \kappa_{\text{eff}})$ at nonvanishing recall.

The section leaves us with two lessons. The universal KL constraint gives the best possible envelope in terms of the researcher’s total information capacity. The short-disclosure bound shows that short reports cannot attain that envelope. Section 5 constructs robustness-check rules with $m = \Theta(1/\gamma)$ that force disclosure at the natural testing-capacity scale and recover $\exp(-\Omega(\kappa_{\text{eff}}))$ screening. Under comparability, this is optimal up to constants in the exponent.

5 Disclosure at the capacity scale

Section 4 showed that binary screening is governed by a single bottleneck: how much information the editor can extract from what is disclosed. Tightening standards can induce search, but short disclosure cannot transmit capacity-scale discrimination. This section shows that a simple robustness-check rule closes the gap.

The rule has two components. First, the journal fixes an evidence window B . In the two-type Gaussian benchmark, B can be a one-sided significance region such as $p < 0.05$. In richer type structures, B can be a bounded window of z-scores, for example a region that is favorable but not so extreme that it is more consistent with a distinct non-credible type. Second, the journal fixes a required number of disclosed passes m . The key scaling is $m(\gamma) = \Theta(1/\gamma)$.

This rule should be read as an editor-defined or referee-audited requirement. The researcher may choose which realized checks to disclose, but the journal defines the admissible evidence region and the specification surface on which checks are evaluated. Referees can add discipline by defining local neighborhoods of required checks around reported specifications. If such neighborhoods are predictable, they become part of the surface. If they are unpredictable and nonlocal, they are a complementary audit technology. For now, assume that the researcher has complete information about the required checks for their chosen specification and that such checks are pre-defined. We discuss the case where the editor can randomize over a region of required checks after seeing the report in Section 7.

We consider two screening objectives. Under a throughput floor, the journal seeks policies with $\rho(\delta) \geq \underline{\rho}$ that minimize $\text{FDR}(\delta)$. Under an FDR target, it seeks policies with $\text{FDR}(\delta) \leq \varepsilon$ that maximize throughput $\rho(\delta)$. If the journal also faces a hard capacity constraint $\rho(\delta) \leq \bar{\rho}$, it can ration among qualifiers. Because rationing changes researchers’ incentives, we treat capacity clearing formally in Appendix A.3.4 and keep the main-text mechanism analysis focused on screening.

Fix a Borel set $B \subset (0, 1)$, interpreted as the editor’s preannounced evidentiary window for a single diagnostic. Fix an integer $m \geq 1$.

Definition 5.1. Given a report R , let $N_B(R)$ be the number of disclosed significant results with respect to the significance region B :

$$N_B(R) \equiv \sum_{p \in R} \mathbf{1}\{p \in B\}$$

The robustness-check rule $\delta_{B,m}$ accepts if and only if the report contains at least m significant results:

$$\delta_{B,m}(R) \equiv \mathbf{1}\{N_B(R) \geq m\}$$

Assumption A.2 provides a calibrated witness window B_0 on which high-impact types have a uniform likelihood-ratio advantage over every non-high type. Economically, B_0 is a region where seeing a significant result is genuinely diagnostic. In the two-type Gaussian case, B_0 can be any window of the form $(0, c)$, with $c < 1$. In more general type structures, the witness window may be an interval of z-scores centered around the high types (see Lemma A.45).

Define the qualification event $Q \equiv \{N_{B_0}(R) \geq m\}$. Under $\delta_{B_0, m}$, acceptance occurs if and only if Q , so (q_H, q_0) are exactly the qualification probabilities:

$$q_H(\delta_{B_0, m}) = \mathbb{P}(Q \mid T \in \mathcal{H}) \quad q_0(\delta_{B_0, m}) = \mathbb{P}(Q \mid T \in \mathcal{T}_0)$$

Theorem 5.2. *Maintain the standing assumptions (Appendix A.1.1), Assumption A.8, and suppose $n_{\text{eff}}(\theta) > 0$. Let B_0 be the witness window from Assumption A.2. Fix a constant c satisfying*

$$\frac{e^{-\ell_0}}{1 - e^{-\ell_0}} < c < p_H(B_0)$$

where $(\ell_0, p_H(B_0))$ are from Assumption A.2. Set

$$m(\gamma) \equiv \left\lceil \frac{c}{\gamma} \right\rceil$$

Then there exist constants $c_0 > 0$ and $c_H \in (0, 1]$ such that for all sufficiently small γ :

$$q_0(\delta_{B_0, m(\gamma)}) \leq \exp(-c_0 \kappa_{\text{eff}}) \quad q_H(\delta_{B_0, m(\gamma)}) \geq c_H$$

Consequently, $\text{FDR}(\delta_{B_0, m(\gamma)}) = \exp(-\Omega(\kappa_{\text{eff}}))$ and throughput is nonvanishing.

The proof has the same logic as the Gaussian benchmark. To qualify, a non-high submission must generate at least $m(\gamma)$ significant results in the witness window. Since $m(\gamma)$ is proportional to $1/\gamma$, qualification requires sustained favorable evidence at the testing-capacity scale. Under dependence, the relevant exponent is not the raw number of attempted checks, but the effective number of independent checks, $n_{\text{eff}}(\theta)n$. With $n \asymp 1/\gamma$, the probability of the required count deviation is exponentially small on the κ_{eff} scale.

Formally, the proof controls the non-high qualification probability by decomposing it into two terms: a posterior-tail term, which captures histories on which a non-high submission nevertheless induces a high posterior belief, and a count-deviation term, which captures histories on which the non-high evidence stream produces at least m significant results in B_0 . Assumption A.8 controls the posterior-tail term, while the standing concentration bound controls the count-deviation term. Appendix A.3 gives the formal argument and constants.

Theorem 5.2 delivers a concrete screening frontier point: it keeps recall bounded away from zero while driving false positives down at rate $\exp(-\Omega(\kappa_{\text{eff}}))$. The constant in the exponent depends on the design choice c and on the calibrated window B_0 . Larger c asks for more corroborating significant results and improves screening, but eventually violates feasibility under high types.

The universal KL constraint implies that, at any nonvanishing throughput floor, FDR cannot decay faster than $\exp(-O(\kappa_{\text{KL}}))$ (Proposition 4.2 combined with Theorem 4.3). Under comparability, $\kappa_{\text{KL}} = \Theta(\kappa_{\text{eff}})$, the achievability bound above therefore matches the information-theoretic envelope up to constants in the exponent.

Corollary 5.3. *Assume $D_{\text{mix}}(\theta) \asymp n_{\text{eff}}(\theta)$ so that $\kappa_{\text{KL}} = \Theta(\kappa_{\text{eff}})$. Then robustness-check policies with $m(\gamma) = \Theta(1/\gamma)$ attain the optimal purification scale on both slices:*

- (i) on any nonvanishing-throughput slice, $\text{FDR} = \exp(-\Theta(\kappa_{\text{KL}}))$ is achievable and unimprovable up to constants in the exponent;
- (ii) on the fixed-FDR slice, sustaining nonvanishing throughput requires $\log(1/\varepsilon) = O(\kappa_{\text{KL}})$; conversely, there exists $c_\star > 0$ such that any target sequence satisfying $\log(1/\varepsilon) \leq c_\star \kappa_{\text{KL}}$ is attainable by an appropriate choice of $m(\gamma) = \Theta(1/\gamma)$ without collapsing throughput.

The policy implication is not that journals should mechanically count thousands of tables. The mechanism identifies the scale of disclosure needed for screening when omission is unverifiable. In practice, the evidence object may be a specification curve, a preregistered surface, a machine-readable audit file, or a referee-defined diagnostic set. What matters for the theory is that the editor conditions on enough of the surface to make private search visible.

6 Specification surfaces and empirical calibration

This section operationalizes the paper’s central empirical construct: the specification surface. A specification surface is a paper-specific *ex ante* commitment that defines the claim under audit and fixes the admissible universe of estimand-preserving variants, reconstructed from the analytical forks the paper itself reveals. It is the empirical counterpart of the journal-defined evidence surface in the theory.

We use this object in two steps. First, we demonstrate an auditable agentic workflow that reproduces published AEA-journal papers and replications of those papers at a fraction of the human-analyst cost reported in [Brodeur et al. \(2026\)](#) (a mega-analysis featuring replications of top empirical papers in Economics and Political Science by the Institute for Replication, henceforth I4R), and validate the workflow on a paired-replication sample of $n = 39$ papers (Sample A). Second, we use the resulting surface-defined specification sets from a broader sample of $n = 103$ recent AEA papers (Sample B) to estimate the primitives that govern screening when science is cheap: a three-type mixture over hypothesis quality, an effective dependence parameter, and a marginal cost index. These ingredients deliver counterfactual operating points for editorial screening policies.

The cost index should be interpreted narrowly. It measures the marginal execution cost of running standardized observational specifications in an audited workflow. It does not measure the full intellectual cost of designing a credible empirical strategy, understanding institutional context, defending identification, or writing the paper. Those costs may remain large. The model’s private-search margin is the incremental cost of trying additional admissible specifications once the claim, data, and surface have been fixed.

Implementation details, estimation procedures, and robustness checks are deferred to [Appendix B](#). Here, we define the empirical objects and summarize the main validation and estimation results.

6.1 Specification surfaces and audited execution

A central question in any replication exercise is what, exactly, is being replicated. In experimental contexts, large-scale efforts distinguish exact replications from conceptual replications ([Crandall and Sherman, 2016](#)): the latter vary implementation details while preserving the outcome concept, treatment concept, and intended estimand. For observational studies, the relevant object is closer to disciplined reanalysis. We do not bring in new data to test the same hypothesis in a new environment. Instead, we treat the paper’s replication package as a fixed computational object and ask whether its central claim is stable across defensible choices that preserve the underlying claim object. This is consistent with the Institute for Replication’s large-scale reanalysis protocol ([Brodeur et al., 2026](#)), which targets a standardized baseline and a structured menu of robustness checks.

We center that exercise on a single object: a per-paper specification surface fixed *ex ante*. The surface defines the baseline claim objects, records the paper’s canonical baseline specifications, and commits to the universe of estimand-preserving variants, together with explicit constraints, budgets, and a canonical inference choice for each specification. It is deliberately conservative: it reconstructs only the forks a reader can verify from the paper itself, namely the analytical dimensions along which the paper’s own reported specifications reveal that defensible alternatives existed. Any protocol-added stress tests beyond these revealed forks are explicit and separable.

Once the surface is fixed, execution is mechanical: only specifications admitted by the surface may be run, and all outputs record the full coefficient vector in a standardized format. A separate verification step then audits those outputs without running new regressions and produces the conservative verified core used in all downstream estimation. Appendix B.1 provides the full implementation details.

Figure 2 illustrates the resulting object for Drobner (2022), which is useful because it has a single verified baseline group, a common outcome and treatment variable, and 78 verified-core specifications. The upper panel sorts coefficient estimates with 95% confidence intervals. The lower panel records the matching analysis choices for each specification: the Table 2 cell, control set, sample rule, fixed effects, and verified specification family.

6.2 Validation against a large scale replication effort

Our primary validation benchmark is the Institute for Replication’s reanalysis protocol in Brodeur et al. (2026). For each paper i in their AEA-journal sample, we target the same central hypothesis and run the automated workflow in Section 6.1 on the paper’s public replication package.¹ This yields an automated baseline reproduction with focal-coefficient t -statistic t_i^{auto} , directly comparable to the I4R benchmark t_i^{i4r} .

Because the theory’s evidence index is the absolute t -statistic, we validate using $|t|$. For hypothesis i , let t_i^{orig} denote the published t -statistic, t_i^{i4r} the I4R reanalysis, and t_i^{auto} our automated reproduction. We assess distributional agreement of $|t^{\text{auto}}|$ with $|t^{\text{i4r}}|$ (Figure 3) and claim-by-claim agreement statistics (Appendix B.2).

We study 39 of the 41 AEA-journal papers from Brodeur et al. (2026).² For each paper we ingest the public replication package, target the I4R central hypothesis, and produce a paired hypothesis-level object: the benchmark t -statistic t_i^{i4r} and our reproduced baseline t -statistic t_i^{auto} .

Figure 3 reports the central distributional comparison. We plot the distribution of the evidence index $|t|$ for original published hypotheses, I4R reanalyses, and automated baseline reproductions, restricting to the verified-comparable subset of Sample A. We also plot a matched reproduction distribution, constructed by selecting within each paper the surface-approved specification that is closest to the I4R reanalysis target.

Two features matter. First, the automated distribution closely tracks the I4R benchmark distribution, supporting the claim that the workflow reproduces canonical reanalysis objects. Second, the gap between original and replicated distributions provides a compact summary of the selection and instability forces motivating the model: the original distribution reflects the equilibrium output of editorial and researcher incentives, while the replicated distributions approximate the underlying evidence conditional on a standardized analysis.

¹We used Anthropic’s Claude Opus 4.5 and Opus 4.6 through the Claude Code CLI. Any frontier model with agentic coding capabilities and access to the user’s computer would suffice, though naturally results would vary.

²We exclude Guerron-Quintana et al. (2023), a structural macroeconomics paper whose hypothesis is incompatible with the specification-search framework, and Cohen and Dechezleprêtre (2022), who used weather data from the Mexican national meteorological database that is available only on request.

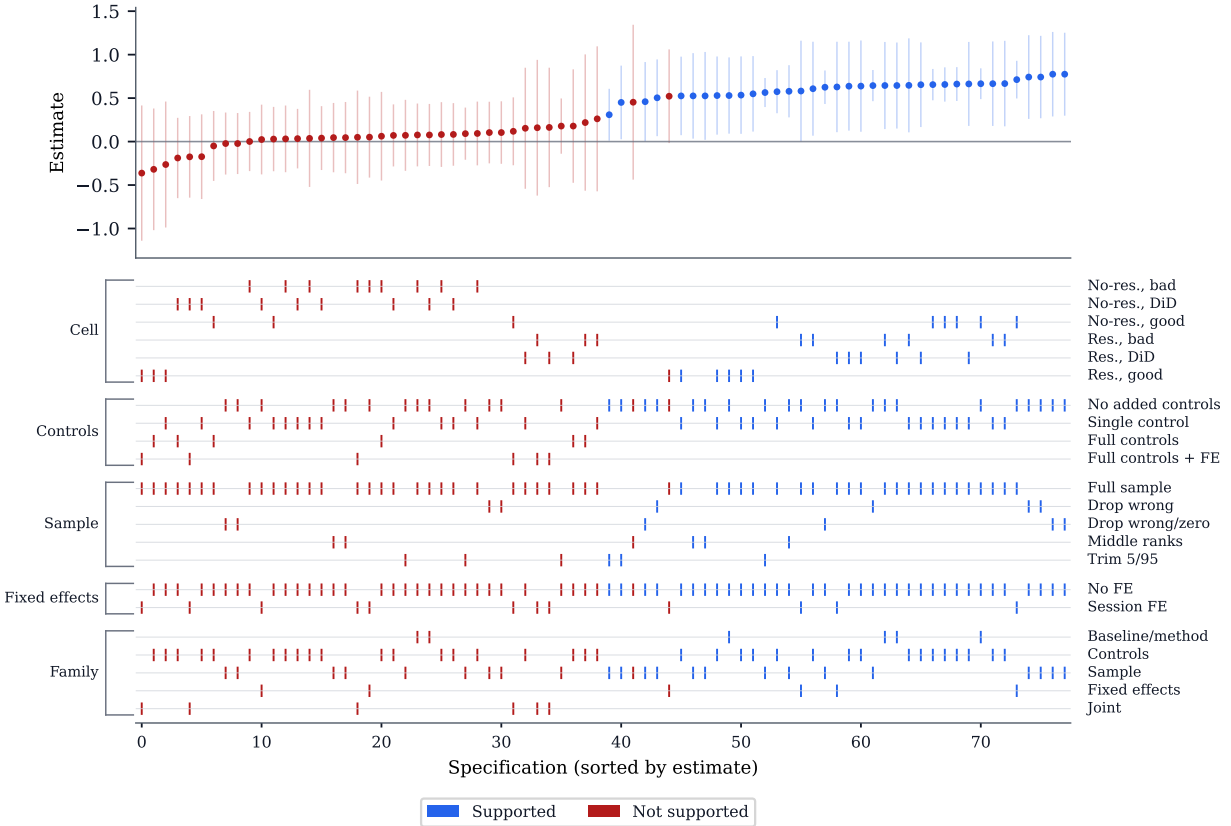


Figure 2: Specification curve for [Drobner \(2022\)](#). Top: each point is one verified-core specification, showing the coefficient on Bayesian belief adjustment in a regression of belief adjustment on Bayesian belief adjustment, with 95% confidence intervals. Specifications are sorted by point estimate; blue marks a positive coefficient significant at the 5% level, and red marks specifications that are not significant in that direction. Bottom: strike plot of the analytic decisions underlying each specification. Each column corresponds to the same specification shown directly above.

We emphasize that the I4R reanalysis is not ground truth. It reflects standardized execution choices that may differ from the original paper. We interpret it as a disciplined reference point under a transparent protocol. Importantly, the same automated system that executes our broader surface-defined specification universes also reproduces the I4R target specification when constrained to do so. This is the sense in which [Figure 3](#) validates that the automated workflow is operating on the intended empirical object rather than producing arbitrary variation.

6.3 Estimating screening primitives from surface-defined specification sets

Having validated the automated workflow against I4R at the hypothesis level, we use the resulting surface-defined specification sets to estimate the evidence and dependence primitives that govern screening when testing is cheap. We then use these estimates to evaluate counterfactual disclosure requirements under a cost shift.

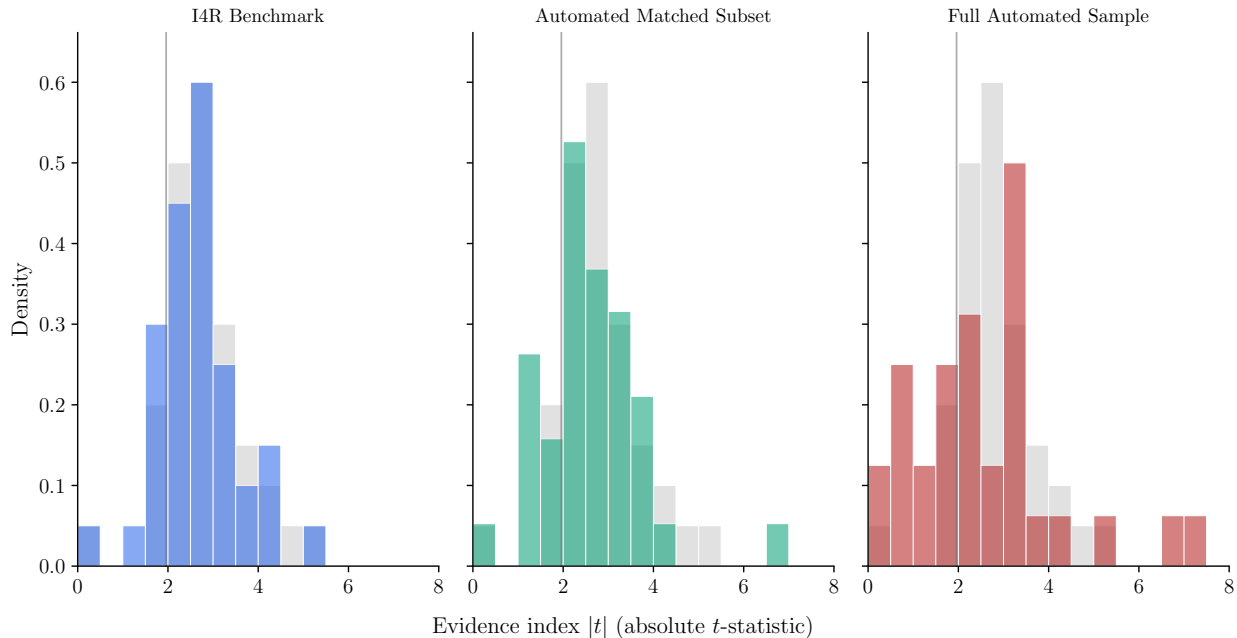


Figure 3: Distributions of the evidence index $|t|$ for hypotheses (Sample A, verified-comparable subset). Gray histogram: original studies. Blue, left panel: I4R reanalyses. Red, center panel: automated reproductions (verified). Green, right panel: matched reproductions (within-paper specification matched to I4R reanalysis target).

6.3.1 Replication artifacts and sample construction

For each paper, the pipeline produces (i) a baseline reproduction, (ii) a standardized specification-level table with one row per robustness run, and (iii) a verification map labeling the estimand-preserving core. These machine-readable artifacts let us fit the model to a large set of comparable specification traversals.

We run automated specification search on a random sample of all papers for which we can ingest the public replication package and construct and execute a standardized specification surface (103 papers total, including the I4R sample). For each paper i we obtain a set of surface-approved specifications $s = 1, \dots, S_i$ with a corresponding evidence index $|t_{is}|$ and verification labels that identify the estimand-preserving core. Sample B provides power for estimating the three-type mixture and dependence primitives.

6.3.2 A three-type model of empirical evidence

To map the data to the model primitives, we adopt a three-type statistical representation aligned with the theory. Let $|t_{is}|$ denote the absolute t -statistic under harmonized inference for specification (i, s) , so that $|t| = 0$ is a null and $|t| \approx 1.96$ corresponds to $p = 0.05$ under a two-sided normal approximation.

Types correspond to null relationships (N), moderate relationships that are stable across defensible analyses but close to conventional thresholds (M), and extreme relationships that generate very large evidence indices (E). The goal is not to claim that any given paper is literally generated by a three-component mixture. The goal is to pin down a prior over types and a disciplined mapping from observed evidence indices to screening primitives.

We model $|t|$ using a three-component mixture of folded Gaussians:

$$|t_{is}| \sim \sum_{k \in \{N, M, E\}} \pi_k \cdot |X_k| \quad X_k \sim \mathcal{N}(\mu_k, 1)$$

for $k \in \{N, M, E\}$. The folded-normal family is natural here: if the latent test statistic is $X \sim \mathcal{N}(\mu_k, 1)$ and we observe only the absolute value, the induced density on $|X|$ is exactly the folded normal. Fixing $\sigma_k = 1$ for all components gives the model a unit-variance interpretation consistent with a well-calibrated t -statistic, and leaves $(\pi_k, \mu_k)_{k \in \{N, M, E\}}$ as the free parameters. Estimation is by maximum likelihood with 50 random initializations on the verified-core baseline specifications restricted to $|t| \leq 10$; components are labeled by sorting fitted means. Appendix B.3 gives details. The fitted components provide two objects used in the counterfactuals: an estimated prior $(\hat{\pi}_N, \hat{\pi}_M, \hat{\pi}_E)$ and an implied diagnostic region on the index scale.

Figure 4 overlays the fitted mixture densities on the Sample B histogram of baseline specification $|t|$ values. The three-component structure captures the multimodality of the evidence distribution: a mass near zero, a concentration around conventional significance thresholds, and a diffuse upper tail. Appendix B.3 reports parameter estimates, goodness-of-fit diagnostics, model selection across $K \in \{2, 3, 4\}$ via AIC and BIC, robustness to alternative distributional families, and trimming checks.

The null component ($\hat{\mu}_N \approx 1.6$, comprising roughly 62 percent of specifications) has a latent mean below conventional significance, so that most realizations of $|t|$ from this type fall short of the 1.96 threshold. The moderate component ($\hat{\mu}_M \approx 3.9$, roughly 31 percent of specifications) concentrates above conventional significance thresholds; these are papers with real effects that reliably pass the $p = 0.05$ boundary but remain sensitive to the number of required passes. The extreme component ($\hat{\mu}_E \approx 7.9$, roughly 8 percent of specifications) captures relationships whose evidence indices are so large that they are essentially unchallenged by any reasonable reanalysis. Neither label carries a normative valence: a moderate paper may study an important effect that happens to be hard to detect, while an extreme paper may exploit a mechanical relationship. What matters for the counterfactual is the screening distinction: moderate-type papers are the ones whose qualification status responds to the number of required passes, whereas null and extreme types are largely inframarginal.

6.3.3 Dependence and effective sample size

The theory emphasizes that when testing is cheap, the screening-relevant quantity is not the number of attempted specifications but the number of effectively independent tests. We estimate an effective dependence primitive using an AR(1) model along the specification traversal from Sample B. Within each baseline group g with $n_g \geq 3$ specifications, we order the specifications according to a chosen ordering and regress $|t_{g,s+1}|$ on $|t_{g,s}|$ to obtain a group-level persistence coefficient $\hat{\phi}_g$. The pooled estimate is a weighted average across groups, with weights proportional to n_g .

Because the AR(1) estimate depends on how specifications are ordered, we estimate $\hat{\phi}$ under six orderings (document order, lexicographic path, breadth-first, depth-first, by verification category, and a random null) and select the ordering with the highest pooled R^2 excluding the random null as the preferred estimate. The preferred ordering, by verification category, yields $\hat{\phi} = 0.151$. We define the effective-independence parameter $\hat{\Delta} \equiv 1 - \hat{\phi} = 0.849$. This maps directly to the Gaussian AR(1) example in Section 3, where $n_{\text{eff}}(\phi) = 1 - \phi$. Appendix B.4 reports all orderings with bootstrap confidence intervals and R^2 values, and all orderings enter the counterfactual sensitivity analysis.

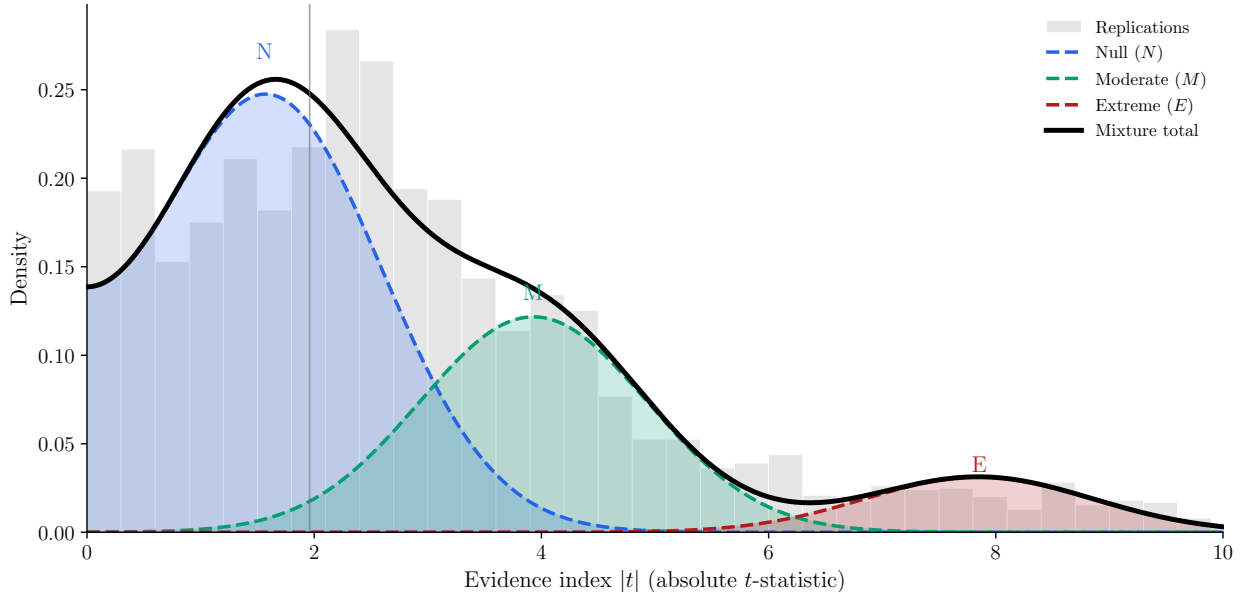


Figure 4: Three-type evidence model: fitted folded-Gaussian mixture on the evidence index $|t|$ (Sample B, verified-core baseline specifications, $|t| \leq 10$). Component densities and total mixture density are overlaid on the histogram.

6.3.4 Cost ratios and disclosure scale

Our counterfactuals vary only the marginal per-specification testing cost γ , holding fixed the evidence environment and dependence estimated from the surface-defined specification sets. The level of γ is not directly observed: omitted exploration is unreported, and replication time mixes large fixed costs with the incremental cost of running additional specifications. We therefore discipline the cost shift with two observable inputs that map to the model.

The first is the cost ratio

$$\lambda \equiv \frac{\gamma^{\text{new}}}{\gamma^{\text{old}}}$$

calibrated using a scope-aligned timing comparison between (i) the wall-clock time of our automated workflow to execute a standardized baseline replication object, including validation, and (ii) a conservative I4R benchmark. Under the I4R protocol, the average replication time was 13 days. To stay conservative, we benchmark against the fastest completed I4R reanalysis reported in [Brodeur et al. \(2026\)](#), which was five working days (40 hours). In our automated workflow, the mean wall-clock time to complete the analogous baseline reproduction object is 14 minutes across 44 successfully completed papers. This yields our baseline $\lambda = 1/172$.³

This comparison is not a claim that automation reduces every component of scientific work by a factor of 172. It is a calibration of the marginal execution cost for standardized specifications after the empirical object is defined. Design, interpretation, identification, and persuasion remain costly. This is why using [Brodeur et al. \(2026\)](#) as a benchmark is useful: the replication teams met several times before the ‘clock’ was started, chose papers in their fields of expertise, and had time prior to the replication to explore and understand the codebase. Our automated workflow, by contrast,

³In the most conservative alternative specification, we assume replicators spent two of those five days purely understanding the codebase and compare 14 minutes to 24 working hours, yielding $\lambda = 1/103$. Additional specifications and assumptions are reported in [Appendix B.5](#); the qualitative results do not change.

performed the entire process of the replication on the clock. We are, if anything, underestimating the marginal cost shift of running a set of empirical replications.

The second input is the pre-shift disclosure scale m^{old} , measured from the number of specifications reported by authors in the papers we study. The median of that set is 50, which we take as the baseline pre-automation disclosure requirement (Appendix B reports alternatives). This is the empirical counterpart of the disclosure requirement in the mechanism: it captures what editors can condition on under unverifiable omission, even if additional private exploration occurred off the record.

6.3.5 Counterfactual screening under a cost shift

We connect the estimated primitives to the model’s counterfactuals, holding the evidence environment fixed and varying only the marginal testing cost γ .

Let γ^{old} denote the pre-automation per-test cost and $\gamma^{\text{new}} = \lambda \gamma^{\text{old}}$ the post-automation cost. Because equilibrium horizons scale like $1/\gamma$, a reduction in γ increases feasible private search and makes short-disclosure rules more vulnerable, exactly as in Sections 4–5.

We evaluate robustness-check rules as studied in Section 5. We fix the evidence window at $B = [1.96, \infty)$ on the $|t|$ scale, so that a specification passes if its absolute t -statistic exceeds the conventional significance threshold.⁴ For a required number of passes m , the type- k qualification probability is $Q_k(m) = \Pr(\text{Bin}(N_{\text{eff}}, p_k(B)) \geq m)$, where $N_{\text{eff}} \approx \hat{\Delta}n$ and $p_k(B) = F_k(z_h) - F_k(z_\ell)$. Under a fixed-capacity interpretation with throughput target $\bar{\rho}$, the editor accepts qualifying papers with probability $a = \bar{\rho}/\bar{Q}$ when feasible. We define the false discovery rate with respect to the null type only:

$$\text{FDR}(m) = \frac{\pi_N Q_N(m)}{\bar{Q}(m)}$$

treating extreme-type papers as true positives for the purpose of screening.

Figure 5 reports the headline comparison. We calibrate the old regime so that $m^{\text{old}} = 50$ qualifying robustness checks achieve a false discovery rate of exactly 0.05, then ask what disclosure is required after the cost shift. Panel A shows the disclosure multiplier $m^{\text{new}}/m^{\text{old}}$ for each baseline requirement: at $m^{\text{old}} = 50$ (highlighted), the new regime requires $m^{\text{new}} = 6,994$, roughly a 140-fold increase. The multiplier grows with the baseline, approaching $1/\lambda \approx 172$ for large m^{old} . Panel B displays the implied FDR as a heatmap over m and the cost ratio λ , zoomed around the baseline calibration point ($m = 6,994$, $\lambda \approx 1/172$, marked by a star). The steep gradient shows that the FDR is sensitive to small changes in the required number of passes near the calibration point. Appendix B.5 reports the full disclosure-scaling schedule for different parameterizations of m^{old} and a sensitivity grid over λ , evidence windows, and mixture model variants.

7 Discussion

Cheap testing changes the meaning of a reported result. In our commitment model, researchers search over a correlated specification space at per-test cost γ , stop strategically, and selectively disclose results. The editor observes only the disclosed report and applies an accept/reject policy. We characterize the attainable FDR–throughput frontier, prove a mechanism-independent information bound governed by a likelihood-ratio budget, and show a sharp asymmetry between tightening

⁴Results are essentially unchanged when we impose a finite upper bound, such as $B = [1.96, 10]$ or $B = [1.96, 15]$, because the extreme component’s pass probability is already near unity. Alternative specifications of B are reported in Appendix B.5.

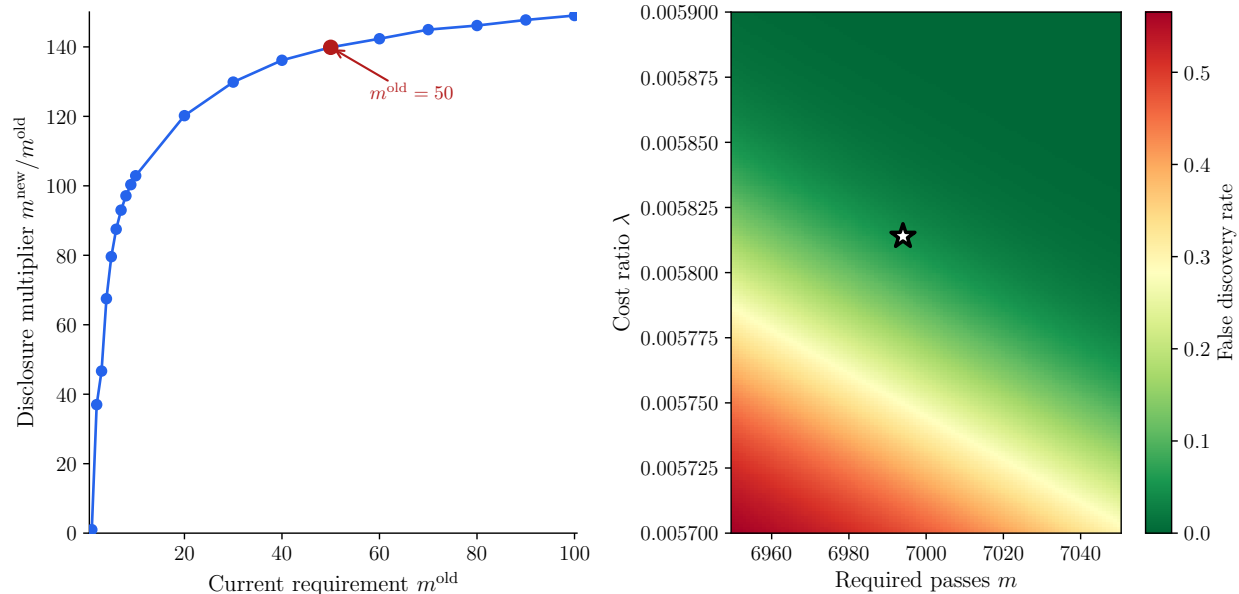


Figure 5: Counterfactual screening under a cost shift ($\lambda \approx 1/172$), calibrated so that $m^{\text{old}} = 50$ achieves FDR = 0.05 in the old regime. Panel A: disclosure multiplier $m^{\text{new}}/m^{\text{old}}$ for each baseline requirement; the highlighted bar at $m^{\text{old}} = 50$ shows a 140-fold increase. Panel B: FDR heatmap over m and cost ratio λ , zoomed around the baseline calibration point; star marks ($m^{\text{new}} = 6,994, \lambda \approx 1/172$).

standards and forcing disclosure. Tightening standards makes researchers search harder. Requiring disclosure makes the search informative to the editor. In the cheap-testing limit, short equilibrium reports cannot transmit capacity-scale discrimination, while robustness-check rules that force disclosure to scale with testing capacity recover exponential screening at a dependence-adjusted effective-sample-size rate.

The empirical calibration makes the scale of the problem concrete. Under conservative assumptions, we estimate a large cost shift from automation ($\lambda \approx 1/172$) and a substantial effective-independence rate in the resulting surface-defined specification sets ($\hat{\Delta} \approx 0.849$). When we ask what it would take to hold fixed a conventional screening target after this cost shift, the answer is on the order of 7,000 qualifying robustness checks (Figure 5). Taken literally, this is a daunting number. No editor can read that many checks, no author can narrate them, and no referee process can adjudicate them one by one. Moreover, for many papers there may not be 7,000 reasonable estimand-preserving specifications on a given dataset.

The point is therefore not that journals should demand thousands of robustness tables. The point is that once private testing capacity becomes large, the unit of editorial interpretation has to change. A main table plus a small number of selected robustness checks is a low-dimensional summary of a high-dimensional search problem. When the marginal cost of search is high, that summary may be a reasonable compromise. When the marginal cost of search is low, the same summary becomes strategically fragile. The relevant object is no longer the best-looking specification together with a few nearby variants. It is the behavior of the claim over an audited specification surface.

Our empirical pipeline implements this shift by separating commitment from execution. For each baseline claim object, we construct and audit an *ex ante* specification surface that commits to a universe of estimand-preserving variants together with explicit constraints, budgets, and a canonical

inference choice. The surface reconstructs the minimal garden of forking paths (Gelman and Loken, 2013) that the paper itself reveals: the analytical dimensions along which the paper’s own reported specifications demonstrate that defensible alternatives existed. Because it captures only revealed forks, it is a conservative lower bound on the researcher’s actual search space and thereby defines the space she most plausibly optimized over. Any protocol-added stress tests beyond these revealed forks are explicit and separable. Execution is then determined by the surface, with independent verification yielding a conservative verified core used in estimation. This operationalization makes the garden-of-forking-paths and multiverse perspectives (Steege et al., 2016; Simonsohn et al., 2020; Kasy, 2021) executable: rather than cataloguing all imaginable analytical degrees of freedom, it reconstructs the minimal garden the paper reveals and asks whether the reported result is a property of the claim or an artifact of the particular path taken.

7.1 Author-selected and referee-requested robustness checks

A useful way to interpret the results is to distinguish three objects that are often grouped together under the label “robustness checks.” The first is an author-selected robustness appendix. The author decides which checks to run, which checks to report, and how to frame them. This is useful as exposition, but it is also selected evidence. If omissions are unverifiable, an author-selected robustness appendix is subject to the same strategic force as the main result: it reports some outcomes from a larger private search process.

The second object is a journal-defined disclosure rule. This is the object in our mechanism. The journal fixes an evidence window, a specification surface, and a qualification rule before evaluating the report. The author may still search and may still choose what to disclose from the realized evidence, but the author does not choose the acceptance rule. In this sense, our robustness-check mechanism should not be read as a defense of unconstrained author-selected robustness checks. It is a defense of robustness checks as an editorial instrument, where the relevant universe is fixed or audited by the journal.

The third object is a referee-requested robustness check. Referee requests can be powerful because they may introduce information that was not part of the author’s private search problem. If a referee can request a genuinely nonlocal and unpredictable check after seeing the submission, then a small number of such checks may be highly informative. In that case, the referee is not merely asking the author to disclose another draw from the author’s searched region; the referee is creating an audit outside the region the author could cheaply optimize over.

The force of this objection depends on unpredictability. Many referee requests are local to the submitted empirical strategy: use a nearby sample restriction, add or remove a natural control, change the clustering rule, use a closely related outcome, or vary a conventional functional form. These checks are valuable, but they are also often predictable. A researcher with cheap testing can anticipate a neighborhood of plausible referee requests and search for a specification whose local neighborhood also looks favorable. In the Gaussian AR(1) benchmark, this amounts to searching not for a single favorable point but for a favorable local block. Holding the local radius fixed, the same scaling logic remains: if the disclosed or audited evidence grows sublinearly in testing capacity, screening cannot recover the dependence-adjusted exponential rate.

This distinction clarifies how referees enter the policy message. Referees matter most by helping define and audit the specification surface. They decide which perturbations are admissible, which perturbations change the estimand, which inference conventions are appropriate, and which parts of the reported surface are credible. A referee request is most informative when it either expands the surface in a way the author could not predict or verifies that the reported surface was not chosen after seeing outcomes. In contrast, a small number of predictable local requests is not a substitute

for disclosure at the scale of private search capacity.

The practical implication is not that editors should remove discretion from referees. It is the opposite. Referee expertise is essential, but its role changes. Under cheap testing, expertise is less about asking for one more table in the neighborhood of the author’s preferred specification and more about defining the neighborhood itself. The editorial object should be a claim, an admissible surface around that claim, and a rule for interpreting the distribution of results over that surface.

7.2 Commitment and the partial-equilibrium interpretation

The model is deliberately a partial-equilibrium theory of editorial screening. We take a journal, a submission pool, an evidence process, and a commitment technology as given. The journal commits *ex ante* to an acceptance rule. Researchers understand the rule, search accordingly, and choose what to disclose. The results characterize the operating points attainable by that committed screening rule.

This commitment assumption is doing real work. Without commitment, the editor’s incentives after observing a report may differ from the incentives used to design the rule. A journal may be tempted to accept a striking short report even when its own announced policy required broader disclosure. Referees may change their requested checks after seeing the direction of the result. Authors may not believe that costly disclosure rules will actually be enforced. If the rule is not credible, then the mechanism is not implemented. Our results therefore should not be read as a prediction that journals will automatically choose the screening frontier, nor as a claim that the publication market as a whole will converge to the socially optimal rule.

The partial-equilibrium interpretation is instead conditional. If a journal can commit to a rule, and if authors believe that rule will be applied, then the paper characterizes how the rule must scale as testing becomes cheap. Under that interpretation, the main theorem is a design constraint. It says that a committed journal cannot restore screening merely by raising statistical thresholds while continuing to evaluate short selected reports. To extract the information generated by cheap search, the journal must condition on evidence whose disclosed or audited dimension grows with the feasible search capacity.

This is also why the paper is not a general equilibrium account of scientific publication. We do not model entry into research questions, journal competition, prestige incentives, referee labor markets, author sorting across journals, or the endogenous choice of which hypotheses are written up. Those forces matter. A stricter journal may lose submissions to a less strict journal. A field may respond to high disclosure requirements by changing what kinds of papers are attempted. Authors may select hypotheses before the paper-level game begins. These are important margins, but they are outside the screening frontier studied here.

The partial-equilibrium benchmark remains useful because it isolates the informational constraint faced by any institution that wants to screen selected evidence. Journal competition can change which policies are adopted. It does not change the fact that a short selected report contains limited information about a high-dimensional private search. Heterogeneous testing costs can change the appropriate weighting of the specification surface. They do not change the distinction between hidden effort and disclosed evidence. Referee expertise can improve the surface and introduce unpredictable audits. It does not remove the need to ask whether the author could have optimized over the checks being requested.

The results should therefore be read as a conditional frontier, not a universal prescription. They answer the question: given a claim, a surface of admissible evidence, a submission pool, and a journal that can commit, what screening is possible when authors can cheaply search and selectively disclose? That question is narrower than a complete theory of publishing, but it is the question an

editor faces when deciding how to interpret robustness evidence for a submitted paper.

7.3 Preregistration, third-party audits, and the object of interpretation

It may seem natural to respond to cheap testing by requiring researchers to register pre-analysis plans (Olken, 2015; Banerjee et al., 2020; Kasy and Spiess, 2024). Pre-analysis plans can be powerful when the researcher cannot privately inspect the relevant data before committing. But many observational studies use public data sources, and researchers can search before submitting a plan. Even if all researchers act in good faith, preregistration under cheap testing can shift selective reporting from the specification level to the project level, because significant preregistered studies are more likely to be written up and submitted (Rosenthal, 1979; Gelman and Loken, 2013). The central issue is not only whether a plan was written before a reported analysis. It is whether the published evidence object reveals the relevant search universe.

A different response is to make papers shorter and rely on subsequent researchers to perform robustness checks. This is valuable for cumulative science. Third-party audits can improve the public record, reveal errors, and discipline future work. But they solve a different problem unless they are connected to the acceptance or certification rule. Our model concerns *ex ante* screening at the point of publication. If robustness checks occur only after publication, they may correct beliefs later, but they do not determine which claims receive the initial journal imprimatur. A journal that wants to screen at the point of acceptance must either condition acceptance on the audited surface or create a credible post-acceptance certification process that affects the status of the published claim.

Historically, the attraction of point estimates, single t -statistics, and a small handful of robustness tables was as much computational as statistical. In a world where each additional specification imposed real human cost, it was infeasible to enumerate the multiverse, so researchers relied on low-dimensional summaries of a much larger set of choices. Cheap computation reverses that logic. When authors can search widely at low marginal cost, a single reported effect is precisely the object most distorted by endogenous exploration, stopping, and omission. Editors should therefore ask for evidence objects whose dimension grows with the feasible search space but whose interpretation remains stable.

This creates a difficult interpretive problem. Specification-curve and multiverse analysis formalize the idea that what matters is the mapping from defensible choices to estimates, together with reasoned arguments about how particular choices move the result (Gelman and Loken, 2013; Steegen et al., 2016; Young and Holsteen, 2017; Simonsohn et al., 2020). Our results give these practices a mechanism-design rationale. When omission is unverifiable, pointwise testing is strategically fragile. Curve-level disclosure is not only descriptive transparency. It is an incentive instrument.

The next methodological challenge is to make these evidence objects interpretable. Editors need rules for evaluating distributions of estimates, dependence across specifications, clustering of specification choices, and the distinction between meaningful robustness failures and changes in estimand. Our theory does not solve that interpretive problem. It shows why the problem becomes unavoidable once science becomes cheap.

References

- Mitra Akhtari, Diana Moreira, and Laura Trucco. Political turnover, bureaucratic turnover, and the quality of public services. *American Economic Review*, 112(2):442–493, 2022. doi: 10.1257/aer.20171867. URL <https://doi.org/10.1257/aer.20171867>.
- Hunt Allcott. The welfare effects of misperceived product costs: Data and calibrations from the automobile market. *American Economic Journal: Economic Policy*, 5(3):30–66, 2013. doi: 10.1257/pol.5.3.30. URL <https://doi.org/10.1257/pol.5.3.30>.
- Hunt Allcott, Matthew Gentzkow, and Lena Song. Digital addiction. *American Economic Review*, 112(7):2424–2463, 2022. doi: 10.1257/aer.20210867. URL <https://doi.org/10.1257/aer.20210867>.
- Onur Altindag, Bilge Erten, and Pinar Keskin. Mental health costs of lockdowns: Evidence from age-specific curfews in turkey. *American Economic Journal: Applied Economics*, 14(2):320–343, 2022. doi: 10.1257/app.20200811. URL <https://doi.org/10.1257/app.20200811>.
- Sandro Ambuehl, B. Douglas Bernheim, and Annamaria Lusardi. Evaluating deliberative competence: A simple method with an application to financial choice. *American Economic Review*, 112(11):3584–3626, 2022. doi: 10.1257/aer.20210290. URL <https://doi.org/10.1257/aer.20210290>.
- Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794, August 2019. doi: 10.1257/aer.20180310. URL <https://doi.org/10.1257/aer.20180310>.
- Isaiah Andrews and Jesse M. Shapiro. A model of scientific communication. *Econometrica*, 89(5):2117–2142, September 2021. doi: 10.3982/ECTA18155. URL <https://doi.org/10.3982/ECTA18155>.
- Vellore Arthi, Brian Beach, and W. Walker Hanlon. Recessions, mortality, and migration bias: Evidence from the lancashire cotton famine. *American Economic Journal: Applied Economics*, 14(2):228–255, 2022. doi: 10.1257/app.20190131. URL <https://doi.org/10.1257/app.20190131>.
- Alicia Atwood. The long-term effects of measles vaccination on earnings and employment. *American Economic Journal: Economic Policy*, 14(2):34–60, 2022. doi: 10.1257/pol.20190509. URL <https://doi.org/10.1257/pol.20190509>.
- Pierre Azoulay, Benjamin F. Jones, J. Daniel Kim, and Javier Miranda. Immigration and entrepreneurship in the united states. *American Economic Review: Insights*, 4(1):71–88, 2022. doi: 10.1257/aeri.20200588. URL <https://doi.org/10.1257/aeri.20200588>.
- Marina Azzimonti, Eva de Francisco, and Vincenzo Quadrini. Financial globalization, inequality, and the rising public debt. *American Economic Review*, 104(8):2267–2302, 2014. doi: 10.1257/aer.104.8.2267. URL <https://doi.org/10.1257/aer.104.8.2267>.
- Linda Babcock, Maria P. Recalde, and Lise Vesterlund. Gender differences in the allocation of low-promotability tasks: The role of backlash. *American Economic Review*, 107(5):131–135, 2017. doi: 10.1257/aer.p20171018. URL <https://doi.org/10.1257/aer.p20171018>.
- Katherine Baicker, Amy Finkelstein, Jae Song, and Sarah Taubman. The impact of medicaid on labor market activity and program participation: Evidence from the oregon health insurance experiment. *American Economic Review*, 104(5):322–328, 2014. doi: 10.1257/aer.104.5.322. URL <https://doi.org/10.1257/aer.104.5.322>.

- Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. Machine learning methods for demand estimation. *American Economic Review*, 105(5):481–485, 2015. doi: 10.1257/aer.p20151021. URL <https://doi.org/10.1257/aer.p20151021>.
- Michael Baker, Yosh Halberstam, Kory Kroft, Alexandre Mas, and Derek Messacar. Pay transparency and the gender gap. *American Economic Journal: Applied Economics*, 15(2):157–183, 2023. doi: 10.1257/app.20210141. URL <https://doi.org/10.1257/app.20210141>.
- Pablo Balán, Augustin Bergeron, Gabriel Tourek, and Jonathan L. Weigel. Local elites as state capacity: How city chiefs use local information to increase tax compliance in the democratic republic of the congo. *American Economic Review*, 112(3):762–797, 2022. doi: 10.1257/aer.20201159. URL <https://doi.org/10.1257/aer.20201159>.
- Abhijit Banerjee, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics. NBER Working Paper 26993, National Bureau of Economic Research, April 2020. URL <https://doi.org/10.3386/w26993>.
- E. Jason Baron. School spending and student outcomes: Evidence from revenue limit elections in wisconsin. *American Economic Journal: Economic Policy*, 14(1):1–39, 2022. doi: 10.1257/pol.20200226. URL <https://doi.org/10.1257/pol.20200226>.
- Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998. URL <https://doi.org/10.1214/aos/1013699998>.
- Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077. URL <https://doi.org/10.1214/12-AOS1077>.
- Martin Bertran, Riccardo Fogliato, and Zhiwei Steven Wu. Many ai analysts, one dataset: Navigating the agentic data science multiverse, 2026. URL <https://arxiv.org/abs/2602.18710>. arXiv:2602.18710v2, revised 11 March 2026.
- Marianne Bitler and Hilary Hoynes. Living arrangements, doubling up, and the great recession: Was this time different? *American Economic Review*, 105(5):166–170, 2015. doi: 10.1257/aer.p20151087. URL <https://doi.org/10.1257/aer.p20151087>.
- Gustavo J. Bobonis, Paul J. Gertler, Marco Gonzalez-Navarro, and Simeon Nichter. Vulnerability and clientelism. *American Economic Review*, 112(11):3627–3659, 2022. doi: 10.1257/aer.20190565. URL <https://doi.org/10.1257/aer.20190565>.
- Iris Bohnet, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states. *American Economic Review*, 98(1):294–310, 2008. doi: 10.1257/aer.98.1.294. URL <https://doi.org/10.1257/aer.98.1.294>.

- Tessa Bold, Selene Ghisolfi, Frances Nsonzi, and Jakob Svensson. Market access and quality upgrading: Evidence from four field experiments. *American Economic Review*, 112(8):2518–2552, 2022. doi: 10.1257/aer.20210122. URL <https://doi.org/10.1257/aer.20210122>.
- Stephen Boyd and Lieven Vandenbergh. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004. doi: 10.1017/CBO9780511804441. URL <https://doi.org/10.1017/CBO9780511804441>.
- J. Carter Braxton and Bledi Taska. Technological change and the consequences of job loss. *American Economic Review*, 113(2):279–316, 2023. doi: 10.1257/aer.20210182. URL <https://doi.org/10.1257/aer.20210182>.
- Abel Brodeur, Derek Mikola, Nikolai Cook, Lenka Fiala, Thomas Brailey, Ryan Briggs, Alexandra de Gendre, Yannick Dupraz, Jacopo Gabani, Romain Gauriot, Joanne Haddad, Goncalo Lima, Jörg Ankel-Peters, Anna Dreber, Douglas Campbell, Lamis Kattan, Diego Marino Fages, Fabian Mierisch, Pu Sun, Taylor Wright, Marie Connolly, Fernando Hoces de la Guardia, Magnus Johannesson, Edward Miguel, Lars Vilhuber, Alejandro Abarca, Mahesh Acharya, Sossou Simplicie Adjisse, Ahwaz Akhtar, Eduardo Alberto Ramirez Lizardi, Sabina Albrecht, Synøve Nygaard Andersen, Zubaria Andlib, Falak Arrora, Thomas Ash, Etienne Bacher, Sebastian Bachler, Félix Bacon, Manuel Bagues, Timea Balogh, Alisher Batmanov, Mara Barschkett, Barış Kaan Basdil, Jaromír Baxa, Sascha O. Becker, Monica Beeder, Louis-Philippe Beland, Abdel-Hamid Bello, Daniel Benenson Markovits, Grant Benjamin, Thomas Bergeron, Moussa P. Blimpo, Marco Binetti, Carl Bonander, Joseph Bonneau, Endre Borbáth, Nicolai Borgen, Solveig Topstad Borgen, Jonathan Borowsky, Elisa Brini, Myriam Brown, Martin Brun, Stephan Bruns, Nino Buliskeria, Andrea Calef, Alistair Cameron, Pamela Campa, Santiago Campos-Rodríguez, Giulio Giacomo Cantone, Fenella Carpena, Perry Jess Carter, Paul Castañeda Dower, Ondrej Cestek, Jill Caviglia-Harris, Gabriella Chauca Strand, Shi Chen, Sya In Chzhen, Jong Chung, Jason Collins, Alexander Coppock, Hugo Cordeau, Ben Couillard, Jonathan Crechet, Lorenzo Crippa, Jing Cui, Christian Czymara, Haley Daarstad, Danh Chi Dao, Daniel Dao, Marco David Schmandt, Astrid de Linde, Lucas De Melo, Lachlan Deer, Micole De Vera, Velichka Dimitrova, Jan Fabian Dollbaum, Jan Matti Dollbaum, Michael Donnelly, Luu Duc Toan Huynh, Tsvetomira Dumbalska, Jamie Duncan, Kiet Tuan Duong, Thibaut Duprey, Christoph Dworschak, Sigmund Ellingsrud, Ali Elminejad, Yasmine Eissa, Andrea Erhart, Giulian Etingin-Frati, Elaheh Fatemipour, Alexa Federice, Jan Feld, Guidon Fenig, Mojtaba Firouzjaeiangalougah, Erlend Fleisje, Alexandre Fortier-Chouinard, Julia Francesca Engel, Nadjim Fréchet, Reid Fortier, Tilman Fries, Michael James Frith, Thomas Galipeau, Sebastian Gallegos, Areez Gangji, Xiaoying Gao, Cloé Garnache, Attila Gáspár, Evelina Gavrilova, Arijit Ghosh, Garreth Gibney, Grant Gibson, Geir Godager, Leonard Goff, Da Gong, Javier González, Jeremy D. Gretton, Cristina Griffa, Idaliya Grigoryeva, Maja Grøtting, Eric Guntermann, Jiaqi Guo, Alexi Gugushvili, Hooman Habibnia, Sonja Häffner, Jonathan D. Hall, Olle Hammar, Amund Hanson Kordt, Barry Hashimoto, Jonathan S. Hartley, Carina I. Hausladen, Tomáš Havránek, Harry He, Matthew Hepplewhite, Mario Herrera-Rodriguez, Felix Heuer, Anthony Heyes, Anson T. Y. Ho, Jonathan Holmes, Armando Holzknecht, Yu-Hsiang Dexter Hsu, Shiang-Hung Hu, Yu-Shiuan Huang, Mathias Huebener, Christoph Huber, Kim P. Huynh, Zuzana Irsova, Ozan Isler, Niklas Jakobsson, Raphaël Jananji, Tharaka A. Jayalath, Michael Jetter, Jenny John, Rachel Joy Forshaw, Felipe Juan, Valon Kadriu, Sunny Karim, Edmund Kelly, Duy Khanh Hoang Dang, Tazia Khushboo, Jin Kim, Gustav Kjellsson, Anders Kjelsrud, Andreas Kotsadam, Jori Korpershoek, Lewis Krashinsky, Suranjana Kundu, Alexander Kustov, Nurlan Lalayev, Audrée Langlois, Jill Laufer, Blake Lee-Whiting, Andreas Leibing, Gabriel Lenz, Joel Levin, Peng Li, Tongzhe Li, Yuchen Lin, Ariel Listo, Dan Liu,

- Xuewen Lu, Elvina Lukmanova, Alex Luscombe, Lester R. Lusher, Ke Lyu, Hai Ma, Nicolas Mäder, Clifton Makate, Alice Malmberg, Adit Maitra, Marco Mandas, Jan Marcus, Shushanik Margaryan, Lili Márk, Andres Martignano, Abigail Marsh, Isabella Masetto, Anthony McCanny, Emma McManus, Ryan McWay, Lennard Metson, Jonas Minet Kinge, Sumit Mishra, Myra Mohnen, Jakob Moeller, Rosalie Montambeault, Sébastien Montpetit, Louis-Philippe Morin, Todd Morris, Scott Moser, Fabio Yoshio Suguri Motoki, Lucija Muehlenbachs, Andreea Musulan, Marco Musumeci, Munirul Nabin, Karim Nchare, Florian Neubauer, Quan M. P. Nguyen, Tuan Nguyen, Viet Nguyen-Tien, Ali Niazi, Giorgi Nikolaishvili, Ardyn Nordstrom, Patrick Nüß, Angela Odermatt, Matt Olson, Henning Øien, Tim Ölkens, Miquel Oliver i Vert, Emre Oral, Christian Oswald, Ali Ousman, Ömer Özak, Shubham Pandey, Alexandre Pavlov, Martino Pelli, Romeo Penheiro, RyuGyung Park, Eva Pérez Martel, Tereza Petrovičová, Linh Phan, Alexa Prettyman, Jakub Procházka, Aqila Putri, Julian Quandt, Kangyu Qiu, Loan Quynh Thi Nguyen, Andaleeb Rahman, Carson H. Rea, Adam Reiremo, Laëtitia Renée, Joseph Richardson, Nicholas Rivers, Bruno Rodrigues, William Roelofs, Tobias Roemer, Ole Rogeberg, Julian Rose, Andrew Roskos-Ewoldsen, Paul Rosmer, Barbara Sabada, Soodeh Saberian, Nicolas Salamanca, Georg Sator, Daniel Scates, Elmar Schlüter, Cameron Sells, Sharmi Sen, Ritika Sethi, Anna Shcherbiak, Moyosore Sogaolu, Matt Soosalu, Erik Ø. Sørensen, Manali Sovani, Noah Spencer, Stefan Staubli, Renske Stans, Anya Stewart, Felix Stips, Kieran Stockley, Stephenson Strobel, Ethan Struby, John P. Tang, Idil Tanrisever, Thomas Tao Yang, Ipek Tastan, Dejan Tatić, Benjamin Tatlow, Féraud Tchuissu Seuyong, Rémi Thériault, Vincent Thivierge, Wenjie Tian, Filip-Mihai Toma, Maddalena Totarelli, Van-Anh Tran, Hung Truong, Nikita Tsoy, Kerem Tuzcuoglu, Diego Ubfal, Laura Villalobos, Julian Walterskirchen, Joseph Tao yi Wang, Vasudha Wattal, Matthew D. Webb, Bryan S. Weber, Reinhard Weisser, Wei-Chien Weng, Christian Westheide, Kimberly White, Jacob Winter, Timo Wochner, Matt Woerman, Jared Wong, Ritchie Woodard, Marcin Wroński, Myra Yazbeck, Gustav Chung Yang, Luther Yap, Kareman Yassin, Hao Ye, Jin Young Yoon, Chris Yurris, Tahreen Zahra, Mirela Zaneva, Aline Zayat, Jonathan Zhang, Ziwei Zhao, and Yaolang Zhong. Reproducibility and robustness of economics and political science research. *Nature*, 652(8108):151–156, April 2026. doi: 10.1038/s41586-026-10251-x. URL <https://doi.org/10.1038/s41586-026-10251-x>.
- Raffaello Bronzini and Eleonora Iachini. Are incentives for r&d effective? evidence from a regression discontinuity approach. *American Economic Journal: Economic Policy*, 6(4):100–134, 2014. doi: 10.1257/pol.6.4.100. URL <https://doi.org/10.1257/pol.6.4.100>.
- Luigi Butera, Robert Metcalfe, William Morrison, and Dmitry Taubinsky. Measuring the welfare effects of shame and pride. *American Economic Review*, 112(1):122–168, 2022. doi: 10.1257/aer.20190433. URL <https://doi.org/10.1257/aer.20190433>.
- Adriana Camacho and Emily Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65, 2011. doi: 10.1257/pol.3.2.41. URL <https://doi.org/10.1257/pol.3.2.41>.
- Yiming Cao and Shuo Chen. Rebel on the canal: Disrupted trade access and social conflict in china, 1650-1911. *American Economic Review*, 112(5):1555–1590, 2022. doi: 10.1257/aer.20201283. URL <https://doi.org/10.1257/aer.20201283>.
- Christoph Carnehl and Johannes Schneider. A quest for knowledge. *Econometrica*, 93(2):623–659, March 2025. doi: 10.3982/ECTA22144. URL <https://doi.org/10.3982/ECTA22144>.

- Matias D Cattaneo, Sebastian Galiani, Paul J Gertler, Sebastian Martinez, and Rocio Titiunik. Housing, health, and happiness. *American Economic Journal: Economic Policy*, 1(1):75–105, 2009. doi: 10.1257/pol.1.1.75. URL <https://doi.org/10.1257/pol.1.1.75>.
- Roberto Chang. Elections, capital flows, and politico-economic equilibria. *American Economic Review*, 100(4):1759–1777, 2010. doi: 10.1257/aer.100.4.1759. URL <https://doi.org/10.1257/aer.100.4.1759>.
- Kerwin Kofi Charles, Erik Hurst, and Matthew J. Notowidigdo. Housing booms and busts, labor market opportunities, and college attendance. *American Economic Review*, 108(10):2947–2994, 2018. doi: 10.1257/aer.20151604. URL <https://doi.org/10.1257/aer.20151604>.
- Valery Charnavoki and Juan J. Dolado. The effects of global shocks on small commodity-exporting economies: Lessons from canada. *American Economic Journal: Macroeconomics*, 6(2):207–237, 2014. doi: 10.1257/mac.6.2.207. URL <https://doi.org/10.1257/mac.6.2.207>.
- François Cohen and Antoine Dechezleprêtre. Mortality, temperature, and public health provision: Evidence from mexico. *American Economic Journal: Economic Policy*, 14(2):161–192, 2022. doi: 10.1257/pol.20180594. URL <https://doi.org/10.1257/pol.20180594>.
- Luke N. Condra, James D. Long, Andrew C. Shaver, and Austin L. Wright. The logic of insurgent electoral violence. *American Economic Review*, 108(11):3199–3231, 2018. doi: 10.1257/aer.20170416. URL <https://doi.org/10.1257/aer.20170416>.
- Lucia Corno, Eliana La Ferrara, and Justine Burns. Interaction, stereotypes, and performance: Evidence from south africa. *American Economic Review*, 112(12):3848–3875, 2022. doi: 10.1257/aer.20181805. URL <https://doi.org/10.1257/aer.20181805>.
- Miguel A Costa-Gomes and Vincent P Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768, 2006. doi: 10.1257/aer.96.5.1737. URL <https://doi.org/10.1257/aer.96.5.1737>.
- Christian S. Crandall and Jeffrey W. Sherman. On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66:93–99, 2016. doi: 10.1016/j.jesp.2015.10.002. URL <https://doi.org/10.1016/j.jesp.2015.10.002>.
- Leemore S Dafny. How do hospitals respond to price changes? *American Economic Review*, 95(5):1525–1547, 2005. doi: 10.1257/000282805775014236. URL <https://doi.org/10.1257/000282805775014236>.
- Guy David, Sara Markowitz, and Seth Richards-Shubik. The effects of pharmaceutical marketing and promotion on adverse drug events and regulation. *American Economic Journal: Economic Policy*, 2(4):1–25, 2010. doi: 10.1257/pol.2.4.1. URL <https://doi.org/10.1257/pol.2.4.1>.
- Lucas W. Davis, Alan Fuchs, and Paul Gertler. Cash for coolers: Evaluating a large-scale appliance replacement program in mexico. *American Economic Journal: Economic Policy*, 6(4):207–238, 2014. doi: 10.1257/pol.6.4.207. URL <https://doi.org/10.1257/pol.6.4.207>.
- Stefano DellaVigna, Ruben Enikolopov, Vera Mironova, Maria Petrova, and Ekaterina Zhuravskaya. Cross-border media and nationalism: Evidence from serbian radio in croatia. *American Economic Journal: Applied Economics*, 6(3):103–132, 2014. doi: 10.1257/app.6.3.103. URL <https://doi.org/10.1257/app.6.3.103>.

- David J. Deming. Using school choice lotteries to test measures of school effectiveness. *American Economic Review*, 104(5):406–411, 2014. doi: 10.1257/aer.104.5.406. URL <https://doi.org/10.1257/aer.104.5.406>.
- Diva Dhar, Tarun Jain, and Seema Jayachandran. Reshaping adolescents’ gender attitudes: Evidence from a school-based experiment in india. *American Economic Review*, 112(3):899–927, 2022. doi: 10.1257/aer.20201112. URL <https://doi.org/10.1257/aer.20201112>.
- Alfredo Di Tillio, Marco Ottaviani, and Peter Norman Sørensen. Persuasion bias in science: Can economics help? *The Economic Journal*, 127(605):F266–F304, 2017. doi: 10.1111/eoj.12515. URL <https://doi.org/10.1111/eoj.12515>.
- Taryn Dinkelman. The effects of rural electrification on employment: New evidence from south africa. *American Economic Review*, 101(7):3078–3108, 2011. doi: 10.1257/aer.101.7.3078. URL <https://doi.org/10.1257/aer.101.7.3078>.
- Thomas Douenne and Adrien Fabre. Yellow vests, pessimistic beliefs, and carbon tax aversion. *American Economic Journal: Economic Policy*, 14(1):81–110, 2022. doi: 10.1257/pol.20200092. URL <https://doi.org/10.1257/pol.20200092>.
- Christoph Drobner. Motivated beliefs and anticipation of uncertainty resolution. *American Economic Review: Insights*, 4(1):89–105, 2022. doi: 10.1257/aeri.20200829. URL <https://doi.org/10.1257/aeri.20200829>.
- Pierre Dubois, Rachel Griffith, and Aviv Nevo. Do prices and attributes explain international differences in food purchases? *American Economic Review*, 104(3):832–867, 2014. doi: 10.1257/aer.104.3.832. URL <https://doi.org/10.1257/aer.104.3.832>.
- Mark Duggan and Fiona Scott Morton. The effect of medicare part d on pharmaceutical prices and utilization. *American Economic Review*, 100(1):590–607, 2010. doi: 10.1257/aer.100.1.590. URL <https://doi.org/10.1257/aer.100.1.590>.
- Ruben Durante, Paolo Pinotti, and Andrea Tesei. The political legacy of entertainment tv. *American Economic Review*, 109(7):2497–2530, 2019. doi: 10.1257/aer.20150958. URL <https://doi.org/10.1257/aer.20150958>.
- Maya Eden. Excessive financing costs in a representative agent framework. *American Economic Journal: Macroeconomics*, 8(2):215–237, 2016. doi: 10.1257/mac.20140147. URL <https://doi.org/10.1257/mac.20140147>.
- Liran Einav, Amy Finkelstein, Stephen P Ryan, Paul Schrimpf, and Mark R Cullen. Selection on moral hazard in health insurance. *American Economic Review*, 103(1):178–219, 2013. doi: 10.1257/aer.103.1.178. URL <https://doi.org/10.1257/aer.103.1.178>.
- Graham Elliott, Nikolay Kudrin, and Kaspar Wüthrich. Detecting p -hacking. *Econometrica*, 90(2): 887–906, March 2022. doi: 10.3982/ECTA18583. URL <https://doi.org/10.3982/ECTA18583>.
- Elena Esposito. The side effects of immunity: Malaria and african slavery in the united states. *American Economic Journal: Applied Economics*, 14(3):290–328, 2022. doi: 10.1257/app.20190372. URL <https://doi.org/10.1257/app.20190372>.

- Gabrielle Fack and Julien Grenet. Improving college access and success for low-income students: Evidence from a large need-based grant program. *American Economic Journal: Applied Economics*, 7(2):1–34, 2015. doi: 10.1257/app.20130423. URL <https://doi.org/10.1257/app.20130423>.
- Michael Faye and Paul Niehaus. Political aid cycles. *American Economic Review*, 102(7):3516–3530, 2012. doi: 10.1257/aer.102.7.3516. URL <https://doi.org/10.1257/aer.102.7.3516>.
- Claudio Ferraz and Frederico Finan. Electoral accountability and corruption: Evidence from the audits of local governments. *American Economic Review*, 101(4):1274–1311, 2011. doi: 10.1257/aer.101.4.1274. URL <https://doi.org/10.1257/aer.101.4.1274>.
- Ana Figueiredo. Wage cyclicality and labor market sorting. *American Economic Review: Insights*, 4(4):425–442, 2022. doi: 10.1257/aeri.20210161. URL <https://doi.org/10.1257/aeri.20210161>.
- William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. arXiv preprint, 2014. URL <https://arxiv.org/abs/1410.2597>.
- Christina M Fong and Erzo F. P Luttmer. What determines giving to hurricane katrina victims? experimental evidence on racial group loyalty. *American Economic Journal: Applied Economics*, 1(2):64–87, 2009. doi: 10.1257/app.1.2.64. URL <https://doi.org/10.1257/app.1.2.64>.
- Meredith Fowlie, Stephen P Holland, and Erin T Mansur. What do emissions markets deliver and to whom? evidence from southern california’s nox trading program. *American Economic Review*, 102(2):965–993, 2012. doi: 10.1257/aer.102.2.965. URL <https://doi.org/10.1257/aer.102.2.965>.
- Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, September 2014. doi: 10.1126/science.1255484. URL <https://doi.org/10.1126/science.1255484>.
- Alexander Frankel and Maximilian Kasy. Which findings should be published? *American Economic Journal: Microeconomics*, 14(1):1–38, February 2022. doi: 10.1257/mic.20190133. URL <https://doi.org/10.1257/mic.20190133>.
- Jeffrey L Furman and Scott Stern. Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review*, 101(5):1933–1963, 2011. doi: 10.1257/aer.101.5.1933. URL <https://doi.org/10.1257/aer.101.5.1933>.
- Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Unpublished manuscript, Department of Statistics, Columbia University (dated 14 November 2013), November 2013. URL https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf.
- Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 2014. doi: 10.1511/2014.111.460. URL <https://doi.org/10.1511/2014.111.460>.
- Marc Goñi. Assortative matching at the top of the distribution: Evidence from the world’s most exclusive marriage market. *American Economic Journal: Applied Economics*, 14(3):445–487, 2022. doi: 10.1257/app.20180463. URL <https://doi.org/10.1257/app.20180463>.
- Christoph Görtz, John D. Tsoukalas, and Francesco Zanetti. News shocks under financial frictions. *American Economic Journal: Macroeconomics*, 14(4):210–243, 2022. doi: 10.1257/mac.20170066. URL <https://doi.org/10.1257/mac.20170066>.

- Gautam Gowrisankaran, Aviv Nevo, and Robert Town. Mergers when prices are negotiated: Evidence from the hospital industry. *American Economic Review*, 105(1):172–203, 2015. doi: 10.1257/aer.20130223. URL <https://doi.org/10.1257/aer.20130223>.
- Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, 1990. doi: 10.1007/978-1-4757-3982-4. URL <http://ee.stanford.edu/~gray/it.pdf>. Corrected first edition (June 26, 2023).
- Michael Greenstone, Guojun He, Ruixue Jia, and Tong Liu. Can technology solve the principal-agent problem? evidence from china’s war on air pollution. *American Economic Review: Insights*, 4(1): 54–70, 2022. doi: 10.1257/aeri.20200373. URL <https://doi.org/10.1257/aeri.20200373>.
- Irena Grosfeld, Alexander Rodnyansky, and Ekaterina Zhuravskaya. Persistent antimarket culture: A legacy of the pale of settlement after the holocaust. *American Economic Journal: Economic Policy*, 5(3):189–226, 2013. doi: 10.1257/pol.5.3.189. URL <https://doi.org/10.1257/pol.5.3.189>.
- Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461–483, 1981. doi: 10.1086/467025. URL <https://doi.org/10.1086/467025>.
- Pablo A. Guerron-Quintana, Tomohiro Hirano, and Ryo Jinnai. Bubbles, crashes, and economic growth: Theory and evidence. *American Economic Journal: Macroeconomics*, 15(2):333–371, April 2023. doi: 10.1257/mac.20220015. URL <https://doi.org/10.1257/mac.20220015>.
- Peter R. Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi: 10.1198/073500105000000063. URL <https://doi.org/10.1198/073500105000000063>.
- Emeric Henry. Strategic disclosure of research results: The cost of proving your honesty. *The Economic Journal*, 119(539):1036–1064, 2009. doi: 10.1111/j.1468-0297.2009.02278.x. URL <https://doi.org/10.1111/j.1468-0297.2009.02278.x>.
- Emeric Henry and Marco Ottaviani. Research and the approval process: The organization of persuasion. *American Economic Review*, 109(3):911–955, 2019. doi: 10.1257/aer.20171919. URL <https://doi.org/10.1257/aer.20171919>.
- Emeric Henry, Ekaterina Zhuravskaya, and Sergei Guriev. Checking and sharing alt-facts. *American Economic Journal: Economic Policy*, 14(3):55–86, 2022. doi: 10.1257/pol.20210037. URL <https://doi.org/10.1257/pol.20210037>.
- Claudia Herresthal. Hidden testing and selective disclosure of evidence. *Journal of Economic Theory*, 200:105312, 2022. doi: 10.1016/j.jet.2021.105312. URL <https://doi.org/10.1016/j.jet.2021.105312>.
- Sandile Hlatshwayo and Michael Spence. Demand and defective growth patterns: The role of the tradable and non-tradable sectors in an open economy. *American Economic Review*, 104(5): 272–277, 2014. doi: 10.1257/aer.104.5.272. URL <https://doi.org/10.1257/aer.104.5.272>.
- Richard Hornbeck and Suresh Naidu. When the levee breaks: Black migration and economic development in the american south. *American Economic Review*, 104(3):963–990, 2014. doi: 10.1257/aer.104.3.963. URL <https://doi.org/10.1257/aer.104.3.963>.

- Caroline M. Hoxby. The economics of online postsecondary education: Moocs, nonselective education, and highly selective education. *American Economic Review*, 104(5):528–533, 2014. doi: 10.1257/aer.104.5.528. URL <https://doi.org/10.1257/aer.104.5.528>.
- Jason Huh and Julian Reif. Teenage driving, mortality, and risky behaviors. *American Economic Review: Insights*, 3(4):523–539, 2021. doi: 10.1257/aeri.20200653. URL <https://doi.org/10.1257/aeri.20200653>.
- Reshmaan Hussam, Natalia Rigol, and Benjamin N. Roth. Targeting high ability entrepreneurs using community information: Mechanism design in the field. *American Economic Review*, 112(3):861–898, 2022. doi: 10.1257/aer.20200751. URL <https://doi.org/10.1257/aer.20200751>.
- John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8):e124, 2005. doi: 10.1371/journal.pmed.0020124. URL <https://doi.org/10.1371/journal.pmed.0020124>.
- C. Kirabo Jackson and Elias Bruegmann. Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, 1(4):85–108, 2009. doi: 10.1257/app.1.4.85. URL <https://doi.org/10.1257/app.1.4.85>.
- Brian A Jacob and Lars Lefgren. The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, 1(3):33–58, 2009. doi: 10.1257/app.1.3.33. URL <https://doi.org/10.1257/app.1.3.33>.
- Ravi Jagadeesan and Davide Viviano. Publication design with incentives in mind. arXiv preprint 2504.21156, arXiv, 2025. URL <https://arxiv.org/abs/2504.21156>. arXiv:2504.21156v2, revised 24 September 2025.
- Damon Jones and Ioana Marinescu. The labor market impacts of universal and permanent cash transfers: Evidence from the alaska permanent fund. *American Economic Journal: Economic Policy*, 14(2):315–340, 2022. doi: 10.1257/pol.20190299. URL <https://doi.org/10.1257/pol.20190299>.
- Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011. doi: 10.1257/aer.101.6.2590. URL <https://doi.org/10.1257/aer.101.6.2590>.
- Maximilian Kasy. Of forking paths and tied hands: Selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives*, 35(3):175–192, 2021. doi: 10.1257/jep.35.3.175. URL <https://doi.org/10.1257/jep.35.3.175>.
- Maximilian Kasy and Jann Spiess. Optimal pre-analysis plans: Statistical decisions subject to implementability. arXiv preprint 2208.09638, arXiv, 2024. URL <https://arxiv.org/abs/2208.09638>. arXiv:2208.09638v3, revised 29 July 2024.
- Benjamin Kohler, David Zollikofer, Johanna Einsiedler, Alexander Hoyle, and Elliott Ash. Read the paper, write the code: Agentic reproduction of social-science results, 2026. URL <https://arxiv.org/abs/2604.21965>.
- Michael Kosfeld and Devesh Rustagi. Leader punishment and cooperation in groups: Experimental field evidence from commons management in ethiopia. *American Economic Review*, 105(2):747–783, 2015. doi: 10.1257/aer.20120700. URL <https://doi.org/10.1257/aer.20120700>.

- Burhanettin Kuruscu. Training and lifetime income. *American Economic Review*, 96(3):832–846, 2006. doi: 10.1257/aer.96.3.832. URL <https://doi.org/10.1257/aer.96.3.832>.
- Sébastien Laffitte and Farid Toubal. Multinationals’ sales and profit shifting in tax havens. *American Economic Journal: Economic Policy*, 14(4):371–396, 2022. doi: 10.1257/pol.20200203. URL <https://doi.org/10.1257/pol.20200203>.
- Morgane Laouénan and Roland Rathelot. Can information reduce ethnic discrimination? evidence from airbnb. *American Economic Journal: Applied Economics*, 14(1):107–132, 2022. doi: 10.1257/app.20190188. URL <https://doi.org/10.1257/app.20190188>.
- Nicholas Lawson. Liquidity constraints, fiscal externalities and optimal tuition subsidies. *American Economic Journal: Economic Policy*, 9(4):313–343, 2017. doi: 10.1257/pol.20150079. URL <https://doi.org/10.1257/pol.20150079>.
- Edward E. Leamer. Let’s take the con out of econometrics. *American Economic Review*, 73(1): 31–43, 1983. URL <https://www.jstor.org/stable/1803924>.
- Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference with the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371. URL <https://doi.org/10.1214/15-AOS1371>.
- Pascal Lézaud. Chernoff-type bound for finite Markov chains. *The Annals of Applied Probability*, 8(3):849–867, 1998. doi: 10.1214/aoap/1028903453. URL <https://doi.org/10.1214/aoap/1028903453>.
- Lee M. Lockwood. Incidental bequests and the choice to self-insure late-life risks. *American Economic Review*, 108(9):2513–2550, 2018. doi: 10.1257/aer.20141651. URL <https://doi.org/10.1257/aer.20141651>.
- Carolina Lopez, Anja Sautmann, and Simone Schaner. Does patient demand contribute to the overuse of prescription drugs? *American Economic Journal: Applied Economics*, 14(1):225–260, 2022. doi: 10.1257/app.20190722. URL <https://doi.org/10.1257/app.20190722>.
- Rodolfo E. Manuelli and Ananth Seshadri. Human capital and the wealth of nations. *American Economic Review*, 104(9):2736–2762, 2014. doi: 10.1257/aer.104.9.2736. URL <https://doi.org/10.1257/aer.104.9.2736>.
- Jan Marcus, Thomas Siedler, and Nicolas R. Ziebarth. The long-run effects of sports club vouchers for primary school children. *American Economic Journal: Economic Policy*, 14(3):128–165, 2022. doi: 10.1257/pol.20200431. URL <https://doi.org/10.1257/pol.20200431>.
- Monica Martinez-Bravo. The role of local officials in new democracies: Evidence from indonesia. *American Economic Review*, 104(4):1244–1287, 2014. doi: 10.1257/aer.104.4.1244. URL <https://doi.org/10.1257/aer.104.4.1244>.
- Benjamin M. Marx and Lesley J. Turner. Student loan nudges: Experimental evidence on borrowing and educational attainment. *American Economic Journal: Economic Policy*, 11(2):108–141, 2019. doi: 10.1257/pol.20180279. URL <https://doi.org/10.1257/pol.20180279>.
- Maurizio Mazzocco and Shiv Saini. Testing efficient risk sharing with heterogeneous risk preferences. *American Economic Review*, 102(1):428–468, 2012. doi: 10.1257/aer.102.1.428. URL <https://doi.org/10.1257/aer.102.1.428>.

- Adam McCloskey and Pascal Michailat. Critical values robust to p-hacking. *Review of Economics and Statistics*, pages 1–35, 2024. doi: 10.1162/rest_a_01456. URL https://doi.org/10.1162/rest_a_01456. Published online April 2024.
- Paul Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391, 1981. doi: 10.2307/3003562. URL <https://www.jstor.org/stable/3003562>.
- Conrad Miller. The persistent effect of temporary affirmative action. *American Economic Journal: Applied Economics*, 9(3):152–190, 2017. doi: 10.1257/app.20160121. URL <https://doi.org/10.1257/app.20160121>.
- Giuseppe Moscarini and Fabien Postel-Vinay. The relative power of employment-to-employment reallocation and unemployment exits in predicting wage growth. *American Economic Review*, 107(5): 364–368, 2017. doi: 10.1257/aer.p20171078. URL <https://doi.org/10.1257/aer.p20171078>.
- J. Peter Neary. Rationalizing the penn world table: True multilateral indices for international comparisons of real income. *American Economic Review*, 94(5):1411–1428, 2004. doi: 10.1257/0002828043052286. URL <https://doi.org/10.1257/0002828043052286>.
- M. Kathleen Ngangoué and Andrew Schotter. The common-probability auction puzzle. *American Economic Review*, 113(6):1572–1599, 2023. doi: 10.1257/aer.20191927. URL <https://doi.org/10.1257/aer.20191927>.
- Martina Björkman Nyqvist, Damien de Walque, and Jakob Svensson. Experimental evidence on the long-run impact of community-based monitoring. *American Economic Journal: Applied Economics*, 9(1):33–69, 2017. doi: 10.1257/app.20150027. URL <https://doi.org/10.1257/app.20150027>.
- Edward N. Okeke. When a doctor falls from the sky: The impact of easing doctor supply constraints on mortality. *American Economic Review*, 113(3):585–627, 2023. doi: 10.1257/aer.20210701. URL <https://doi.org/10.1257/aer.20210701>.
- Benjamin A. Olken. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80, 2015. doi: 10.1257/jep.29.3.61. URL <https://doi.org/10.1257/jep.29.3.61>.
- Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79):1–32, 2015. doi: 10.1214/EJP.v20-4039. URL <https://doi.org/10.1214/EJP.v20-4039>.
- Michael J. Pries and Richard Rogerson. Declining worker turnover: The role of short-duration employment spells. *American Economic Journal: Macroeconomics*, 14(1):260–300, 2022. doi: 10.1257/mac.20190230. URL <https://doi.org/10.1257/mac.20190230>.
- Carmen M Reinhart and Kenneth S Rogoff. From financial crash to debt crisis. *American Economic Review*, 101(5):1676–1706, 2011. doi: 10.1257/aer.101.5.1676. URL <https://doi.org/10.1257/aer.101.5.1676>.
- Elena Reutskaja, Rosemarie Nagel, Colin F Camerer, and Antonio Rangel. Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, 101(2): 900–926, 2011. doi: 10.1257/aer.101.2.900. URL <https://doi.org/10.1257/aer.101.2.900>.
- R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970. doi: 10.1515/9781400873173. URL <https://doi.org/10.1515/9781400873173>.

- Christina D. Romer and David H. Romer. New evidence on the aftermath of financial crises in advanced countries. *American Economic Review*, 107(10):3072–3118, 2017. doi: 10.1257/aer.20150320. URL <https://doi.org/10.1257/aer.20150320>.
- Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3): 638–641, 1979. doi: 10.1037/0033-2909.86.3.638. URL <https://doi.org/10.1037/0033-2909.86.3.638>.
- Federico Rossi. The relative efficiency of skilled labor across countries: Measurement and interpretation. *American Economic Review*, 112(1):235–266, 2022. doi: 10.1257/aer.20191852. URL <https://doi.org/10.1257/aer.20191852>.
- Ivan Rudik. Optimal climate policy when damages are unknown. *American Economic Journal: Economic Policy*, 12(2):340–373, 2020. doi: 10.1257/pol.20160541. URL <https://doi.org/10.1257/pol.20160541>.
- Silvia Saccardo and Marta Serra-Garcia. Enabling or limiting cognitive flexibility? evidence of demand for moral commitment. *American Economic Review*, 113(2):396–429, 2023. doi: 10.1257/aer.20201333. URL <https://doi.org/10.1257/aer.20201333>.
- Anya Samek and Chuck Longfield. Do thank-you calls increase charitable giving? expert forecasts and field experimental evidence. *American Economic Journal: Applied Economics*, 15(2):103–124, 2023. doi: 10.1257/app.20210068. URL <https://doi.org/10.1257/app.20210068>.
- Sonja Settele. How do beliefs about the gender wage gap affect the demand for public policy? *American Economic Journal: Economic Policy*, 14(2):475–508, 2022. doi: 10.1257/pol.20200559. URL <https://doi.org/10.1257/pol.20200559>.
- Albert N. Shiryaev. *Optimal Stopping Rules*, volume 8 of *Stochastic Modelling and Applied Probability*. Springer, 1978. doi: 10.1007/978-3-540-74011-7. URL <https://doi.org/10.1007/978-3-540-74011-7>. Reprinted by Springer, 2008.
- Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Proceedings of the National Academy of Sciences*, 108(34):13514–13519, 2011. doi: 10.1073/pnas.1018034108. URL <https://doi.org/10.1073/pnas.1018034108>.
- Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: 10.1038/s41562-020-0912-z. URL <https://doi.org/10.1038/s41562-020-0912-z>.
- J. L. Snell. Applications of martingale system theorems. *Transactions of the American Mathematical Society*, 73(2):293–312, 1952. doi: 10.1090/S0002-9947-1952-0050209-9. URL <https://doi.org/10.1090/S0002-9947-1952-0050209-9>.
- Jann Spiess. Optimal estimation when researcher and social preferences are misaligned. *Econometrica*, 93(5):1779–1810, September 2025. doi: 10.3982/ECTA18640. URL <https://doi.org/10.3982/ECTA18640>.
- Kevin M Stange. An empirical investigation of the option value of college enrollment. *American Economic Journal: Applied Economics*, 4(1):49–84, 2012. doi: 10.1257/app.4.1.49. URL <https://doi.org/10.1257/app.4.1.49>.

- Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637. URL <https://doi.org/10.1177/1745691616658637>.
- Claudia Steinwender. Real effects of information frictions: When the states and the kingdom became united. *American Economic Review*, 108(3):657–696, 2018. doi: 10.1257/aer.20150681. URL <https://doi.org/10.1257/aer.20150681>.
- Wendy A. Stock. Trends in economics and other undergraduate majors. *American Economic Review*, 107(5):644–649, 2017. doi: 10.1257/aer.p20171067. URL <https://doi.org/10.1257/aer.p20171067>.
- John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002. doi: 10.1111/1467-9868.00346. URL <https://doi.org/10.1111/1467-9868.00346>.
- Aleksey Tetenov. An economic theory of statistical testing. CeMMAP Working Paper CWP50/16, Centre for Microdata Methods and Practice (CeMMAP), September 2016. URL <https://doi.org/10.1920/wp.cem.2016.5016>.
- Davide Viviano, Kaspar Wüthrich, and Paul Niehaus. A model of multiple hypothesis testing. arXiv preprint 2104.13367, arXiv, 2026. URL <https://arxiv.org/abs/2104.13367>. arXiv:2104.13367v10, revised 18 February 2026.
- Abraham Wald. *Statistical Decision Functions*. John Wiley & Sons, New York, 1950. URL https://openlibrary.org/works/OL5729594W/Statistical_decision_functions.
- Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152. URL <https://doi.org/10.1111/1468-0262.00152>.
- Jhacova Williams. Historical lynchings and the contemporary voting behavior of blacks. *American Economic Journal: Applied Economics*, 14(3):224–253, 2022. doi: 10.1257/app.20190549. URL <https://doi.org/10.1257/app.20190549>.
- Nicholas Wilson. Child marriage bans and female schooling and labor market outcomes: Evidence from natural experiments in 17 low- and middle-income countries. *American Economic Journal: Economic Policy*, 14(3):449–477, 2022. doi: 10.1257/pol.20200008. URL <https://doi.org/10.1257/pol.20200008>.
- Jessica Wisdom, Julie S Downs, and George Loewenstein. Promoting healthy choices: Information versus convenience. *American Economic Journal: Applied Economics*, 2(2):164–178, 2010. doi: 10.1257/app.2.2.164. URL <https://doi.org/10.1257/app.2.2.164>.
- Jia Wu, Junsen Zhang, and Chunchao Wang. Student performance, peer effects, and friend networks: Evidence from a randomized peer intervention. *American Economic Journal: Economic Policy*, 15(1):510–542, 2023. doi: 10.1257/pol.20200563. URL <https://doi.org/10.1257/pol.20200563>.
- Wei Xiong and Jialin Yu. The chinese warrants bubble. *American Economic Review*, 101(6):2723–2753, 2011. doi: 10.1257/aer.101.6.2723. URL <https://doi.org/10.1257/aer.101.6.2723>.
- Cristobal Young and Katherine Holsteen. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, 46(1):3–40, January 2017. doi: 10.1177/0049124115610347. URL <https://doi.org/10.1177/0049124115610347>.

George Zanjani. Regulation, capital, and the evolution of organizational form in us life insurance. *American Economic Review*, 97(3):973–983, 2007. doi: 10.1257/aer.97.3.973. URL <https://doi.org/10.1257/aer.97.3.973>.

Appendix roadmap

The theoretical appendices (Appendix A) collect the technical constructions and proofs referenced in the main text. Appendix A.1 states standing assumptions, defines the primitives $(D_{\text{mix}}, n_{\text{eff}})$ and the report space, and gives a sufficient condition for the Bayes-factor tail assumptions in contractive location AR(1) models. Appendix A.2 develops envelope reduction, optimal-stopping tools, and reusable truncation inequalities. Appendix A.3 presents the robustness-check mechanism and proves Theorem 5.2. Appendix A.4 collects auxiliary inequalities and proves Theorem 4.3. Appendix A.5 develops the short-report divergence budget, isolates the information-from-selection term, and proves Theorem 4.5 and Corollary A.43. Appendix A.6 provides Gaussian-specific calculations, including posterior-tail control for robustness checks and verification routes for the AR(1) sufficient condition and the short-report selection-control bound. Appendix A.7 gives modular generalizations of the lower bound, short-report bound, and robustness-check achievability under weaker conditions.

The empirical appendix (Appendix B) documents the replication artifacts, verification protocol, and estimation procedures supporting Section 6. Appendix B.1 describes the surface-driven replication and specification-search pipeline, including typing and mechanical validation. Appendix B.2 defines the samples, inference harmonization, and validation statistics. Appendix B.3 reports estimation and diagnostics for the three-type evidence model. Appendix B.4 estimates within-paper dependence. Appendix B.5 reports the counterfactual disclosure-scaling analysis and robustness checks.

We use large language models as part of the replication-and-verification workflow described in Appendix B.1, primarily to read heterogeneous replication packages, draft and edit runner scripts, and produce structured configuration artifacts. We also used language models for proofreading, proof-checking, and general editing assistance during the development of the manuscript.

A Theoretical Appendix

A.1 Environment assumptions and primitives

This subsection states the regularity assumptions and primitives used in Sections 3 and 4, and fixes the report-space measurability conventions used throughout Appendix A.2 and Appendix A.5.

A.1.1 Formal standing assumptions and primitives

This appendix collects the full regularity conditions that underlie the main-text primitives. In the main text we emphasize the economic content (bounded informativeness, effective sample size, and no perfect revelation) and use the resulting indices κ_{KL} and κ_{eff} .

We briefly restate the environment and notation used throughout the paper. A submission has an unobserved type T in a finite set \mathcal{T} , with prior $\mathbb{P}(T = t) = \pi_t$. Types are partitioned into high-impact and non-high,

$$\mathcal{H} \subseteq \mathcal{T} \quad \mathcal{T}_0 \equiv \mathcal{T} \setminus \mathcal{H} \quad \pi_H \equiv \mathbb{P}(T \in \mathcal{H}), \quad \pi_0 \equiv 1 - \pi_H$$

and we write $\mathbb{P}_H(\cdot)$ and $\mathbb{P}_0(\cdot)$ for laws conditional on $T \in \mathcal{H}$ and $T \in \mathcal{T}_0$ respectively.

Assumption A.1. Fix θ . There is a Polish state space X , a measurable map $\mathbf{p} : \mathsf{X} \rightarrow (0, 1)$, and for each type $t \in \mathcal{T}$ a Markov kernel K_t^θ on X with stationary distribution ν_t such that

- (i) Conditional on $T = t$, the latent chain $(X_n)_{n \geq 1}$ is stationary with law ν_t and transition kernel K_t^θ , and for every type t , $\nu_t \circ \mathbf{p}^{-1}$ is atomless on $(0, 1)$

- (ii) The observed p -values satisfy $P_n = \mathbf{p}(X_n)$ for all n
- (iii) There exists $n_{\text{eff}}(\theta) \in (0, 1]$ and a universal constant $C_H > 0$ such that for every type $t \in \mathcal{T}$, every Borel set $B \subset (0, 1)$, every $n \geq 1$, and every $\varepsilon \in (0, 1)$,

$$\begin{aligned}\mathbb{P}_t\left(\frac{1}{n}S_n(B) \geq p_t(B) + \varepsilon\right) &\leq \exp(-C_H n_{\text{eff}}(\theta) n \varepsilon^2) \\ \mathbb{P}_t\left(\frac{1}{n}S_n(B) \leq p_t(B) - \varepsilon\right) &\leq \exp(-C_H n_{\text{eff}}(\theta) n \varepsilon^2)\end{aligned}$$

where $p_t(B) \equiv \mathbb{P}_t(P_1 \in B)$ and $S_n(B) \equiv \sum_{k=1}^n \mathbf{1}\{P_k \in B\}$.

- (iv) There exists a sigma-finite dominating measure λ such that $\nu_t \ll \lambda$ and $K_t^\theta(x, \cdot) \ll \lambda$ for all x
- (v) For each high type $t_H \in \mathcal{H}$ and non-high type $t_0 \in \mathcal{T}_0$, we have $K_{t_H}^\theta(x, \cdot) \ll K_{t_0}^\theta(x, \cdot)$ for all x and $\nu_{t_H} \ll \nu_{t_0}$
- (vi) Define the one-step KL envelope from a high type to a non-high type by

$$D(t_H \rightarrow t_0; \theta) \equiv \sup_{x \in \mathbf{X}} \text{KL}\left(K_{t_H}^\theta(x, \cdot) \parallel K_{t_0}^\theta(x, \cdot)\right).$$

Let μ_0 denote the conditional law of T given $T \in \mathcal{T}_0$, and define

$$D_{\text{mix}}(\theta) \equiv \max_{t_H \in \mathcal{H}} \int D(t_H \rightarrow t_0; \theta) \mu_0(dt_0) < \infty. \quad (5)$$

Assume also there exists $C_\nu < \infty$ such that

$$\sup_{t_H \in \mathcal{H}} \sup_{t_0 \in \mathcal{T}_0} \text{KL}(\nu_{t_H} \parallel \nu_{t_0}) \leq C_\nu. \quad (6)$$

Stationarity in Assumption A.1(i) is a normalization. If the initial evidence state is not stationary, one can enlarge the type space \mathcal{T} to incorporate a finite set of initial conditions or allow a finite burn-in; this affects only additive constants (such as C_ν) and does not change the scaling results.

Assumption A.1(iii) is stated directly in terms of a concentration exponent that can be interpreted as an *effective number of independent draws*. It can be verified via many tools; for example, if each latent chain is reversible with a type-uniform L^2 spectral gap lower bound, then the standard spectral-gap concentration inequality (Appendix A.2, Lemma A.25) implies (iii) (up to constants).

We use Assumption A.1(iii) only for indicator functions of evidence windows, i.e. for $f = \mathbf{1}\{P \in B\}$.

Assumption A.2. Maintain Assumption A.1. There exist a Borel set $B_0 \subset (0, 1)$ and constants $\ell_0 > 0$ and $p_H(B_0) \in (0, 1]$ such that:

- (i) For every non-high type $t \in \mathcal{T}_0$,

$$\frac{d\mathcal{L}_H(P_1)}{d\mathcal{L}_t(P_1)}(p) \geq e^{\ell_0} \quad \text{for } \mathcal{L}_t(P_1)\text{-a.e. } p \in B_0$$

- (ii) For every high-impact type $t \in \mathcal{H}$,

$$\mathbb{P}_t(P_1 \in B_0) \geq p_H(B_0)$$

Definition A.3. Fix a strictly monotone continuous bijection $\psi : (0, 1) \rightarrow \mathbb{R}$ and define the singleton index by

$$Z_n \equiv \psi(P_n)$$

We adopt the convention that larger Z_n corresponds to more favorable evidence; for p -values this means ψ is decreasing. In Example ?? we take $\psi(p) = \Phi^{-1}(1 - p)$ so that Z_n is the usual z-score.

Definition A.4. Maintain the standing regularity condition (Appendix A.1.1). Let $\mathcal{L}_H(P_1)$ and $\mathcal{L}_0(P_1)$ denote the induced one-test laws of the singleton report under $\mathbb{P}(\cdot \mid T \in \mathcal{H})$ and $\mathbb{P}(\cdot \mid T \in \mathcal{T}_0)$. By Assumption A.1(v) and finiteness of \mathcal{T}_0 , we have $\mathcal{L}_H(P_1) \ll \mathcal{L}_0(P_1)$, so the Radon–Nikodym derivative below is well-defined. Let

$$\ell(p) \equiv \log \left(\frac{d\mathcal{L}_H(P_1)}{d\mathcal{L}_0(P_1)}(p) \right)$$

denote a version of the singleton log-likelihood ratio. For $c \in \mathbb{R}$, define the LR superlevel set

$$B(c) \equiv \{p \in (0, 1) : \ell(p) \geq c\}$$

Definition A.5. Maintain the standing regularity condition (Appendix A.1.1). Let $n_{\text{eff}}(\theta) \in (0, 1]$ denote the effective-sample-size rate from Assumption A.1(iii). Let $D_{\text{mix}}(\theta)$ denote the per-test information bound defined in (5). Define the indices

$$\kappa_{\text{eff}} \equiv \frac{n_{\text{eff}}(\theta)}{\gamma} \quad \kappa_{\text{KL}} \equiv \frac{D_{\text{mix}}(\theta)}{\gamma}$$

Assumption A.6 (Subset Rényi growth). Fix $\alpha > 1$. There exists $d_\alpha(\theta) < \infty$ such that for every $n \geq 1$, every index set $J \subseteq \{1, \dots, n\}$, and every non-high type $t_0 \in \mathcal{T}_0$,

$$D_\alpha(\mathcal{L}_H(Z_J) \parallel \mathcal{L}_{t_0}(Z_J)) \leq |J| d_\alpha(\theta),$$

where D_α is the order- α Rényi divergence (Appendix A.4.1). In particular, since $\mathcal{L}_0(Z_J)$ is a mixture over $t_0 \in \mathcal{T}_0$, the same bound holds with \mathcal{L}_0 in place of \mathcal{L}_{t_0} (by Jensen’s inequality, using convexity of $x \mapsto x^{1-\alpha}$ for $\alpha > 1$).

Assumption A.6 controls how much H -vs-0 likelihood-ratio information can be carried by a short reported subset of coordinates, and it is the marginal input in the truncated-report Rényi budget (Appendix A.5).

A.1.2 Report space and measurability

We represent a report as a finite counting measure on $(0, 1)$. Let \mathcal{R} denote the space of finite counting measures on $(0, 1)$ equipped with the σ -field generated by evaluation maps $r \mapsto r(B)$ for Borel $B \subseteq (0, 1)$. We write $r \preceq r'$ if r is a subreport of r' (equivalently, a submultiset), i.e. if $r(B) \leq r'(B)$ for every Borel set B .

For a measurable policy $\delta : \mathcal{R} \rightarrow [0, 1]$, define its *disclosure envelope* $\delta^\dagger : \mathcal{R} \rightarrow [0, 1]$ by

$$\delta^\dagger(r) \equiv \sup_{u \preceq r} \delta(u)$$

In our setting reports are finite, so the supremum is a maximum.

Lemma A.7. *If $\delta : \mathcal{R} \rightarrow [0, 1]$ is measurable, then its disclosure envelope δ^\dagger is measurable (and hence is a valid acceptance policy).*

Proof. For $n \geq 0$ let $\mathcal{R}_n \equiv \{r \in \mathcal{R} : r((0, 1)) = n\}$ denote the measurable subset of reports with total mass n . On \mathcal{R}_n , define the i th order statistic of a report by the quantile formula

$$p_{(i)}(r) \equiv \inf\{x \in (0, 1) : r((0, x]) \geq i\} \quad i = 1, \dots, n$$

This is measurable because for any $q \in (0, 1)$ we have $\{p_{(i)}(r) \leq q\} = \{r((0, q]) \geq i\}$, and $r \mapsto r((0, q])$ is an evaluation map. For each subset $I \subseteq \{1, \dots, n\}$, define the corresponding subreport $r_I \in \mathcal{R}$ by

$$r_I(B) \equiv \sum_{i \in I} \mathbf{1}\{p_{(i)}(r) \in B\} \quad B \subseteq (0, 1) \text{ Borel}$$

Then $r \mapsto r_I$ is measurable on \mathcal{R}_n , and every subreport $u \preceq r$ is equal to r_I for some I (select the indices of the atoms included, counting multiplicity). Therefore on \mathcal{R}_n ,

$$\delta^\uparrow(r) = \max_{I \subseteq \{1, \dots, n\}} \delta(r_I)$$

which is a finite maximum of measurable functions, hence measurable. Since $\mathcal{R} = \bigsqcup_{n \geq 0} \mathcal{R}_n$ is a measurable partition, δ^\uparrow is measurable on all of \mathcal{R} . \square

A.1.3 Posterior tails for robustness checks

The robustness-check analysis requires a long-horizon tail bound for the Bayes factor (equivalently, for the posterior probability of being high under non-high data generating processes). We state it separately because it is only used in Appendix A.3 and related robustness-check arguments.

Assumption A.8 (Posterior tails). Fix $\alpha > 1$. For each $t \in \mathcal{T}_0$, write $\pi_t^0 \equiv \mathbb{P}(T = t \mid T \in \mathcal{T}_0)$ and let

$$\pi_{\min}^0 \equiv \min_{t \in \mathcal{T}_0} \pi_t^0 > 0.$$

Define the prior-dependent constant

$$C_{\pi, \alpha} \equiv |\mathcal{T}_0| \cdot \left(\frac{\pi_H}{\pi_0 \pi_{\min}^0} \right)^{(\alpha-1)/\alpha}.$$

There exists $C_\alpha(\theta) > 0$ such that for every $n \geq 1$ and every $\varepsilon \in (0, 1)$,

$$\mathbb{P}_0(\pi_n(H) > \varepsilon) \leq C_{\pi, \alpha} \varepsilon^{-(\alpha-1)/\alpha} \exp\left(-\frac{\alpha-1}{\alpha} C_\alpha(\theta) n_{\text{eff}}(\theta) (n-1)\right),$$

where $\pi_n(H) \equiv \mathbb{P}(T \in \mathcal{H} \mid \mathcal{F}_n)$ is the posterior probability that the submission is high after n observations.

Assumption A.8 controls how unlikely it is, under non-high data generating processes, for the posterior to place nontrivial mass on being high after many observations, and it closes the posterior-tail term in the robustness-check bound (Appendix A.3).

A.1.4 A sufficient condition in contractive location AR(1) models

Assumption A.6 is stated abstractly because it enters several arguments. In contractive location AR(1) models, Assumptions A.6 and A.8 follow from one-dimensional quadratic bounds for the innovation translate family. Proposition A.9 records one convenient sufficient condition. Appendix A.6 verifies its translate inequalities in the Gaussian example. Selection control is treated separately: Appendix A.5.2 gives a general sufficient condition for Assumption A.38, and Appendix A.6.5 verifies it in the Gaussian AR(1) example.

Proposition A.9. Consider a stationary mean-shift AR(1) evidence process of the form

$$Z_{n+1} = \phi Z_n + (1 - \phi)\mu_t + \varepsilon_{n+1}, \quad \phi \in [0, 1),$$

where $(\varepsilon_n)_{n \geq 1}$ are i.i.d., independent of T , and their law does not depend on t . For $u \in \mathbb{R}$, write $F_u \equiv \mathcal{L}(\varepsilon_1 + u)$. Fix $\alpha > 1$ and write $\lambda \equiv (\alpha - 1)/\alpha$. Assume $\mathcal{H} = \{H\}$ is a singleton and denote the high-type mean by μ_H . Assume there exist constants $v_\alpha(\theta) \in (0, \infty)$ and $v_\lambda(\theta) \in (0, \infty)$ such that for all $u, v \in \mathbb{R}$,

$$D_\alpha(F_u \| F_v) \leq \frac{\alpha}{2v_\alpha(\theta)}(u - v)^2$$

$$\mathbb{E}_{X \sim F_v} \left[\left(\frac{dF_u}{dF_v}(X) \right)^\lambda \right] \leq \exp\left(-\frac{\lambda}{2\alpha v_\lambda(\theta)}(u - v)^2 \right)$$

Then Assumption A.6 holds with

$$d_\alpha(\theta) \equiv \frac{\alpha}{2v_\alpha(\theta)} \Delta_{\max}^2 \quad \Delta_{\max} \equiv \max_{t_0 \in \mathcal{T}_0} |\mu_H - \mu_{t_0}|.$$

Moreover, define the minimum mean gap

$$\Delta_{\min} \equiv \min_{t_0 \in \mathcal{T}_0} |\mu_H - \mu_{t_0}|.$$

Then Assumption A.8 holds with

$$C_\alpha(\theta) = \frac{(1 - \phi)^2 \Delta_{\min}^2}{2\alpha v_\lambda(\theta) n_{\text{eff}}(\theta)}.$$

Proof. We verify the two claims in turn.

Subset Rényi growth. Fix $n \geq 1$, fix $J \subseteq \{1, \dots, n\}$ with $J = \{j_1 < \dots < j_k\}$, and fix a non-high type $t_0 \in \mathcal{T}_0$. Write $\Delta \equiv \mu_H - \mu_{t_0}$. Let $P_J \equiv \mathcal{L}_H(Z_J)$ and $Q_J \equiv \mathcal{L}_{t_0}(Z_J)$ and write $L_J \equiv dP_J/dQ_J(Z_J)$. By definition,

$$\mathbb{E}_{t_0}[L_J^\alpha] = \exp((\alpha - 1)D_\alpha(P_J \| Q_J)).$$

We bound this α -moment by iterating along the Markov subsequence.

First consider Z_{j_1} . Under any type t , stationarity yields

$$Z_{j_1} = \mu_t + \varepsilon_{j_1} + \sum_{r=1}^{\infty} \phi^r \varepsilon_{j_1-r},$$

where the infinite sum is independent of ε_{j_1} and has a type-independent law. Therefore, by the data-processing inequality for Rényi divergence, adding the independent noise term $\sum_{r=1}^{\infty} \phi^r \varepsilon_{j_1-r}$ cannot increase D_α , hence

$$D_\alpha(\mathcal{L}_H(Z_{j_1}) \| \mathcal{L}_{t_0}(Z_{j_1})) \leq D_\alpha(F_{\mu_H} \| F_{\mu_{t_0}}) \leq \frac{\alpha}{2v_\alpha(\theta)} \Delta^2.$$

Next, for each $\ell \geq 2$ with gap $m = j_\ell - j_{\ell-1}$, write $c_m \equiv 1 - \phi^m$. The recursion yields

$$Z_{j_\ell} = \phi^m Z_{j_{\ell-1}} + c_m \mu_t + \sum_{r=0}^{m-1} \phi^r \varepsilon_{j_\ell-r},$$

so conditional on $Z_{j_{\ell-1}}$ the m -step transition differs across types only by the translate shift $c_m \Delta$. Writing $B_\ell \equiv \sum_{r=0}^{m-1} \phi^r \varepsilon_{j_\ell-r} = \varepsilon_{j_\ell} + W_\ell$ with W_ℓ independent of ε_{j_ℓ} , the same data-processing argument gives

$$D_\alpha(\mathcal{L}(B_\ell + c_m \mu_H) \parallel \mathcal{L}(B_\ell + c_m \mu_{t_0})) \leq D_\alpha(F_{c_m \mu_H} \parallel F_{c_m \mu_{t_0}}) \leq \frac{\alpha}{2v_\alpha(\theta)} c_m^2 \Delta^2 \leq \frac{\alpha}{2v_\alpha(\theta)} \Delta^2.$$

This bound is uniform in $Z_{j_{\ell-1}}$ because shifting both laws by $\phi^m Z_{j_{\ell-1}}$ does not change Rényi divergence.

Let $L_{j_1} \equiv d\mathcal{L}_H(Z_{j_1})/d\mathcal{L}_{t_0}(Z_{j_1})(Z_{j_1})$ and for each $\ell \geq 2$ let

$$L_{j_\ell} \equiv \frac{d\mathcal{L}_H(Z_{j_\ell} \mid Z_{j_{\ell-1}})}{d\mathcal{L}_{t_0}(Z_{j_\ell} \mid Z_{j_{\ell-1}})}(Z_{j_\ell}).$$

By the Markov property, $L_J = L_{j_1} \prod_{\ell=2}^k L_{j_\ell}$ almost surely under \mathbb{P}_{t_0} . Moreover, for each $\ell \geq 2$ the conditional moment satisfies

$$\mathbb{E}_{t_0}[L_{j_\ell}^\alpha \mid Z_{j_{\ell-1}}] = \exp\left((\alpha - 1)D_\alpha(\mathcal{L}_H(Z_{j_\ell} \mid Z_{j_{\ell-1}}) \parallel \mathcal{L}_{t_0}(Z_{j_\ell} \mid Z_{j_{\ell-1}}))\right),$$

and similarly $\mathbb{E}_{t_0}[L_{j_1}^\alpha] = \exp((\alpha - 1)D_\alpha(\mathcal{L}_H(Z_{j_1}) \parallel \mathcal{L}_{t_0}(Z_{j_1})))$. Combining the divergence bounds above and iterating the tower property yields

$$\mathbb{E}_{t_0}[L_J^\alpha] \leq \exp\left((\alpha - 1)|J| \cdot \frac{\alpha}{2v_\alpha(\theta)} \Delta^2\right),$$

and taking logs gives

$$D_\alpha(\mathcal{L}_H(Z_J) \parallel \mathcal{L}_{t_0}(Z_J)) \leq |J| \cdot \frac{\alpha}{2v_\alpha(\theta)} \Delta^2 \leq |J| \cdot d_\alpha(\theta),$$

where the last inequality uses $|\Delta| \leq \Delta_{\max}$.

Posterior tails. Fix $n \geq 1$, $\varepsilon \in (0, 1)$, and $t_0 \in \mathcal{T}_0$. By Bayes' rule,

$$\pi_n(H) = \frac{\pi_H p_H(Z_{1:n})}{\pi_H p_H(Z_{1:n}) + \sum_{u \in \mathcal{T}_0} \pi_u p_u(Z_{1:n})} \leq \frac{\pi_H}{\pi_{t_0}} \cdot \frac{p_H(Z_{1:n})}{p_{t_0}(Z_{1:n})}.$$

Applying Markov's inequality with exponent λ yields

$$\mathbb{P}_{t_0}(\pi_n(H) > \varepsilon) \leq \left(\frac{\pi_H}{\pi_{t_0} \varepsilon}\right)^\lambda \mathbb{E}_{t_0} \left[\left(\frac{p_H(Z_{1:n})}{p_{t_0}(Z_{1:n})} \right)^\lambda \right].$$

Under t_0 , the likelihood ratio $p_H(Z_{1:n})/p_{t_0}(Z_{1:n})$ factors into a marginal term and $n - 1$ one-step transition terms. Write $L_1 \equiv p_H(Z_1)/p_{t_0}(Z_1)$ and $L_{t+1} \equiv p_H(Z_{t+1} \mid Z_t)/p_{t_0}(Z_{t+1} \mid Z_t)$ for $t \geq 1$, so $p_H(Z_{1:n})/p_{t_0}(Z_{1:n}) = L_1 \prod_{t=1}^{n-1} L_{t+1}$. Since $\lambda \in (0, 1)$, the map $x \mapsto x^\lambda$ is concave. Because $\mathbb{E}_{t_0}[L_1] = 1$, Jensen's inequality gives $\mathbb{E}_{t_0}[L_1^\lambda] \leq 1$. Conditional on Z_t , the $(t+1)$ st transition compares the translate laws $F_{\phi Z_t + (1-\phi)\mu_H}$ and $F_{\phi Z_t + (1-\phi)\mu_{t_0}}$. Since their shift difference is $(1-\phi)(\mu_H - \mu_{t_0})$, the translate-Chernoff bound gives

$$\mathbb{E}_{t_0}[L_{t+1}^\lambda \mid Z_t] \leq \exp\left(-\frac{\lambda}{2\alpha v_\lambda(\theta)} (1-\phi)^2 (\mu_H - \mu_{t_0})^2\right).$$

The right-hand side is constant in Z_t , so iterating conditional expectations yields

$$\mathbb{E}_{t_0} \left[\left(\frac{p_H(Z_{1:n})}{p_{t_0}(Z_{1:n})} \right)^\lambda \right] \leq \exp\left(-\frac{\lambda}{2\alpha v_\lambda(\theta)} (1-\phi)^2 (\mu_H - \mu_{t_0})^2 (n-1)\right).$$

Finally, summing over $t_0 \in \mathcal{T}_0$ yields the \mathbb{P}_0 bound in Assumption A.8 with the minimum gap Δ_{\min} . \square

A.2 Equilibrium and stopping-time tools

This subsection records the equilibrium reduction, stopping-time inequalities, and monotonicity tools used in Section 3 and in the proofs of the frontier bounds (Appendices A.4–A.6). Appendix A.3 collects the mechanism-specific constructions and proofs used in Section 5. Throughout, ‘throughput’ refers to the journal’s expected acceptance rate, while $1/\gamma$ is the researcher’s testing-capacity scale.

A.2.1 Equilibrium selection and reporting

The researcher chooses a stopping time τ and a report $R \in \mathcal{R}$ (a finite multiset of realized p -values; Appendix A.1.2). Because omission is free, after stopping she discloses an envelope-attaining subreport. Lemma A.15 formalizes that, for any policy δ , the best-response problem reduces to an optimal stopping problem with payoff $M_\tau^\delta - \gamma\tau$, and Proposition A.16 provides an earliest optimal stopping-time tie-break.

A.2.2 Histories and down-sets

Recall the singleton index $Z_n = \psi(P_n)$ from Definition A.3. Write the length- n singleton history as $Z_{1:n} \equiv (Z_1, \dots, Z_n) \in \mathbb{R}^n$. We use the coordinatewise order on \mathbb{R}^n : for $x, y \in \mathbb{R}^n$, write $x \leq y$ if $x_i \leq y_i$ for all i .

Definition A.10. Fix $n \geq 1$. A measurable set $D \subseteq \mathbb{R}^n$ is a *down-set* if $x \in D$ and $y \leq x$ imply $y \in D$. Equivalently, D is coordinatewise decreasing. An *up-set* is a coordinatewise increasing set.

Definition A.11. Fix $n \geq 1$. A measurable set $E \subseteq \mathbb{R}^n$ is *two-sided monotone* if there exist a down-set D and an up-set U such that $E = D \cup U$.

In a one-dimensional ‘‘window’’ rule, continuation often occurs in either tail: e.g. $\{Z \notin [a, b]\} = \{Z \leq a\} \cup \{Z \geq b\}$ is two-sided monotone. More generally, if continuation is triggered by being outside an interval in a scalar statistic that is increasing in the history, then conditioning on any subvector preserves the two-sided monotone structure in the omitted coordinates.

Lemma A.12. Fix $n \geq 1$ and an index set $J \subseteq \{1, \dots, n\}$. Let $S : \mathbb{R}^n \rightarrow \mathbb{R}$ be coordinatewise nondecreasing and fix $a \leq b$. For each realization $z_J \in \mathbb{R}^{|J|}$, define the induced event in omitted coordinates

$$E(z_J) \equiv \left\{ z_{-J} \in \mathbb{R}^{n-|J|} : S(z_J, z_{-J}) \notin (a, b) \right\}$$

Then $E(z_J)$ is two-sided monotone in $\mathbb{R}^{n-|J|}$ (Definition A.11).

Proof. Fix z_J . Since $z_{-J} \mapsto S(z_J, z_{-J})$ is coordinatewise nondecreasing, the sets

$$D \equiv \{z_{-J} \in \mathbb{R}^{n-|J|} : S(z_J, z_{-J}) \leq a\} \quad U \equiv \{z_{-J} \in \mathbb{R}^{n-|J|} : S(z_J, z_{-J}) \geq b\}$$

are a down-set and an up-set respectively. Their union is $E(z_J)$. □

A.2.3 Gaussian AR(1): belief state is one-dimensional

In the ordered three-type setting $\mu_N < \mu_H < \mu_L$, extremely small p -values (very large $Z = \Phi^{-1}(1-p)$) can be more indicative of a biased (exogeneity-failure) type L than of the credible causal type H . Equivalently, the singleton log-likelihood ratio $\ell(p)$ from Definition A.4 need not be monotone in Z . Thus likelihood-ratio superlevel acceptance regions can correspond to an interval in the raw index Z (‘‘significant but not too significant’’) rather than a one-sided threshold.

In the Gaussian AR(1) mean-shift model, let $M \equiv \mu_T$ denote the (static) mean parameter and write $Z_{1:n} = (Z_1, \dots, Z_n)$. Conditional on $M = \mu$, we have $Z_{1:n} \sim \mathcal{N}(\mu \mathbf{1}, \Sigma_\phi)$, where Σ_ϕ is the $n \times n$ AR(1) correlation matrix with (i, j) entry $\phi^{|i-j|}$. Therefore the likelihood of μ given $Z_{1:n} = z_{1:n}$ is of exponential-family form

$$L_n(\mu; z_{1:n}) \propto \exp\left(\mu s_n(z_{1:n}) - \frac{1}{2}\mu^2 b_n\right) \quad s_n(z_{1:n}) = \mathbf{1}^\top \Sigma_\phi^{-1} z_{1:n} \quad b_n = \mathbf{1}^\top \Sigma_\phi^{-1} \mathbf{1}$$

Hence the posterior $\mathcal{L}(M \mid Z_{1:n})$ depends on the realized singleton history only through (n, s_n) . This scalar statistic is the natural Markov state for the researcher's POMDP in the Gaussian AR(1) example, including the three-type case $\mu_N < \mu_H < \mu_L$. Moreover, for $\phi \in [0, 1)$ the weight vector $\Sigma_\phi^{-1} \mathbf{1}$ has nonnegative entries (the row sums of the AR(1) precision matrix), so s_n is coordinatewise increasing in $z_{1:n}$.

A.2.4 Reusable stopping-time inequalities

Lemma A.13. *Fix any acceptance policy δ and let $\sigma^*(\delta)$ be a best response with stopping time τ^* and induced report R . Then $\mathbb{E}[\tau^*] \leq 1/\gamma$ (ex ante, under the prior), hence $\tau^* < \infty$ almost surely. Moreover,*

$$\mathbb{E}[\tau^* \mid T \in \mathcal{H}] \leq \frac{1}{\pi_H} \mathbb{E}[\tau^*] \leq \frac{1}{\pi_H \gamma} \quad \mathbb{E}[\tau^* \mid T \in \mathcal{T}_0] \leq \frac{1}{\pi_0} \mathbb{E}[\tau^*] \leq \frac{1}{\pi_0 \gamma}$$

Proof. The researcher can always stop immediately and receive a payoff in $[0, 1]$, so any best response has nonnegative payoff. For any strategy, $\mathbb{E}[\delta(R)] - \gamma \mathbb{E}[\tau] \leq 1 - \gamma \mathbb{E}[\tau]$, hence $0 \leq 1 - \gamma \mathbb{E}[\tau^*]$ and $\mathbb{E}[\tau^*] \leq 1/\gamma$. Since $\mathbb{E}[\tau^*] < \infty$, necessarily $\mathbb{P}(\tau^* = \infty) = 0$. Finally,

$$\mathbb{E}[\tau^*] = \pi_H \mathbb{E}[\tau^* \mid T \in \mathcal{H}] + \pi_0 \mathbb{E}[\tau^* \mid T \in \mathcal{T}_0]$$

and since both conditional expectations are nonnegative,

$$\pi_H \mathbb{E}[\tau^* \mid T \in \mathcal{H}] \leq \mathbb{E}[\tau^*] \quad \pi_0 \mathbb{E}[\tau^* \mid T \in \mathcal{T}_0] \leq \mathbb{E}[\tau^*]$$

Rearranging yields the conditional-time bounds. \square

Lemma A.14. *Let $(E_n)_{n \geq 0}$ be an increasing sequence of events ($E_n \subseteq E_{n+1}$). For any stopping time τ and any deterministic n ,*

$$\mathbf{1}\{E_\tau\} \leq \mathbf{1}\{E_n\} + \mathbf{1}\{\tau > n\}$$

Consequently, under any law \mathbb{P} ,

$$\mathbb{P}(E_\tau) \leq \mathbb{P}(E_n) + \mathbb{P}(\tau > n)$$

Proof. If $\tau \leq n$, then $E_\tau \subseteq E_n$ by monotonicity; if $\tau > n$, the second indicator equals 1. \square

A.2.5 Proofs for Section 3

Recall the report space \mathcal{R} and the subreport order \preceq from Appendix A.1.2. For $n \geq 1$, define the full report (counting measure)

$$R_n^{\text{full}}(B) \equiv \sum_{i=1}^n \mathbf{1}\{P_i \in B\}, \quad B \subseteq (0, 1) \text{ Borel}$$

and let \emptyset denote the empty report (the zero counting measure).

Because omission is unverifiable, after any realized history the researcher can discard unfavorable outcomes and submit whichever subreport maximizes acceptance probability. Given a policy $\delta : \mathcal{R} \rightarrow [0, 1]$, define the *envelope* acceptance value

$$M_n^\delta \equiv \max_{R \preceq R_n^{\text{full}}} \delta(R) \quad M_0^\delta \equiv \delta(\emptyset)$$

Equivalently, $M_n^\delta = \delta^\uparrow(R_n^{\text{full}})$ for the disclosure envelope δ^\uparrow from Appendix A.1.2. Lemma A.15 formalizes that the researcher can always attain M_τ^δ at any stopping time τ , so her best-response problem reduces to an optimal stopping problem with payoff $M_\tau^\delta - \gamma\tau$.

Lemma A.15. *Fix δ and let (τ, R) be any researcher strategy. Then*

$$\delta(R) \leq M_\tau^\delta \quad \text{almost surely}$$

Conversely, for every stopping time τ there exists an \mathcal{F}_τ -measurable report R^{max} such that

$$\delta(R^{\text{max}}) = M_\tau^\delta \quad \text{almost surely}$$

Consequently, the researcher's problem under δ reduces to the optimal stopping problem

$$\sup_{\tau} \mathbb{E}[M_\tau^\delta - \gamma\tau]$$

In particular, equilibrium outcomes depend on δ only through the envelope process $(M_n^\delta)_{n \geq 0}$.

Proof. For the first claim, $\delta(R)$ is one feasible subreport acceptance value at time τ , so it cannot exceed the maximum over feasible subreports, i.e. $\delta(R) \leq M_\tau^\delta$.

For the second claim, fix n and enumerate the subsets $I \subseteq \{1, \dots, n\}$ in a deterministic order (e.g. by increasing $|I|$ and then lexicographically). Each I induces a feasible subreport $R_{n,I} \in \mathcal{R}$ defined by $R_{n,I}(B) \equiv \sum_{i \in I} \mathbf{1}\{P_i \in B\}$. Let R_n^{max} be the first subreport in this list that attains $\max_{R \preceq R_n^{\text{full}}} \delta(R)$. This selection is measurable as a function of (P_1, \dots, P_n) because it is obtained by finitely many comparisons of measurable maps. Defining $R^{\text{max}} \equiv R_\tau^{\text{max}}$ yields an \mathcal{F}_τ -measurable report satisfying $\delta(R^{\text{max}}) = M_\tau^\delta$ almost surely.

The reduction to optimal stopping follows immediately: given τ , the researcher can achieve M_τ^δ by reporting R^{max} and cannot do better by the first claim. \square

Fix δ and write $M_n \equiv M_n^\delta$. Define the value-to-go process

$$\mathcal{V}_n \equiv \text{ess sup}_{\tau \geq n} \mathbb{E}[M_\tau - \gamma(\tau - n) \mid \mathcal{F}_n],$$

where the essential supremum is over stopping times τ with $\tau \geq n$ almost surely. Define the earliest time at which continuation has no value by

$$\tau^{\text{early}}(\delta) \equiv \inf\{n \geq 0 : \mathcal{V}_n = M_n\}.$$

Proposition A.16. *Fix δ and write $M_n \equiv M_n^\delta$. Then $\tau^{\text{early}}(\delta)$ is optimal for $\sup_{\tau} \mathbb{E}[M_\tau - \gamma\tau]$ and is the earliest optimal stopping time among all optimizers. Moreover, every optimal stopping time τ satisfies $\mathbb{E}[\tau] \leq 1/\gamma$.*

Proof. This is a standard Snell-envelope argument for discrete-time optimal stopping with a running cost; see, e.g., [Snell \(1952\)](#) or [Shiryaev \(1978, Ch. 2, especially §2.14\)](#). Define $X_n := M_n - \gamma n$ and its Snell envelope

$$\mathcal{W}_n := \operatorname{ess\,sup}_{\tau \geq n} \mathbb{E}[X_\tau \mid \mathcal{F}_n].$$

Standard results imply (\mathcal{W}_n) is the smallest supermartingale dominating (X_n) , satisfies

$$\mathcal{W}_n = \max \left(X_n, \mathbb{E}[\mathcal{W}_{n+1} \mid \mathcal{F}_n] \right),$$

and that the first hitting time $\inf\{n \geq 0 : \mathcal{W}_n = X_n\}$ is optimal and earliest among all optimizers. Translating back via $\mathcal{V}_n = \mathcal{W}_n + \gamma n$ gives the claim.

For the expectation bound, note $0 \leq M_n \leq 1$. If τ is optimal then $0 \leq \mathbb{E}[M_\tau - \gamma\tau] \leq 1 - \gamma\mathbb{E}[\tau]$, hence $\mathbb{E}[\tau] \leq 1/\gamma$. \square

A.3 Robustness-check mechanism analysis

This subsection collects the mechanism-specific definitions and proofs for the robustness-check (certificate) policies studied in [Section 5](#).

A.3.1 Robustness-check rules and separation

Definition A.17. Fix an integer $m \geq 1$, a Borel set $B \subset (0, 1)$, and $a \in [0, 1]$. A disclosed check is significant if its p -value lies in B . Given a report R , define the number of disclosed significant results with respect to B by

$$N_B(R) \equiv \sum_{p \in R} \mathbf{1}\{p \in B\}$$

Formally, [Appendix A.1.2](#) represents reports as counting measures r , in which case $N_B(r) = r(B)$. The robustness-check policy $\delta_{B,m,a}$ accepts if and only if the report contains at least m significant results:

$$\delta_{B,m,a}(R) \equiv a \mathbf{1}\{N_B(R) \geq m\}$$

Setting $a = 1$ recovers the main-text robustness-check rule $\delta_{B,m}$ from [Definition 5.1](#). We retain a because it is useful when implementing a throughput target by rationing among qualifying reports.

For analysis, define the associated value function for a general reward level $a \in [0, \infty)$ by

$$V_{m,B}(a) \equiv \sup_{\tau} \mathbb{E} \left[a \mathbf{1}\{S_\tau(B) \geq m\} - \gamma\tau \right] \quad S_n(B) \equiv \sum_{k=1}^n \mathbf{1}\{P_k \in B\}$$

When $a \in [0, 1]$, this coincides with the researcher's best-response value under the robustness-check policy $\delta_{B,m,a}$.

Definition A.18. Fix (m, B) . Define $Q_n(m, B) \equiv \{S_n(B) \geq m\}$ and $\tau_B^{(m)} \equiv \inf\{n \geq 1 : S_n(B) \geq m\}$.

A.3.2 Two levers: standards and disclosure

Robustness checks have two core design levers: the window B (standards) and the required count m (disclosure). For a cost-share parameter $c > 0$ we set

$$m(\gamma; c) \equiv \left\lceil \frac{c}{\gamma} \right\rceil$$

We interpret c as the fraction of the $1/\gamma$ testing-capacity scale that must be converted into disclosed significant results in B before qualification. The acceptance probability a can be used to ration among qualifiers when the journal imposes a binding throughput constraint (Appendix A.3.4).

Definition A.19. Fix a cost share $c > 0$, an acceptance probability $a \in (0, 1]$, and a Borel set $B \subset (0, 1)$. Let $m(\gamma; c) \equiv \lceil c/\gamma \rceil$ as in Section A.3.2. Define $k(\gamma; a) \equiv \lceil a/\gamma \rceil$, $n(\gamma; c, a) \equiv m(\gamma; c) + k(\gamma; a)$, and $r(\gamma; c, a) \equiv m(\gamma; c)/n(\gamma; c, a)$. Define the worst-case non-high one-test significance probability

$$p_0^{\max}(B) \equiv \sup_{t \in \mathcal{T}_0} \mathbb{P}_t(P_1 \in B)$$

We say (c, a, B) satisfies the separation condition if, for all sufficiently small γ ,

$$p_0^{\max}(B) < r(\gamma; c, a) \tag{7}$$

If there exist constants $\eta_0 > 0$ and $\gamma_0 > 0$ such that $p_0^{\max}(B) \leq r(\gamma; c, a) - \eta_0$ for all $\gamma \in (0, \gamma_0]$, we say separation holds *with uniform slack*.

A.3.3 Witness windows imply separation

Lemma A.20. *Maintain Assumption A.2. Then $p_0^{\max}(B_0) \leq e^{-\ell_0}$. Consequently, for any acceptance probability $a \in (0, 1]$ and any cost share*

$$c > \frac{e^{-\ell_0}}{1 - e^{-\ell_0}}$$

the triple (c, a, B_0) satisfies Definition A.19 with uniform slack for all sufficiently small γ .

Proof. Fix any non-high type $t \in \mathcal{T}_0$. Assumption A.2 gives $d\mathcal{L}_H(P_1)/d\mathcal{L}_t(P_1) \geq e^{\ell_0}$ on B_0 , $\mathcal{L}_t(P_1)$ -almost everywhere, hence

$$\mathbb{P}_H(P_1 \in B_0) \geq e^{\ell_0} \mathbb{P}_t(P_1 \in B_0)$$

Since $\mathbb{P}_H(P_1 \in B_0) \leq 1$, this implies $\mathbb{P}_t(P_1 \in B_0) \leq e^{-\ell_0}$ and therefore $p_0^{\max}(B_0) \leq e^{-\ell_0}$.

Pick $\eta_0 > 0$ such that $e^{-\ell_0} \leq c/(c+1) - 2\eta_0$. For any $a \in (0, 1]$, Definition A.19 uses $r(\gamma; c, a) = m(\gamma; c)/(m(\gamma; c) + k(\gamma; a))$ with $k(\gamma; a) = \lceil a/\gamma \rceil$. Using $m(\gamma; c) \geq c/\gamma$ and $k(\gamma; a) \leq a/\gamma + 1 \leq 1/\gamma + 1$, we have for $\gamma \in (0, 1)$

$$r(\gamma; c, a) \geq \frac{c/\gamma}{(c+1)/\gamma + 2} = \frac{c}{c+1+2\gamma}$$

Choose $\gamma_0 > 0$ such that $c/(c+1+2\gamma) \geq c/(c+1) - \eta_0$ for all $\gamma \in (0, \gamma_0]$. Then for all $\gamma \in (0, \gamma_0]$,

$$p_0^{\max}(B_0) \leq \frac{c}{c+1} - 2\eta_0 \leq r(\gamma; c, a) - \eta_0$$

which is uniform slack. □

A.3.4 Throughput implementation for robustness checks

Proposition A.21. *Maintain Assumption A.1 and fix a target throughput $\rho \in (0, 1)$ and a cost $\gamma > 0$. Fix an integer $m \geq 1$ and a Borel set $B \subset (0, 1)$. Assume there exists a (possibly mixed) stopping time τ attaining $V_{m,B}(1)$ such that $\mathbb{P}(S_\tau(B) \geq m) \geq \rho$.*

Then there exist $a^ \in [\rho, 1]$ and a (possibly mixed) best response to δ_{B,m,a^*} such that*

$$\mathbb{E}[\delta_{B,m,a^*}(R)] = \rho$$

Proof. Fix $\rho \in (0, 1)$ and $\gamma > 0$. Maintain Assumption A.1. Fix $m \geq 1$ and Borel $B \subset (0, 1)$. For $n \geq 0$ define the count of significant results (with respect to B) in the realized history by

$$S_n^B \equiv \sum_{k=1}^n \mathbf{1}\{P_k \in B\}$$

and define the qualification indicator

$$\bar{M}_n^{m,B} \equiv \mathbf{1}\{S_n^B \geq m\} \tag{8}$$

Fix $a \in [0, 1]$ and write $\delta \equiv \delta_{B,m,a}$. We claim that the envelope acceptance value satisfies

$$M_n^\delta \equiv \max_{R \preceq R_n^{\text{full}}} \delta(R) = a \bar{M}_n^{m,B} \tag{9}$$

Indeed, if $S_n^B < m$ then no feasible subreport can contain m significant results in B , so $\delta(R) = 0$ for all feasible R and $M_n^\delta = 0$. If $S_n^B \geq m$, then there exists a feasible subreport consisting of m significant results in B , for which $\delta(R) = a$, and since $\delta \leq a$ everywhere we obtain $M_n^\delta = a$. This proves (9).

In particular, $(\bar{M}_n^{m,B})$ and (M_n^δ) are adapted and take values in $[0, 1]$. Fix $a \in [0, 1]$. By Lemma A.15, the researcher's problem depends on δ only through (M_n^δ) . By (9), the stopping problem is

$$\sup_{\tau} \mathbb{E}[M_\tau^\delta - \gamma\tau] = \sup_{\tau} \mathbb{E}[a \bar{M}_\tau^{m,B} - \gamma\tau]$$

The reward process $a \bar{M}_n^{m,B} - \gamma n$ is bounded and adapted, so an optimal stopping time exists (Proposition A.16), and we allow time-0 mixing between optimal stopping times.

Define the value function $V_{m,B}$ as above:

$$V_{m,B}(a) \equiv \sup_{\tau} \mathbb{E}[a \bar{M}_\tau^{m,B} - \gamma\tau]$$

where $a \in [0, \infty)$ is a prize level and we allow time-0 randomization over stopping times.

For each (possibly mixed) stopping strategy τ , define

$$m(\tau) \equiv \mathbb{E}[\bar{M}_\tau^{m,B}] \in [0, 1] \quad c(\tau) \equiv \gamma \mathbb{E}[\tau] \in [0, \infty]$$

Then

$$V_{m,B}(a) = \sup_{\tau} (a m(\tau) - c(\tau))$$

so $V_{m,B} : [0, \infty) \rightarrow \mathbb{R}$ is convex as a pointwise supremum of affine functions of a (see, e.g., (Boyd and Vandenberghe, 2004, §3.2.3)). Hence for each $a > 0$ the subdifferential $\partial V_{m,B}(a)$ is a nonempty closed interval

$$\partial V_{m,B}(a) = [V'_-(a), V'_+(a)] \subseteq [0, 1]$$

We record the standard supporting-hyperplane inclusion: if τ is a -optimal then for all $a' \in [0, \infty)$,

$$V_{m,B}(a') \geq a' m(\tau) - c(\tau) = V_{m,B}(a) + m(\tau)(a' - a)$$

so $m(\tau) \in \partial V_{m,B}(a)$. Conversely, for pointwise suprema of affine functions, every $g \in \partial V_{m,B}(a)$ can be implemented by time-0 mixing between at most two a -optimal stopping times; see, e.g., (Rockafellar, 1970).

Under $\delta_{B,m,a}$ and a stopping rule τ , the induced acceptance probability equals

$$\mathbb{E}[\delta_{B,m,a}(R)] = \mathbb{E}[M_\tau^\delta] = a \cdot \mathbb{E}[\bar{M}_\tau^{m,B}] = a \cdot m(\tau)$$

In particular, under an a -optimal (possibly mixed) response, the set of achievable acceptance rates is

$$a \cdot \partial V_{m,B}(a) = \{a \cdot g : g \in \partial V_{m,B}(a)\} \quad (10)$$

Thus the throughput target $\mathbb{E}[\delta] = \rho$ is satisfiable at parameter a if and only if there exists $g \in \partial V_{m,B}(a)$ with $ag = \rho$, equivalently

$$\frac{\rho}{a} \in \partial V_{m,B}(a) \quad (11)$$

Note that any such a must satisfy $a \geq \rho$, since $\partial V_{m,B}(a) \subseteq [0, 1]$.

Define $h(a) \equiv \rho/a$ on $[\rho, 1]$. This function is continuous and strictly decreasing, with $h(\rho) = 1$ and $h(1) = \rho$. Define

$$g_+(a) \equiv \sup \partial V_{m,B}(a) = V'_+(a)$$

For convex $V_{m,B}$, the map $a \mapsto g_+(a)$ is nondecreasing, and $g_+(a) \leq 1$ for all a . By the proposition's assumption, there exists a 1-optimal rule with qualification probability at least ρ , hence $g_+(1) \geq \rho = h(1)$.

Define

$$a^* \equiv \inf\{a \in [\rho, 1] : g_+(a) \geq h(a)\}$$

This set is nonempty because $g_+(1) \geq h(1)$, and it is bounded below by ρ . By monotonicity of g_+ and continuity of h , we have

$$\inf \partial V_{m,B}(a^*) \leq h(a^*) \leq \sup \partial V_{m,B}(a^*)$$

which is exactly (11). Hence there exists $g^* \in \partial V_{m,B}(a^*)$ with $g^* = h(a^*)$.

By the subgradient implementation above, there exists an a^* -optimal (possibly mixed) stopping rule τ^* such that

$$m(\tau^*) = g^* = \frac{\rho}{a^*}$$

Therefore the induced acceptance rate satisfies

$$\mathbb{E}[\delta_{B,m,a^*}(R)] = a^* m(\tau^*) = a^* \frac{\rho}{a^*} = \rho$$

Define an associated best response strategy at a^* by using τ^* and specifying the report at τ^* as follows. If $\bar{M}_{\tau^*}^{m,B} = 1$, report any feasible subreport $R \preceq R_{\tau^*}^{\text{full}}$ with exactly m significant results in B (e.g. the first m such results in chronological order). If $\bar{M}_{\tau^*}^{m,B} = 0$, report \emptyset . Either choice attains $M_{\tau^*}^\delta$ by (9). Because the rule refers only to the finite set of indices $\{1, \dots, \tau^*\}$ and uses chronological order, it is measurable and avoids any global ordering of reports.

This completes the proof. \square

A.3.5 Proofs for Section 5

Lemma A.22. Fix (m, B, a) and let $\delta = \delta_{B,m,a}$. Let $\tau_B^{(m)}$ be the qualification time from Definition A.18. For any strategy $\sigma = (\tau, S)$, define $\tilde{\tau} \equiv \min\{\tau, \tau_B^{(m)}\}$ and define \tilde{S} as follows: on $\{\tilde{\tau} = \tau\}$ set $\tilde{S} = S$, and on $\{\tilde{\tau} = \tau_B^{(m)} < \tau\}$ choose any m indices $i \leq \tau_B^{(m)}$ such that $P_i \in B$. Then

$$\mathbb{E}[\delta(\tilde{R}) - \gamma\tilde{\tau}] \geq \mathbb{E}[\delta(R) - \gamma\tau]$$

In particular, the selected best response to $\delta_{B,m,a}$ can be taken to stop upon qualification, and its stopping time has the form

$$\tau^* = \tau_B^{(m)} \wedge \tau_{\text{quit}}$$

for some quitting time τ_{quit} . Hence for all $n \geq 1$,

$$\{\tau^* > n\} \subseteq Q_n(m, B)^c$$

Proof. On $\{\tau_B^{(m)} < \tau\}$, qualification holds at $\tau_B^{(m)}$ by definition, so the researcher can report a qualifying report and obtain acceptance probability a . Stopping earlier cannot reduce acceptance (the policy never exceeds a), and strictly lowers cost by $\gamma(\tau - \tau_B^{(m)}) > 0$. On $\{\tau \leq \tau_B^{(m)}\}$ nothing changes. Thus the modified strategy weakly improves payoff.

For the final claim, apply the improvement argument to the selected best response and define $\tau_{\text{quit}} \equiv \tau^*$ on $\{\tau^* < \tau_B^{(m)}\}$ and $\tau_{\text{quit}} \equiv \infty$ otherwise. If $\tau^* > n$ then necessarily $\tau_B^{(m)} > n$, hence $S_n(B) < m$ and $Q_n(m, B)^c$ holds. \square

Lemma A.23. Fix (m, B, a) with $a \in (0, 1]$ and let τ^* be the selected best response stopping time to $\delta_{B,m,a}$. Fix any $n \geq m$ and define

$$K \equiv \left\lceil \frac{a}{\gamma} \right\rceil \quad \text{so that} \quad \gamma \cdot K \geq a$$

Then on $\{\tau^* > n\}$,

$$\mathbb{P}(Q_{n+K}(m, B) \mid \mathcal{F}_n) \geq \frac{\gamma}{a}$$

In particular, the bound is nontrivial only when $\gamma \leq a$; if $\gamma > a$ then $\{\tau^* > n\}$ is empty. where $\mathbb{P}(\cdot \mid \mathcal{F}_n)$ is the researcher's subjective posterior predictive probability.

Proof. Work on the event $\{\tau^* > n\}$, which is \mathcal{F}_n -measurable. By Lemma A.22, $Q_n(m, B)^c$ holds on $\{\tau^* > n\}$, so stopping at n yields payoff 0.

Because τ^* is optimal, the conditional continuation value at time n cannot be negative on $\{\tau^* > n\}$: otherwise stopping at n on a subset where it is negative would strictly improve payoff. Thus on $\{\tau^* > n\}$,

$$0 \leq \mathbb{E}[a \mathbf{1}\{Q_{\tau^*}(m, B)\} - \gamma(\tau^* - n) \mid \mathcal{F}_n]$$

Since $Q_k(m, B)$ is increasing in k , Lemma A.14 gives

$$\mathbf{1}\{Q_{\tau^*}(m, B)\} \leq \mathbf{1}\{Q_{n+K}(m, B)\} + \mathbf{1}\{\tau^* > n + K\}$$

Also $\tau^* - n \geq 1 + K \mathbf{1}\{\tau^* > n + K\}$ pointwise on $\{\tau^* > n\}$. Taking conditional expectations and combining yields

$$0 \leq a \mathbb{P}(Q_{n+K}(m, B) \mid \mathcal{F}_n) - \gamma + (a - \gamma K) \mathbb{P}(\tau^* > n + K \mid \mathcal{F}_n)$$

Since $a - \gamma K \leq 0$, we obtain

$$0 \leq a \mathbb{P}(Q_{n+K}(m, B) \mid \mathcal{F}_n) - \gamma$$

which is the claimed inequality. \square

Lemma A.24. Fix (m, B, a) with $a \in (0, 1]$ and let τ^* be the selected best response stopping time to $\delta_{B,m,a}$. Let

$$\pi_n(H) \equiv \mathbb{P}(T \in \mathcal{H} \mid \mathcal{F}_n)$$

denote the posterior probability that the submission is high-impact at time n . Fix any $n \geq m$ and define $K = \lceil a/\gamma \rceil$ as in Lemma A.23. Define the “ K -step non-high opportunity” event and its non-high conditional probability by

$$\Delta_{n,K} \equiv Q_{n+K}(m, B) \setminus Q_n(m, B) \quad q_{n,K}^0 \equiv \mathbb{P}_0(\Delta_{n,K} \mid \mathcal{F}_n)$$

Then on $\{\tau^* > n\}$,

$$\pi_n(H) + q_{n,K}^0 \geq \frac{\gamma}{a}$$

Consequently,

$$\{\tau^* > n\} \subseteq \left\{ \pi_n(H) \geq \frac{\gamma}{2a} \right\} \cup \left\{ q_{n,K}^0 \geq \frac{\gamma}{2a} \right\}$$

Proof. On the event $\{\tau^* > n\}$ we have $Q_n(m, B)^c$ by Lemma A.22. Lemma A.23 gives

$$\mathbb{P}(Q_{n+K}(m, B) \mid \mathcal{F}_n) \geq \frac{\gamma}{a} \quad \text{on } \{\tau^* > n\}$$

On $Q_n(m, B)^c$, the event $Q_{n+K}(m, B)$ is equivalent to $\Delta_{n,K}$, so it suffices to bound $\mathbb{P}(\Delta_{n,K} \mid \mathcal{F}_n)$.

Decompose the posterior predictive probability by the coarse type class:

$$\mathbb{P}(\Delta_{n,K} \mid \mathcal{F}_n) = \pi_n(H)\mathbb{P}(\Delta_{n,K} \mid \mathcal{F}_n, T \in \mathcal{H}) + (1 - \pi_n(H))\mathbb{P}(\Delta_{n,K} \mid \mathcal{F}_n, T \in \mathcal{T}_0)$$

The first conditional probability is at most 1. For the second, by definition of $\mathbb{P}_0(\cdot) = \mathbb{P}(\cdot \mid T \in \mathcal{T}_0)$ we have $\mathbb{P}(\Delta_{n,K} \mid \mathcal{F}_n, T \in \mathcal{T}_0) = \mathbb{P}_0(\Delta_{n,K} \mid \mathcal{F}_n) = q_{n,K}^0$. Therefore,

$$\mathbb{P}(\Delta_{n,K} \mid \mathcal{F}_n) \leq \pi_n(H) + q_{n,K}^0$$

Combining yields $\pi_n(H) + q_{n,K}^0 \geq \gamma/a$ on $\{\tau^* > n\}$, and the union bound is immediate. \square

Lemma A.25. Let (X_n) be a stationary reversible Markov chain with L^2 spectral gap $\text{gap} \in (0, 1]$. Let $f : \mathsf{X} \rightarrow [0, 1]$ be measurable and define

$$S_n \equiv \sum_{k=1}^n (f(X_k) - \mathbb{E}[f(X_1)])$$

Then there exists a universal constant $C_H < \infty$ such that for all $n \geq 1$ and all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}S_n \geq \varepsilon\right) \leq \exp(-C_H \text{gap } n \varepsilon^2) \quad \mathbb{P}\left(\frac{1}{n}S_n \leq -\varepsilon\right) \leq \exp(-C_H \text{gap } n \varepsilon^2)$$

Proof. See Lézaud (1998) or Paulin (2015). We use only the existence of a universal constant C_H and the dependence of the exponent on the spectral gap. \square

Proposition A.26. Maintain Assumption A.1 and Definition A.5. Fix a type $t \in \mathcal{T}$ and a Borel evidence window $B \subset (0, 1)$. Define the one-test significance probability

$$p_t(B) \equiv \mathbb{P}_t(P_1 \in B) \in [0, 1]$$

Let $Y_n \equiv \mathbf{1}\{P_n \in B\}$ and $S_n(B) \equiv \sum_{k=1}^n Y_k$. Then there exists a universal constant $C < \infty$ such that for all $n \geq 1$ and all $\varepsilon \in (0, 1)$,

$$\begin{aligned} \mathbb{P}_t\left(\frac{1}{n}S_n(B) \geq p_t(B) + \varepsilon\right) &\leq \exp(-C n_{\text{eff}}(\theta) n \varepsilon^2) \\ \mathbb{P}_t\left(\frac{1}{n}S_n(B) \leq p_t(B) - \varepsilon\right) &\leq \exp(-C n_{\text{eff}}(\theta) n \varepsilon^2) \end{aligned}$$

Proof. This is immediate from Assumption A.1(iii) with $C = C_H$. \square

Lemma A.27. *Maintain Assumption A.1 and Definition A.18. Fix a type $t \in \mathcal{T}$ and a Borel evidence window $B \subset (0, 1)$ with $p_t(B) > 0$. Fix any $\delta \in (0, 1)$ and define*

$$n_m \equiv \left\lceil \frac{1 + \delta}{p_t(B)} m \right\rceil$$

Then there exists a constant $C_{\delta,t,B} < \infty$ such that for all integers $m \geq 1$,

$$\mathbb{E}_t[\tau_B^{(m)}] \leq n_m + C_{\delta,t,B}$$

In particular, for each fixed $\delta \in (0, 1)$,

$$\mathbb{E}_t[\tau_B^{(m)}] \leq \frac{1 + \delta}{p_t(B)} m + O_{\delta,t,B}(1)$$

so the expected qualification time grows at most linearly in m with slope arbitrarily close to $1/p_t(B)$. Moreover, if $\inf_{t \in \mathcal{H}} p_t(B) > 0$, then for every $\delta \in (0, 1)$ there exists $C_{\delta,B} < \infty$ such that

$$\sup_{t \in \mathcal{H}} \mathbb{E}_t[\tau_B^{(m)}] \leq \left\lceil \frac{1 + \delta}{\inf_{t \in \mathcal{H}} p_t(B)} m \right\rceil + C_{\delta,B} \quad \text{for all } m \geq 1$$

Proof. Fix t, B , and $\delta \in (0, 1)$ and write $p \equiv p_t(B) > 0$. Recall that $\{\tau_B^{(m)} > n\} = \{S_n(B) < m\}$. By the tail-sum formula,

$$\mathbb{E}_t[\tau_B^{(m)}] = \sum_{n \geq 0} \mathbb{P}_t(\tau_B^{(m)} > n) = \sum_{n \geq 0} \mathbb{P}_t(S_n(B) < m)$$

Split the sum at n_m :

$$\mathbb{E}_t[\tau_B^{(m)}] \leq n_m + \sum_{n \geq n_m} \mathbb{P}_t(S_n(B) < m)$$

For $n \geq n_m$, we have $m/n \leq p/(1 + \delta)$, so

$$p - \frac{m}{n} \geq p - \frac{p}{1 + \delta} = \frac{\delta}{1 + \delta} p$$

Therefore,

$$\{S_n(B) < m\} \subseteq \left\{ \frac{1}{n} S_n(B) \leq p - \frac{\delta}{1 + \delta} p \right\}$$

Apply Proposition A.26 with $\varepsilon \equiv \frac{\delta}{1 + \delta} p$ to obtain

$$\mathbb{P}_t(S_n(B) < m) \leq \exp(-C n_{\text{eff}}(\theta) n \varepsilon^2) \quad \text{for all } n \geq n_m$$

for the universal constant C from Proposition A.26. The tail sum is therefore bounded by a finite geometric series:

$$\sum_{n \geq n_m} \mathbb{P}_t(S_n(B) < m) \leq \sum_{n \geq n_m} \exp(-c_{\delta,t,B} n) \leq \frac{\exp(-c_{\delta,t,B} n_m)}{1 - \exp(-c_{\delta,t,B})} \leq C_{\delta,t,B} \equiv \frac{1}{1 - \exp(-c_{\delta,t,B})}$$

where $c_{\delta,t,B} \equiv C n_{\text{eff}}(\theta) \left(\frac{\delta}{1 + \delta} p\right)^2 > 0$ and $C_{\delta,t,B} < \infty$ depends only on (δ, t, B) and the standing primitives. This proves the first claim. The uniform bound over $t \in \mathcal{H}$ follows by replacing $p_t(B)$ with $\inf_{t \in \mathcal{H}} p_t(B)$. \square

Theorem A.28. Fix $c > 0$, $a \in (0, 1]$, and a Borel evidence window $B \subset (0, 1)$. Let $m = m(\gamma; c)$ as in Section A.3.2 and consider the robustness-check rule $\delta_{B,m,a}$. Define

$$K \equiv \left\lceil \frac{a}{\gamma} \right\rceil \quad n \equiv m + K \quad r(\gamma; c, a) \equiv \frac{m}{n}$$

Let τ^* be the selected best response stopping time. Define the worst-case non-high one-test significance probability

$$p_0^{\max}(B) \equiv \sup_{t \in \mathcal{T}_0} \mathbb{P}_t(P_1 \in B) \in [0, 1]$$

Assume the separation condition $p_0^{\max}(B) < r(\gamma; c, a)$ (Definition A.19). Then there exist constants $C_1, c_2 > 0$ depending only on primitives and (c, a, B) such that:

(i) The non-high qualification probability satisfies

$$\mathbb{P}_0(Q_{\tau^*}(m, B)) \leq \mathbb{P}_0\left(\pi_m(H) > \frac{\gamma}{2a}\right) + C_1 \left(1 + \frac{a}{\gamma}\right) \exp\left(-c_2 n_{\text{eff}}(\theta) n \left(r(\gamma; c, a) - p_0^{\max}(B)\right)^2\right)$$

(ii) The induced non-high acceptance probability satisfies

$$q_0(\delta_{B,m,a}) = a \mathbb{P}_0(Q_{\tau^*}(m, B))$$

and hence $q_0(\delta_{B,m,a})$ is bounded by a times the right-hand side in part (i).

Proof. We bound the non-high qualification probability in three steps. First, we reduce the random horizon to a deterministic one. Since $Q_k(m, B)$ is increasing in k , Lemma A.14 gives

$$\mathbb{P}_0(Q_{\tau^*}(m, B)) \leq \mathbb{P}_0(Q_m(m, B)) + \mathbb{P}_0(\tau^* > m)$$

Next, we bound $\mathbb{P}_0(\tau^* > m)$ using the K -step lemma. Set $n = m + K$. Lemma A.24 (with $n = m$) implies

$$\mathbb{P}_0(\tau^* > m) \leq \mathbb{P}_0\left(\pi_m(H) > \frac{\gamma}{2a}\right) + \mathbb{P}_0\left(q_{m,K}^0 > \frac{\gamma}{2a}\right)$$

For the opportunity term, $q_{m,K}^0 = \mathbb{P}_0(\Delta_{m,K} | \mathcal{F}_m)$, so $\mathbb{E}_0[q_{m,K}^0] = \mathbb{P}_0(\Delta_{m,K})$, and Markov's inequality gives

$$\mathbb{P}_0\left(q_{m,K}^0 > \frac{\gamma}{2a}\right) \leq \frac{2a}{\gamma} \mathbb{P}_0(\Delta_{m,K}) \leq \frac{2a}{\gamma} \mathbb{P}_0(Q_{m+K}(m, B)),$$

where the final inequality uses $\Delta_{m,K} \subseteq Q_{m+K}(m, B)$. Since $Q_m(m, B) \subseteq Q_{m+K}(m, B)$, combining yields

$$\mathbb{P}_0(Q_{\tau^*}(m, B)) \leq \mathbb{P}_0\left(\pi_m(H) > \frac{\gamma}{2a}\right) + \left(1 + \frac{2a}{\gamma}\right) \mathbb{P}_0(Q_{m+K}(m, B)).$$

Finally, we apply concentration at the deterministic horizon $n = m + K$. Set

$$\varepsilon \equiv \frac{m}{m+K} - p_0^{\max}(B) = r(\gamma; c, a) - p_0^{\max}(B) > 0,$$

where positivity is the separation condition. Fix any non-high type $t \in \mathcal{T}_0$ and write $p_t(B) \equiv \mathbb{P}_t(P_1 \in B) \leq p_0^{\max}(B)$. Since $Q_n(m, B) = \{S_n(B) \geq m\}$, we have

$$\mathbb{P}_t(Q_n(m, B)) = \mathbb{P}_t\left(\frac{1}{n}S_n(B) \geq \frac{m}{n}\right) \leq \mathbb{P}_t\left(\frac{1}{n}S_n(B) \geq p_t(B) + \varepsilon\right).$$

Applying Proposition A.26 yields

$$\mathbb{P}_t(Q_n(m, B)) \leq \exp(-C n_{\text{eff}}(\theta) n \varepsilon^2)$$

for a universal constant $C > 0$. Since \mathbb{P}_0 is a mixture over non-high types, the same bound holds for $\mathbb{P}_0(Q_n(m, B))$. Absorb constants into C_1, c_2 to obtain (i).

Part (ii) is immediate from Definition A.17. \square

Theorem 5.2 (restated). Maintain the standing assumptions (Appendix A.1.1), Assumption A.8, and suppose $n_{\text{eff}}(\theta) > 0$. Let B_0 be the witness window from Assumption A.2. Fix c satisfying $e^{-\ell_0}/(1 - e^{-\ell_0}) < c < p_H(B_0)$ and set $m(\gamma) = \lceil c/\gamma \rceil$. Then there exist constants $c_0 > 0$ and $c_H \in (0, 1]$ such that for all sufficiently small γ ,

$$q_0(\delta_{B_0, m(\gamma)}) \leq \exp(-c_0 \kappa_{\text{eff}}) \quad q_H(\delta_{B_0, m(\gamma)}) \geq c_H$$

so $\text{FDR}(\delta_{B_0, m(\gamma)}) = \exp(-\Omega(\kappa_{\text{eff}}))$ and throughput is nonvanishing.

Proof. We prove the bound for the more general rationed rule $\delta_{B, m, a}$ from Definition A.17; the main-text policy is the special case $a = 1$.

Fix $a \in (0, 1]$ and a constant c such that separation holds with uniform slack and the design is feasible under high types:

$$c > \frac{e^{-\ell_0}}{1 - e^{-\ell_0}} \quad \text{and} \quad c < a p_H(B_0)$$

where $(\ell_0, p_H(B_0))$ are from Assumption A.2. Write $m = m(\gamma) = \lceil c/\gamma \rceil$ and $B = B_0$. Let $\delta = \delta_{B, m, a}$ and let τ^* be the selected best-response stopping time.

Theorem A.28(i) yields

$$\mathbb{P}_0(Q_{\tau^*}(m, B)) \leq \mathbb{P}_0\left(\pi_m(H) > \frac{\gamma}{2a}\right) + C_1 \left(1 + \frac{a}{\gamma}\right) \exp\left(-c_2 n_{\text{eff}}(\theta) n \left(r(\gamma; c, a) - p_0^{\max}(B)\right)^2\right) \quad (12)$$

where $n = m + \lceil a/\gamma \rceil$ and $r(\gamma; c, a) = m/n$.

Suppose separation holds with uniform slack (Definition A.19). For the witness window B_0 , Lemma A.20 implies this whenever $c > e^{-\ell_0}/(1 - e^{-\ell_0})$ (and any $a \in (0, 1]$). Under this slack, $r(\gamma; c, a) - p_0^{\max}(B) \geq \eta_0 > 0$ for all sufficiently small γ , so the exponential term in (12) is $\exp(-\Omega(n_{\text{eff}}(\theta) n)) = \exp(-\Omega(\kappa_{\text{eff}}))$. Moreover, $\log(1 + a/\gamma) = O(\log(1/\gamma)) = o(\kappa_{\text{eff}})$, so the prefactor $(1 + a/\gamma)$ can be absorbed into the exponent. Thus (12) gives

$$\mathbb{P}_0(Q_{\tau^*}(m, B)) \leq \mathbb{P}_0\left(\pi_m(H) > \frac{\gamma}{2a}\right) + \exp(-\Omega(\kappa_{\text{eff}}))$$

Since $q_0(\delta) = a \mathbb{P}_0(Q_{\tau^*}(m, B))$, we obtain

$$q_0(\delta) \leq a \mathbb{P}_0\left(\pi_m(H) > \frac{\gamma}{2a}\right) + \exp(-\Omega(\kappa_{\text{eff}}))$$

Assumption A.8 bounds the posterior tail term with $n = m$ and $\varepsilon = \gamma/(2a)$:

$$\mathbb{P}_0\left(\pi_m(H) > \frac{\gamma}{2a}\right) \leq C_{\pi, \alpha} \left(\frac{2a}{\gamma}\right)^{(\alpha-1)/\alpha} \exp\left(-\frac{\alpha-1}{\alpha} C_\alpha(\theta) n_{\text{eff}}(\theta) (m-1)\right)$$

Since $m = \lceil c/\gamma \rceil$ we have $n_{\text{eff}}(\theta) (m-1) = \Theta(\kappa_{\text{eff}})$, while $\log(2a/\gamma) = O(\log(1/\gamma)) = o(\kappa_{\text{eff}})$, so the right-hand side is $\exp(-\Omega(\kappa_{\text{eff}}))$. Therefore $q_0(\delta) \leq a \exp(-c_0 \kappa_{\text{eff}})$ for some $c_0 > 0$ and all

sufficiently small γ . Appendix A.6 shows that in the ordered Gaussian AR(1) example the posterior tail term is itself $\exp(-\Omega(\kappa_{\text{eff}}))$ (Lemma A.49), which is sufficient for the robustness-check bound.

By Definition A.17, the best-response value satisfies

$$V_{m,B}(a) = \mathbb{E}[\delta(R) - \gamma\tau^*] \leq \mathbb{E}[\delta(R)] = \rho(\delta) = \pi_H q_H(\delta) + \pi_0 q_0(\delta)$$

so $\pi_H q_H(\delta) \geq V_{m,B}(a) - \pi_0 q_0(\delta)$.

To lower bound $V_{m,B}(a)$, fix $\eta \in (0, 1)$ such that $c < (1 - \eta)a p_H(B_0)$. Let $p_0^{\max} \equiv p_0^{\max}(B_0)$ and set $r_1 \equiv (p_H(B_0) + p_0^{\max})/2 \in (0, 1)$. Choose $n_1 = n_1(\gamma) \equiv \lceil L \log(1/\gamma) \rceil$ for a large constant $L > 0$, and define the stage-1 event

$$E_1 \equiv \left\{ \frac{1}{n_1} S_{n_1}(B_0) \geq r_1 \right\}$$

By Proposition A.26, $\mathbb{P}(E_1 | T \in \mathcal{H}) \rightarrow 1$ and $\mathbb{P}(E_1 | T \in \mathcal{T}_0) \rightarrow 0$, and moreover $\gamma n_1 = o(1)$.

Next define a stage-2 horizon

$$n_2 = n_2(\gamma) \equiv \left\lceil \frac{1 + \eta}{p_H(B_0)} m(\gamma) \right\rceil$$

so $\gamma n_2 \leq (1 + \eta)c/p_H(B_0) + o(1) < (1 - \eta^2)a$ for all sufficiently small γ . Consider the feasible strategy that stops at time n_1 if E_1 fails, and otherwise continues to time n_2 and then stops, disclosing all significant results in B_0 . Under $T \in \mathcal{H}$, Proposition A.26 also gives $\mathbb{P}(S_{n_2}(B_0) \geq m | T \in \mathcal{H}) \rightarrow 1$ because $m/n_2 \leq p_H(B_0)/(1 + \eta) < p_H(B_0)$. Therefore this strategy yields (for all sufficiently small γ)

$$V_{m,B}(a) \geq a \mathbb{P}(E_1 \cap \{S_{n_2}(B_0) \geq m\}) - \gamma n_1 - \gamma(n_2 - n_1) \mathbb{P}(E_1) \geq \pi_H \cdot \frac{\eta^2 a}{2}$$

after using $\mathbb{P}(E_1 \cap \{S_{n_2}(B_0) \geq m\}) \geq \pi_H(1 - o(1))$, $\mathbb{P}(E_1) = \pi_H + o(1)$, $\gamma n_1 = o(1)$, and $\gamma n_2 \leq (1 - \eta^2)a$ for all sufficiently small γ .

Combining with $q_0(\delta) = \exp(-\Omega(\kappa_{\text{eff}}))$ gives, for all sufficiently small γ ,

$$q_H(\delta) \geq \frac{\eta^2 a}{4}$$

Finally, for the false discovery rate, since $\rho(\delta) \geq \pi_H q_H(\delta) \geq \pi_H \eta^2 a/4$, we have

$$\text{FDR}(\delta) = \frac{\pi_0 q_0(\delta)}{\rho(\delta)} \leq \frac{4\pi_0}{\pi_H \eta^2 a} \exp(-c_0 \kappa_{\text{eff}}) = \exp(-\Omega(\kappa_{\text{eff}}))$$

□

A.4 Information-theoretic lower bound

This appendix collects auxiliary inequalities and proofs for the mechanism-independent frontier bounds in Section 4, including Theorems 4.3 and 4.4. We first record auxiliary inequalities, then establish a stopped-process KL budget under Assumption A.1, and finally assemble the main lower bound and its slice corollaries. The KL budget appeals to the standing assumptions in Appendix A.1.1 and the universal effort bound in Lemma A.13 (Appendix A.2).

For a policy δ , let $q_H(\delta) \equiv \mathbb{P}(A = 1 | T \in \mathcal{H})$ and $q_0(\delta) \equiv \mathbb{P}(A = 1 | T \in \mathcal{T}_0)$ denote the induced acceptance probabilities under high and non-high types. Write $\rho(\delta) = \pi_H q_H(\delta) + \pi_0 q_0(\delta)$ and $\text{FDR}(\delta) = \pi_0 q_0(\delta)/\rho(\delta)$, and let

$$K(\delta) \equiv \text{KL}(\text{Bern}(q_H(\delta)) \parallel \text{Bern}(q_0(\delta)))$$

A.4.1 Auxiliary inequalities

Fix $\alpha > 1$. For probability measures $P \ll Q$ on a measurable space, we use the order- α Rényi divergence

$$D_\alpha(P\|Q) \equiv \frac{1}{\alpha - 1} \log \mathbb{E}_Q \left[\left(\frac{dP}{dQ} \right)^\alpha \right].$$

Lemma A.29. *Let $P \ll Q$ be probability measures on a measurable space and fix $\alpha > 1$. Write $L \equiv dP/dQ$. Then for every $x > 0$,*

$$Q(L \geq x) \leq x^{-\alpha} \exp((\alpha - 1)D_\alpha(P\|Q))$$

Proof. By Markov's inequality,

$$Q(L \geq x) = Q(L^\alpha \geq x^\alpha) \leq \mathbb{E}_Q[L^\alpha] x^{-\alpha}$$

By definition of Rényi divergence, $\mathbb{E}_Q[L^\alpha] = \exp((\alpha - 1)D_\alpha(P\|Q))$. □

Lemma A.30. *Let $P \ll Q$ be probability measures on a measurable space and fix $\alpha > 1$. Assume $D_\alpha(P\|Q) < \infty$. Then for every measurable function f with values in $[0, 1]$,*

$$\mathbb{E}_Q[f] \geq \mathbb{E}_P[f]^{\alpha/(\alpha-1)} \exp(-D_\alpha(P\|Q))$$

Proof. Let $L \equiv dP/dQ$. Then $\mathbb{E}_P[f] = \mathbb{E}_Q[Lf]$. Hölder's inequality with conjugate exponents α and $\alpha/(\alpha - 1)$ gives

$$\mathbb{E}_Q[Lf] \leq \mathbb{E}_Q[L^\alpha]^{1/\alpha} \mathbb{E}_Q[f^{\alpha/(\alpha-1)}]^{(\alpha-1)/\alpha}$$

Since $f \in [0, 1]$, we have $f^{\alpha/(\alpha-1)} \leq f$, so

$$\mathbb{E}_P[f] \leq \mathbb{E}_Q[L^\alpha]^{1/\alpha} \mathbb{E}_Q[f]^{(\alpha-1)/\alpha}$$

Rearranging yields

$$\mathbb{E}_Q[f] \geq \mathbb{E}_P[f]^{\alpha/(\alpha-1)} \mathbb{E}_Q[L^\alpha]^{-1/(\alpha-1)}$$

By definition of Rényi divergence, $\mathbb{E}_Q[L^\alpha] = \exp((\alpha - 1)D_\alpha(P\|Q))$. Substituting proves the claim. □

Lemma A.31. *For $p \in (0, 1]$, define $C(p) \equiv -(1 - p) \log(1 - p)$. Then for all $p \in (0, 1]$ and all $q \in (0, 1)$,*

$$\text{KL}(p\|q) \geq p \log \frac{p}{q} - C(p)$$

Consequently, if $\text{KL}(p\|q) \leq K$ then

$$q \geq p \exp\left(-\frac{K + C(p)}{p}\right)$$

Proof. For $p \in (0, 1]$ and $q \in (0, 1)$,

$$\text{KL}(p\|q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}$$

Since $1 - q \leq 1$

$$\log \frac{1 - p}{1 - q} \geq \log(1 - p)$$

Thus

$$\text{KL}(p||q) \geq p \log \frac{p}{q} + (1-p) \log(1-p)$$

Rearranging gives $\text{KL}(p||q) \geq p \log(p/q) - C(p)$.

For the inversion, if $\text{KL}(p||q) \leq K$ then

$$p \log \frac{p}{q} \leq K + C(p)$$

Hence

$$\log \frac{p}{q} \leq \frac{K + C(p)}{p}$$

Exponentiating yields $q \geq p \exp(-(K + C(p))/p)$. □

Corollary A.32. Fix (γ, θ) and any policy δ and write $q_H \equiv q_H(\delta)$ and $q_0 \equiv q_0(\delta)$. For $q_H \in (0, 1]$, define $C(q_H) \equiv -(1 - q_H) \log(1 - q_H)$. Then

$$q_0 \geq q_H \exp\left(-\frac{C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^*(\delta)] + C(q_H)}{q_H}\right)$$

Proof. Combine Theorem 4.3 with Lemma A.31. □

A.4.2 Fixed-throughput and fixed-FDR slices

Define the false discovery rate

$$\text{FDR}(\delta) \equiv \mathbb{P}(T \in \mathcal{T}_0 \mid A = 1) = \frac{\pi_0 q_0(\delta)}{\rho(\delta)}$$

with the convention that $\text{FDR}(\delta) = 0$ when $\rho(\delta) = 0$. For $\varepsilon \in (0, 1)$ define the slope

$$\eta(\varepsilon) \equiv \frac{\pi_H}{\pi_0} \cdot \frac{\varepsilon}{1 - \varepsilon}$$

and the associated log-stringency scalar

$$L(\varepsilon) \equiv \log \frac{1}{\eta(\varepsilon)} = \log \left(\frac{\pi_0(1 - \varepsilon)}{\pi_H \varepsilon} \right)$$

For the dual fixed-FDR view, define the maximal recall on an FDR slice:

$$G_{\gamma, \theta}(\varepsilon) \equiv \sup \{q_H(\delta) : \delta \text{ measurable and } \text{FDR}(\delta) \leq \varepsilon\}$$

Lemma A.33. Fix $\varepsilon \in (0, 1)$. For any policy δ ,

$$\text{FDR}(\delta) \leq \varepsilon \quad \iff \quad q_0(\delta) \leq \eta(\varepsilon) q_H(\delta)$$

Proof. If $\rho(\delta) = 0$ then $q_H(\delta) = q_0(\delta) = 0$, so both sides hold. Otherwise the equivalence is immediate from $\text{FDR} = \pi_0 q_0 / (\pi_H q_H + \pi_0 q_0)$ by rearranging. □

Corollary A.34. Fix a target throughput $\bar{\rho} \in (0, \pi_H]$. There exist constants $c, C > 0$ depending only on $(\bar{\rho}, \pi_H, \pi_0)$ and standing constants such that for every policy δ with $\rho(\delta) = \bar{\rho}$,

$$\text{FDR}(\delta) \geq c \exp(-C \cdot \kappa_{\text{KL}}) \quad \text{equivalently} \quad q_0(\delta) \geq \frac{\bar{\rho}}{\pi_0} c \exp(-C \cdot \kappa_{\text{KL}})$$

Proof. Write $q_H \equiv q_H(\delta)$ and $q_0 \equiv q_0(\delta)$. Under $\rho(\delta) = \bar{\rho}$ we have $\bar{\rho} = \pi_H q_H + \pi_0 q_0$ and $\text{FDR}(\delta) = \pi_0 q_0 / \bar{\rho}$.

If $q_H < \bar{\rho} / (2\pi_H)$ then $\pi_0 q_0 > \bar{\rho} / 2$ and $\text{FDR}(\delta) > 1/2$, so the claim holds after adjusting constants. Otherwise $q_H \geq \bar{\rho} / (2\pi_H)$.

By Lemma A.31 and $C(q_H) \leq 1/e$,

$$\text{KL}(\text{Bern}(q_H) \parallel \text{Bern}(q_0)) \geq q_H \log \frac{q_H}{q_0} - \frac{1}{e}$$

Combine with Theorem 4.3 to obtain

$$q_H \log \frac{q_H}{q_0} \leq C_\nu + \frac{\kappa_{\text{KL}}}{\pi_H} + \frac{1}{e}$$

Since $q_H \geq \bar{\rho} / (2\pi_H)$, this implies $\log(q_H/q_0) \leq C_0 + C_1 \kappa_{\text{KL}}$ for constants $C_0, C_1 > 0$ depending only on $(\bar{\rho}, \pi_H)$ and standing constants. Therefore

$$q_0 \geq q_H \exp(-C_0 - C_1 \kappa_{\text{KL}}) \geq \frac{\bar{\rho}}{2\pi_H} \exp(-C_0 - C_1 \kappa_{\text{KL}})$$

and hence

$$\text{FDR}(\delta) = \frac{\pi_0 q_0}{\bar{\rho}} \geq \frac{\pi_0}{2\pi_H} \exp(-C_0 - C_1 \kappa_{\text{KL}})$$

which is the desired $\exp(-\Theta(\kappa_{\text{KL}}))$ lower bound after relabeling constants. \square

Corollary A.35. Fix $\varepsilon \in (0, 1)$ with $L(\varepsilon) > 0$. For every policy δ with $\text{FDR}(\delta) \leq \varepsilon$,

$$q_H(\delta) L(\varepsilon) \leq C_\nu + \frac{\kappa_{\text{KL}}}{\pi_H} + \frac{1}{e} \quad \rho(\delta) \leq \frac{\pi_H}{1 - \varepsilon} \cdot \frac{C_\nu + \kappa_{\text{KL}}/\pi_H + 1/e}{L(\varepsilon)}$$

Consequently,

$$G_{\gamma, \theta}(\varepsilon) \leq 1 \wedge \frac{C_\nu + \kappa_{\text{KL}}/\pi_H + 1/e}{L(\varepsilon)}$$

Proof. Fix any policy δ with $\text{FDR}(\delta) \leq \varepsilon$. By Lemma A.33, $q_0 \leq \eta(\varepsilon) q_H$. If $q_H = 0$ then $\rho(\delta) = 0$ and the claim is trivial, so assume $q_H \in (0, 1]$. Then $\log(q_H/q_0) \geq \log(1/\eta(\varepsilon)) = L(\varepsilon)$.

By Lemma A.31 and $C(q_H) \leq 1/e$,

$$\text{KL}(\text{Bern}(q_H) \parallel \text{Bern}(q_0)) \geq q_H \log \frac{q_H}{q_0} - \frac{1}{e} \geq q_H L(\varepsilon) - \frac{1}{e}$$

Combine with Theorem 4.3 to obtain

$$q_H(\delta) \cdot L(\varepsilon) \leq C_\nu + \frac{\kappa_{\text{KL}}}{\pi_H} + \frac{1}{e}$$

Taking the supremum over policies yields the bound on $G_{\gamma, \theta}(\varepsilon)$. Finally, $\text{FDR}(\delta) \leq \varepsilon$ implies $\pi_0 q_0 \leq \varepsilon \rho$, hence $\pi_H q_H = \rho - \pi_0 q_0 \geq (1 - \varepsilon)\rho$ and $\rho \leq \pi_H q_H / (1 - \varepsilon)$. \square

A.4.3 Proof of Theorem 4.4

Theorem 4.4 (restated). Fix θ and consider $\gamma \downarrow 0$. Let δ^γ be any sequence of policies with selected best-response stopping times τ^γ and induced operating points (q_H^γ, q_0^γ) . Assume $\liminf_{\gamma \downarrow 0} q_H^\gamma \geq c_H > 0$. If $\mathbb{E}_H[\tau^\gamma] = o(1/\gamma)$ then $K(\delta^\gamma) = o(\kappa_{\text{KL}})$ and hence $-\log q_0^\gamma = o(\kappa_{\text{KL}})$ (equivalently $q_0^\gamma = \exp(-o(\kappa_{\text{KL}}))$), implying $\text{FDR}(\delta^\gamma) \geq \exp(-o(\kappa_{\text{KL}}))$ on any throughput floor $\rho(\delta^\gamma) \geq \bar{\rho} > 0$. If instead $\mathbb{E}_H[\tau^\gamma] = O(1)$ then q_0^γ is bounded away from 0 for all sufficiently small γ (and hence $\text{FDR}(\delta^\gamma)$ is bounded away from 0 as well).

Proof. Fix θ and a sequence $\gamma \downarrow 0$. Let δ^γ be any sequence of policies with induced operating points (q_H^γ, q_0^γ) under the selected best response τ^γ . Assume $\liminf_{\gamma \downarrow 0} q_H^\gamma \geq c_H > 0$.

By the universal KL budget (Theorem 4.3),

$$K(\delta^\gamma) = \text{KL}(\text{Bern}(q_H^\gamma) \parallel \text{Bern}(q_0^\gamma)) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^\gamma]$$

By Lemma A.31 and $C(q_H^\gamma) \leq 1/e$,

$$K(\delta^\gamma) \geq q_H^\gamma \log \frac{q_H^\gamma}{q_0^\gamma} - \frac{1}{e}$$

Therefore,

$$\log \frac{q_H^\gamma}{q_0^\gamma} \leq \frac{K(\delta^\gamma) + 1/e}{q_H^\gamma}$$

By the recall assumption, for all sufficiently small γ we have $q_H^\gamma \geq c_H/2$, so

$$-\log q_0^\gamma \leq \log \frac{1}{q_H^\gamma} + \frac{K(\delta^\gamma) + 1/e}{q_H^\gamma} \leq O(1) + O(K(\delta^\gamma))$$

If $\mathbb{E}_H[\tau^\gamma] = o(1/\gamma)$, then $D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^\gamma] = o(D_{\text{mix}}(\theta)/\gamma) = o(\kappa_{\text{KL}})$ and hence $K(\delta^\gamma) = C_\nu + o(\kappa_{\text{KL}}) = o(\kappa_{\text{KL}})$. Substituting into the bound above yields $-\log q_0^\gamma = o(\kappa_{\text{KL}})$. On any throughput floor $\rho(\delta^\gamma) \geq \bar{\rho} > 0$ we therefore have

$$\text{FDR}(\delta^\gamma) = \frac{\pi_0 q_0^\gamma}{\rho(\delta^\gamma)} \geq \frac{\pi_0}{\bar{\rho}} q_0^\gamma = \exp(-o(\kappa_{\text{KL}}))$$

If instead $\mathbb{E}_H[\tau^\gamma] = O(1)$, then $K(\delta^\gamma) = O(1)$ and the same bound implies q_0^γ is bounded away from 0 for all sufficiently small γ (and hence $\text{FDR}(\delta^\gamma)$ is bounded away from 0 as well). \square

A.4.4 Proof of Corollary 5.3

Corollary 5.3 (restated). Assume $D_{\text{mix}}(\theta) \asymp n_{\text{eff}}(\theta)$ so $\kappa_{\text{KL}} = \Theta(\kappa_{\text{eff}})$. Then robustness-check policies with $m(\gamma) = \Theta(1/\gamma)$ attain the optimal purification scale on both slices:

1. on any nonvanishing-throughput slice, $\text{FDR} = \exp(-\Theta(\kappa_{\text{KL}}))$ is achievable and unimprovable up to constants in the exponent;
2. on the fixed-FDR slice, sustaining nonvanishing throughput requires $\log(1/\varepsilon) = O(\kappa_{\text{KL}})$; conversely, there exists $c_\star > 0$ such that any target sequence satisfying $\log(1/\varepsilon) \leq c_\star \kappa_{\text{KL}}$ is attainable by an appropriate choice of $m(\gamma) = \Theta(1/\gamma)$ without collapsing throughput.

Proof. Assume comparability $D_{\text{mix}}(\theta) \asymp n_{\text{eff}}(\theta)$ so $\kappa_{\text{KL}} = \Theta(\kappa_{\text{eff}})$.

On any fixed-throughput slice $\rho(\delta) = \bar{\rho} > 0$, Corollary A.34 implies

$$\text{FDR}(\delta) \geq \exp(-O(\kappa_{\text{KL}}))$$

so no policy can purify faster than $\exp(-O(\kappa_{\text{KL}}))$ at nonvanishing throughput. On the other hand, Theorem 5.2 gives a robustness-check design with nonvanishing recall and

$$\text{FDR} = \exp(-\Omega(\kappa_{\text{eff}})) = \exp(-\Omega(\kappa_{\text{KL}}))$$

so under comparability robustness checks attain $\exp(-\Theta(\kappa_{\text{KL}}))$ purification up to constants in the exponent.

If a policy satisfies $\text{FDR}(\delta) \leq \varepsilon$, Corollary A.35 implies

$$\rho(\delta) \leq \frac{\pi_H}{1-\varepsilon} \cdot \frac{C_\nu + \kappa_{\text{KL}}/\pi_H + 1/e}{L(\varepsilon)} \quad \text{where } L(\varepsilon) = \log\left(\frac{\pi_0(1-\varepsilon)}{\pi_H\varepsilon}\right) \asymp \log\frac{1}{\varepsilon}$$

Thus sustaining nonvanishing throughput while $\varepsilon \downarrow 0$ requires $\log(1/\varepsilon) = O(\kappa_{\text{KL}})$, and no policy can substantially improve the exponent beyond the κ_{KL} -scale without collapsing throughput. Conversely, under comparability Theorem 5.2 gives a robustness-check design with nonvanishing throughput and

$$\text{FDR} \leq \exp(-c_\star \kappa_{\text{KL}})$$

for some constant $c_\star > 0$. Therefore any target sequence satisfying $\log(1/\varepsilon) \leq c_\star \kappa_{\text{KL}}$ is attained by such a $m(\gamma) = \Theta(1/\gamma)$ design. \square

A.4.5 A stopped-process KL budget

Lemma A.36. *Let P and Q be Markov laws for a latent process $(X_n)_{n \geq 1}$ with initial distributions ν_P, ν_Q and transition kernels K_P, K_Q . Let τ be a stopping time for the natural filtration $\mathcal{G}_n \equiv \sigma(X_1, \dots, X_n)$. Assume $\nu_P \ll \nu_Q$ and $K_P(x, \cdot) \ll K_Q(x, \cdot)$ for all x . Then*

$$\text{KL}(\mathcal{L}_P(X_{1:\tau}) \parallel \mathcal{L}_Q(X_{1:\tau})) \leq \text{KL}(\nu_P \parallel \nu_Q) + \mathbb{E}_P \left[\sum_{n=1}^{\tau-1} \text{KL}(K_P(X_n, \cdot) \parallel K_Q(X_n, \cdot)) \right]$$

Proof. Let Δ be a cemetery symbol and define the stopped process $Y_n \equiv X_n$ on $\{n \leq \tau\}$ and $Y_n \equiv \Delta$ on $\{n > \tau\}$. For $N \geq 1$, define the truncated stopping time $\tau_N \equiv \tau \wedge N$. Then $X_{1:\tau_N}$ is a measurable function of $Y_{1:N}$ (erase trailing Δ 's), hence by data processing,

$$\text{KL}(\mathcal{L}_P(X_{1:\tau_N}) \parallel \mathcal{L}_Q(X_{1:\tau_N})) \leq \text{KL}(\mathcal{L}_P(Y_{1:N}) \parallel \mathcal{L}_Q(Y_{1:N})) \quad \text{for every } N \geq 1$$

By the chain rule for relative entropy (e.g. Gray (1990, Eq. (2.24))),

$$\text{KL}(\mathcal{L}_P(Y_{1:N}) \parallel \mathcal{L}_Q(Y_{1:N})) = \text{KL}(\nu_P \parallel \nu_Q) + \sum_{n=1}^{N-1} \mathbb{E}_P [\text{KL}(\mathcal{L}_P(Y_{n+1} \mid Y_{1:n}) \parallel \mathcal{L}_Q(Y_{n+1} \mid Y_{1:n}))]$$

On the event $\{\tau \leq n\}$ we have $Y_{n+1} \equiv \Delta$ under both P and Q , so the conditional KL is 0. On $\{\tau > n\}$, we have $Y_{1:n} = X_{1:n}$ and $\{\tau > n\} \in \mathcal{G}_n$, so by the Markov property,

$$\mathcal{L}_P(Y_{n+1} \mid Y_{1:n}) = \mathcal{L}_P(X_{n+1} \mid X_n) = K_P(X_n, \cdot) \quad \mathcal{L}_Q(Y_{n+1} \mid Y_{1:n}) = K_Q(X_n, \cdot)$$

Therefore,

$$\text{KL}(\mathcal{L}_P(Y_{1:N}) \parallel \mathcal{L}_Q(Y_{1:N})) \leq \text{KL}(\nu_P \parallel \nu_Q) + \mathbb{E}_P \left[\sum_{n=1}^{N-1} \mathbf{1}\{\tau > n\} \text{KL}(K_P(X_n, \cdot) \parallel K_Q(X_n, \cdot)) \right]$$

Combining with the data-processing inequality above yields, for every N ,

$$\text{KL}(\mathcal{L}_P(X_{1:\tau_N}) \parallel \mathcal{L}_Q(X_{1:\tau_N})) \leq \text{KL}(\nu_P \parallel \nu_Q) + \mathbb{E}_P \left[\sum_{n=1}^{\tau_N-1} \text{KL}(K_P(X_n, \cdot) \parallel K_Q(X_n, \cdot)) \right]$$

Finally, as $N \rightarrow \infty$ we have $\tau_N \uparrow \tau$ and the stopped history $X_{1:\tau}$ is generated by the increasing truncations $(X_{1:\tau_N})_{N \geq 1}$. Therefore $\text{KL}(\mathcal{L}_P(X_{1:\tau_N}) \parallel \mathcal{L}_Q(X_{1:\tau_N})) \uparrow \text{KL}(\mathcal{L}_P(X_{1:\tau}) \parallel \mathcal{L}_Q(X_{1:\tau}))$, and monotone convergence yields the claim. \square

Lemma A.37. *Maintain Assumption A.1. Fix any policy δ and let $\tau^*(\delta)$ be the selected optimal stopping time. Let R be the induced report and let \mathcal{L}_H and \mathcal{L}_0 denote the induced laws under $T \in \mathcal{H}$ and $T \in \mathcal{T}_0$. Then*

$$\text{KL}(\mathcal{L}_H(R) \parallel \mathcal{L}_0(R)) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^*(\delta)]$$

Proof. By data processing, the report is a measurable function of the latent stopped history

$$(X_1, \dots, X_{\tau^*(\delta)})$$

Hence

$$\text{KL}(\mathcal{L}_H(R) \parallel \mathcal{L}_0(R)) \leq \text{KL}(\mathcal{L}_H(X_{1:\tau^*}) \parallel \mathcal{L}_0(X_{1:\tau^*}))$$

We bound the right-hand side by handling the non-high mixture at the path level. Let μ_H denote the conditional law of T given $T \in \mathcal{H}$ and let μ_0 denote the conditional law given $T \in \mathcal{T}_0$. For each high type $t_H \in \mathcal{H}$ and non-high type $t_0 \in \mathcal{T}_0$, write

$$P_{t_H} \equiv \mathcal{L}(X_{1:\tau^*} \mid T = t_H) \quad Q_{t_0} \equiv \mathcal{L}(X_{1:\tau^*} \mid T = t_0)$$

Then

$$\mathcal{L}_H(X_{1:\tau^*}) = \int P_{t_H} \mu_H(dt_H) \quad \mathcal{L}_0(X_{1:\tau^*}) = \int Q_{t_0} \mu_0(dt_0)$$

For fixed t_H , by convexity of KL in its second argument,

$$\text{KL}(P_{t_H} \parallel \mathcal{L}_0(X_{1:\tau^*})) \leq \int \text{KL}(P_{t_H} \parallel Q_{t_0}) \mu_0(dt_0)$$

By convexity of KL in its first argument,

$$\text{KL}(\mathcal{L}_H(X_{1:\tau^*}) \parallel \mathcal{L}_0(X_{1:\tau^*})) \leq \int \text{KL}(P_{t_H} \parallel \mathcal{L}_0(X_{1:\tau^*})) \mu_H(dt_H)$$

Combining,

$$\text{KL}(\mathcal{L}_H(X_{1:\tau^*}) \parallel \mathcal{L}_0(X_{1:\tau^*})) \leq \int \int \text{KL}(P_{t_H} \parallel Q_{t_0}) \mu_0(dt_0) \mu_H(dt_H)$$

Fix (t_H, t_0) . Since $P_n = \mathbf{p}(X_n)$, the observation filtration satisfies $\sigma(P_1, \dots, P_n) \subseteq \sigma(X_1, \dots, X_n)$, so τ^* is also a stopping time for the latent chain filtration. Under Assumption A.1, both type-conditional laws are Markov with stationary distributions ν_{t_H}, ν_{t_0} and transition kernels $K_{t_H}^\theta, K_{t_0}^\theta$. By Lemma A.36,

$$\text{KL}(P_{t_H} \parallel Q_{t_0}) \leq \text{KL}(\nu_{t_H} \parallel \nu_{t_0}) + \mathbb{E}_{t_H} \left[\sum_{n=1}^{\tau^*-1} \text{KL}(K_{t_H}^\theta(X_n, \cdot) \parallel K_{t_0}^\theta(X_n, \cdot)) \right]$$

By (6), $\text{KL}(\nu_{t_H} \parallel \nu_{t_0}) \leq C_\nu$. Moreover, by definition of $D(t_H \rightarrow t_0; \theta)$ and since $\sum_{n=1}^{\tau^*-1} 1 \leq \tau^*$,

$$\mathbb{E}_{t_H} \left[\sum_{n=1}^{\tau^*-1} \text{KL} \left(K_{t_H}^\theta(X_n, \cdot) \parallel K_{t_0}^\theta(X_n, \cdot) \right) \right] \leq D(t_H \rightarrow t_0; \theta) \mathbb{E}_{t_H}[\tau^*]$$

Therefore,

$$\text{KL}(P_{t_H} \parallel Q_{t_0}) \leq C_\nu + D(t_H \rightarrow t_0; \theta) \mathbb{E}_{t_H}[\tau^*]$$

Integrating over μ_0 and using (5) gives, for μ_H -a.e. t_H ,

$$\int \text{KL}(P_{t_H} \parallel Q_{t_0}) \mu_0(dt_0) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_{t_H}[\tau^*]$$

Integrating over μ_H and using $\mathbb{E}_H[\tau^*] = \int \mathbb{E}_{t_H}[\tau^*] \mu_H(dt_H)$ yields

$$\text{KL}(\mathcal{L}_H(X_{1:\tau^*}) \parallel \mathcal{L}_0(X_{1:\tau^*})) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^*]$$

Combining with the initial data-processing step proves the claim. \square

A.4.6 Proof of Theorem 4.3

Theorem 4.3 (restated). There exists a constant $C_\nu < \infty$ from Assumption A.1 such that for every policy δ with selected best-response stopping time $\tau^*(\delta)$,

$$K(\delta) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^*(\delta)] \leq C_\nu + \frac{\kappa_{\text{KL}}}{\pi_H}$$

Proof. Fix an environment (γ, θ) and an acceptance policy δ with selected best response $\sigma^*(\delta)$. Let R denote the induced report and let $A \in \{0, 1\}$ denote the induced acceptance decision. By construction, conditional on R the journal draws $U \sim \text{Unif}[0, 1]$ independently and sets $A = \mathbf{1}\{U \leq \delta(R)\}$, so A is generated from R by a Markov kernel. Therefore, data processing yields

$$\text{KL}(\mathcal{L}_H(A) \parallel \mathcal{L}_0(A)) \leq \text{KL}(\mathcal{L}_H(R) \parallel \mathcal{L}_0(R))$$

Since $\mathcal{L}_H(A) = \text{Bern}(q_H(\delta))$ and $\mathcal{L}_0(A) = \text{Bern}(q_0(\delta))$, the left-hand side equals

$$\text{KL}(\text{Bern}(q_H(\delta)) \parallel \text{Bern}(q_0(\delta)))$$

By Lemma A.37,

$$\text{KL}(\mathcal{L}_H(R) \parallel \mathcal{L}_0(R)) \leq C_\nu + D_{\text{mix}}(\theta) \mathbb{E}_H[\tau^*(\delta)]$$

Combining the displays gives the first inequality in Theorem 4.3. Finally, Lemma A.13 implies $\mathbb{E}_H[\tau^*(\delta)] \leq 1/(\pi_H \gamma)$, which yields the second display. \square

A.5 Short reports and extra information from selection

This appendix provides the technical change-of-measure machinery behind Theorem 4.5. We formalize a truncated-report divergence bound and an extra-information-from-selection term Λ_{sel} , and then use them to prove the theorem and its disclosure-scaling implication. The argument relies on the report-space formalization in Appendix A.1.2 and uses the monotonicity tools in Appendix A.2 to obtain verifiable selection-control conditions (verified in Appendix A.6.5).

A.5.1 A short-report (truncated-report) Rényi budget

Fix an integer horizon $n \geq 1$. Let τ be the (selected) stopping time and let $R \in \mathcal{R}$ be the reported counting measure on $(0, 1)$ (equivalently, the reported multiset of p -values; see Section A.2.1). Define the truncated report

$$\tilde{R}^{(n)} \equiv \begin{cases} R & \tau \leq n \\ \perp & \tau > n \end{cases}$$

where \perp is a cemetery symbol.

The proof of Theorem 4.5 proceeds by applying a Rényi change-of-measure bound to the acceptance event truncated at a large horizon n_γ . The key input is that, under short disclosure and negligible extra information from selection, the truncated report $\tilde{R}^{(n_\gamma)}$ cannot carry κ_{eff} -scale separation between H and 0. Accordingly we work with an explicit *truncated-report* Rényi divergence bound for $\tilde{R}^{(n)}$. This is the natural object because the journal observes only the disclosed report; bounding the divergence of an augmented output that reveals unobserved selection objects (such as (τ, I)) can be overly conservative in multi-type window environments.

We obtain a sufficient condition for the truncated-report budget by upper bounding the divergence of $\tilde{R}^{(n)}$ by that of an augmented transcript that includes the stopping time and disclosed indices. The only place where this augmentation can be conservative is in comparing the induced selection weights under H versus 0; we isolate this as a belief-state extra-information-from-selection term $\Lambda_{\text{sel}}(n, k)$.

The bound is written in terms of a marginal Rényi rate $d_\alpha(\theta)$. We maintain the subset Rényi growth condition in Assumption A.6.

Fix a horizon n and a report cap m (so $|R| \leq m$ almost surely). On $\{\tau \leq n\}$ let $I \subseteq \{1, \dots, \tau\}$ be the (random) set of disclosed indices, so $|I| \leq m$. Define $J \equiv I \cup \{\tau\}$, so $|J| \leq m + 1$. Define the augmented truncated transcript

$$\hat{R}^{(n)} \equiv \begin{cases} (\tau, I, Z_J) & \tau \leq n \\ \perp & \tau > n \end{cases}$$

where Z_J is listed in increasing index order and \perp is a cemetery symbol. Then $\tilde{R}^{(n)}$ is a measurable function of $\hat{R}^{(n)}$, hence by data processing,

$$D_\alpha(\mathcal{L}_H(\tilde{R}^{(n)}) \parallel \mathcal{L}_0(\tilde{R}^{(n)})) \leq D_\alpha(\mathcal{L}_H(\hat{R}^{(n)}) \parallel \mathcal{L}_0(\hat{R}^{(n)}))$$

For each component $c \equiv (t, I)$ of $\hat{R}^{(n)}$ with $t \leq n$, write $J(c) \equiv I \cup \{t\}$ and let $w_u^{(c)}(z_{J(c)})$ denote the conditional probability under $u \in \{H, 0\}$ that the transcript equals (t, I) given $Z_{J(c)} = z_{J(c)}$. For the cemetery component, set $J(\perp) \equiv \emptyset$ and write $w_u^{(\perp)} \equiv \mathbb{P}_u(\tau > n)$. We define the order-0 (log-LR) extra-information-from-selection term as

$$\Lambda_{\text{sel}}(n, k) \equiv \sup_{\substack{c \in \{\perp\} \cup \{(t, I): t \leq n\} \\ |J(c)| \leq k \\ \mathbb{P}_0(\hat{R}^{(n)} \text{ is in component } c) > 0}} \text{ess sup}_{z_{J(c)}} \log \frac{w_H^{(c)}(z_{J(c)})}{w_0^{(c)}(z_{J(c)})} \in [0, \infty] \quad (13)$$

where the essential supremum is taken with respect to the conditional law

$$\mathcal{L}_0(Z_{J(c)} \mid \hat{R}^{(n)} \text{ is in component } c)$$

Equivalently, it is taken with respect to the unnormalized measure $w_0^{(c)}(z_{J(c)}) \mathcal{L}_0(Z_{J(c)})(dz_{J(c)})$, so $\Lambda_{\text{sel}}(n, k)$ ignores zero-probability transcript components under \mathcal{L}_0 . Intuitively, $\Lambda_{\text{sel}}(n, k)$ quantifies

how much additional H -vs-0 likelihood ratio can be created by endogenous stopping and selective disclosure beyond what is already carried by the disclosed coordinates.

The definition (13) is endogenous: it depends on the equilibrium stopping and reporting rule induced by the journal's policy. Theorem 4.5 therefore treats Λ_{sel} as a primitive "extra information from selection" object and assumes only that it is negligible on the κ_{eff} scale along the relevant equilibrium sequence.

Assumption A.38. Along the equilibrium sequence considered in Theorem 4.5, with truncation horizon

$$n_\gamma \equiv \left\lceil \frac{8}{\gamma} \log \frac{1}{\gamma} \right\rceil$$

we have

$$\Lambda_{\text{sel}}(n_\gamma, m(\gamma) + 1) = o(\kappa_{\text{eff}}) \quad \text{as } \gamma \downarrow 0$$

Since $n \mapsto \Lambda_{\text{sel}}(n, k)$ is nondecreasing, Assumption A.38 also implies $\Lambda_{\text{sel}}(n, m(\gamma) + 1) = o(\kappa_{\text{eff}})$ for every $n = O(1/\gamma)$. Appendix A.6.5 verifies Assumption A.38 in the ordered three-type Gaussian AR(1) example by checking the ordered-sandwiching sufficient condition in Proposition A.41; in particular, selection contributes only an $O(1)$ additive term (hence is negligible on the κ_{eff} scale).

Proposition A.39. Fix $\alpha > 1$ and integers $n \geq 1$ and $m \geq 0$. Maintain Assumption A.6. Then for every equilibrium in which $|R| \leq m$ almost surely,

$$D_\alpha(\mathcal{L}_H(\tilde{R}^{(n)}) \parallel \mathcal{L}_0(\tilde{R}^{(n)})) \leq (m+1)d_\alpha(\theta) + \frac{1}{\alpha-1} \log \left(\sum_{k=0}^{m+1} \binom{n+1}{k} \right) + \Lambda_{\text{sel}}(n, m+1).$$

Proof. We upper bound the Rényi divergence of the truncated report by that of the augmented transcript $\hat{R}^{(n)}$. For $\hat{R}^{(n)}$, we expand the α -moment of the likelihood ratio over its disjoint components c . On each component c , the likelihood ratio factors into (i) the marginal likelihood ratio of the disclosed coordinates Z_J and (ii) a ratio of conditional selection weights $w_H^{(c)}/w_0^{(c)}$ capturing the extra information carried by stopping and disclosure. The marginal term is controlled by Assumption A.6, while the selection term is controlled by $\Lambda_{\text{sel}}(n, m+1)$. Counting the number of possible components yields the combinatorial term.

By data processing,

$$D_\alpha(\mathcal{L}_H(\tilde{R}^{(n)}) \parallel \mathcal{L}_0(\tilde{R}^{(n)})) \leq D_\alpha(\mathcal{L}_H(\hat{R}^{(n)}) \parallel \mathcal{L}_0(\hat{R}^{(n)}))$$

Let $P \equiv \mathcal{L}_H(\hat{R}^{(n)})$ and $Q \equiv \mathcal{L}_0(\hat{R}^{(n)})$, and let $L \equiv dP/dQ$. Then

$$\exp\left((\alpha-1)D_\alpha(\mathcal{L}_H(\hat{R}^{(n)}) \parallel \mathcal{L}_0(\hat{R}^{(n)}))\right) = \mathbb{E}_Q[L^\alpha]$$

Since the components of $\hat{R}^{(n)}$ are disjoint (including the cemetery component \perp),

$$\mathbb{E}_Q[L^\alpha] = \sum_c \mathbb{E}_Q\left[L^\alpha \mathbf{1}\{\hat{R}^{(n)} \text{ is in component } c\}\right]$$

Fix a component c . If $c = \perp$, set $J \equiv \emptyset$ and recall $w_u^{(\perp)} = \mathbb{P}_u(\tau > n)$. If $c \equiv (t, I)$ with $t \leq n$, set $J \equiv J(c) = I \cup \{t\}$ and recall $w_u^{(c)}(z_J) = \mathbb{P}_u((\tau, I) = c \mid Z_J = z_J)$.

Write $\ell_J(z_J) \equiv d\mathcal{L}_H(Z_J)/d\mathcal{L}_0(Z_J)(z_J)$ for the marginal likelihood ratio on the disclosed coordinates (with the convention $\ell_\emptyset \equiv 1$). On component c , the likelihood ratio factors as

$$L = \ell_J(Z_J) \cdot \frac{w_H^{(c)}(Z_J)}{w_0^{(c)}(Z_J)} \quad Q\text{-almost surely}$$

Therefore,

$$\mathbb{E}_Q \left[L^\alpha \mathbf{1}\{\widehat{R}^{(n)} \text{ is in component } c\} \right] = \int \ell_J(z_J)^\alpha \left(\frac{w_H^{(c)}(z_J)}{w_0^{(c)}(z_J)} \right)^\alpha w_0^{(c)}(z_J) \mathcal{L}_0(Z_J)(dz_J)$$

where we interpret the integrand as zero on $\{w_0^{(c)}(z_J) = 0\}$.

Since $w_H^{(c)} \in [0, 1]$ and $\alpha > 1$, on $\{w_0^{(c)}(z_J) > 0\}$ we have

$$\left(\frac{w_H^{(c)}}{w_0^{(c)}} \right)^\alpha w_0^{(c)} = \left(\frac{w_H^{(c)}}{w_0^{(c)}} \right)^{\alpha-1} w_H^{(c)} \leq \left(\frac{w_H^{(c)}}{w_0^{(c)}} \right)^{\alpha-1}$$

By definition of $\Lambda_{\text{sel}}(n, m+1)$ and since $|J| \leq m+1$ on every component,

$$\left(\frac{w_H^{(c)}(z_J)}{w_0^{(c)}(z_J)} \right)^{\alpha-1} \leq e^{(\alpha-1)\Lambda_{\text{sel}}(n, m+1)}$$

for $(w_0^{(c)}(z_J) \mathcal{L}_0(Z_J)(dz_J))$ -a.e. z_J . Thus each component satisfies

$$\begin{aligned} \mathbb{E}_Q \left[L^\alpha \mathbf{1}\{\widehat{R}^{(n)} \text{ is in component } c\} \right] &\leq e^{(\alpha-1)\Lambda_{\text{sel}}(n, m+1)} \int \ell_J(z_J)^\alpha \mathcal{L}_0(Z_J)(dz_J) \\ &= \exp\left((\alpha-1)\{\Lambda_{\text{sel}}(n, m+1) + D_\alpha(\mathcal{L}_H(Z_J)\|\mathcal{L}_0(Z_J))\} \right) \end{aligned}$$

By Assumption A.6 and $|J| \leq m+1$,

$$D_\alpha(\mathcal{L}_H(Z_J)\|\mathcal{L}_0(Z_J)) \leq (m+1)d_\alpha(\theta)$$

For each $t \leq n$ and each $k \leq \min\{m, t\}$ there are $\binom{t}{k}$ index sets $I \subseteq \{1, \dots, t\}$ of size k . Thus the number of non-cemetery components of the form (t, I) is at most

$$\sum_{t=0}^n \sum_{k=0}^{\min\{m, t\}} \binom{t}{k}$$

and including the cemetery component gives at most

$$1 + \sum_{t=0}^n \sum_{k=0}^{\min\{m, t\}} \binom{t}{k} = \sum_{k=0}^{m+1} \binom{n+1}{k}$$

using $\sum_{t=k}^n \binom{t}{k} = \binom{n+1}{k+1}$. Summing the component bounds therefore yields

$$\exp\left((\alpha-1)D_\alpha(\mathcal{L}_H(\widehat{R}^{(n)})\|\mathcal{L}_0(\widehat{R}^{(n)})) \right) \leq \left(\sum_{k=0}^{m+1} \binom{n+1}{k} \right) e^{(\alpha-1)\{(m+1)d_\alpha(\theta) + \Lambda_{\text{sel}}(n, m+1)\}}$$

Taking logs and dividing by $\alpha-1$ gives the stated bound. \square

A.5.2 A sufficient condition for Assumption A.38

Assumption A.38 is the most opaque primitive in the short-report analysis because it is endogenous: it depends on the equilibrium stopping and disclosure rule induced by the journal’s policy. The sufficient condition below is stated for a generic finite environment (it does not assume an AR(1) structure); Appendix A.6.5 verifies it in the Gaussian AR(1) running example. In many ordered environments, however, selection can create at most a constant amount of additional H -vs-0 likelihood ratio beyond what is already carried by the disclosed coordinates. The next proposition records one sufficient condition.

Assumption A.40. Fix $n \geq 1$ and consider an equilibrium with $|R| \leq m$ almost surely for some $m \geq 0$. Let $\widehat{R}^{(n)}$ be the augmented truncated transcript from Appendix A.5.1 with components $c \equiv (t, I)$ and $J(c) = I \cup \{t\}$. We assume that for every component c with $\mathbb{P}_0(\widehat{R}^{(n)} \text{ is in component } c) > 0$ and $\mathcal{L}_0(Z_{J(c)} \mid \widehat{R}^{(n)} \text{ is in component } c)$ -a.e. $z_{J(c)}$, the conditional selection event $\{(\tau, I) = c\}$ viewed as a subset of the omitted coordinates $Z_{-J(c)}$ given $Z_{J(c)} = z_{J(c)}$ is two-sided monotone (Definition A.11) in the coordinatewise order.

Assumption A.40 holds, for example, if conditional on disclosed coordinates the transcript component is determined by whether a scalar belief statistic of the omitted coordinates (e.g. posterior odds or a likelihood-ratio statistic) lies outside an interval, and that statistic is coordinatewise increasing; Lemma A.12 records this scalar sufficient condition.

The key structural requirement is a conditional stochastic ordering: given any disclosed subvector $Z_J = z_J$, the conditional law of the omitted coordinates under H is sandwiched between the laws under two non-high types in the down-set order. Intuitively, this is a conditional form of “ H lies between two extremes” on the evidence scale. It holds in many ordered multi-type settings; Appendix A.6.5 verifies it explicitly in the ordered three-type Gaussian AR(1) example via an exact monotone-shift property.

In the i.i.d. benchmark, ordered sandwiching reduces to a transparent parameter ordering in many one-dimensional exponential-family models with the monotone likelihood ratio property: (14) holds whenever the high type is bracketed by two non-high extremes (e.g. $\mu_N < \mu_H < \mu_L$ in the Gaussian location model). The polar case where all high types lie on one side of all non-high types corresponds to one-sided (threshold) screening, and can be treated by a one-sided variant under one-sided transcript events; Corollary A.42 records such a bound.

A common route to ordered sandwiching is a conditional mean-shift property: if, given $Z_J = z_J$, the conditional law of Z_{-J} has a type-independent covariance and its conditional mean shifts coordinatewise with the type parameter, then down-set probabilities are ordered by translation.

Proposition A.41. Fix $\alpha > 1$ and integers $n \geq 1$ and $m \geq 0$. Consider any finite environment with \mathcal{H} possibly multi-valued and non-high set \mathcal{T}_0 . Assume there exist two non-high types $t^-, t^+ \in \mathcal{T}_0$ with prior weights

$$\pi_{t^-}^0 \equiv \mathbb{P}(T = t^- \mid T \in \mathcal{T}_0) > 0 \quad \pi_{t^+}^0 \equiv \mathbb{P}(T = t^+ \mid T \in \mathcal{T}_0) > 0 \quad \pi_{\min}^{0,\pm} \equiv \min\{\pi_{t^-}^0, \pi_{t^+}^0\}$$

such that for every index set $J \subseteq \{1, \dots, n\}$, every z_J , and every down-set $D \subseteq \mathbb{R}^{|-J|}$,

$$\mathbb{P}_{t^+}(Z_{-J} \in D \mid Z_J = z_J) \leq \mathbb{P}_H(Z_{-J} \in D \mid Z_J = z_J) \leq \mathbb{P}_{t^-}(Z_{-J} \in D \mid Z_J = z_J) \quad (14)$$

where $\mathbb{P}_H(\cdot \mid Z_J = z_J)$ denotes the conditional law under $T \in \mathcal{H}$.

Assume an equilibrium in which $|R| \leq m$ almost surely, and consider the augmented truncated transcript $\widehat{R}^{(n)}$ with components $c \equiv (t, I)$ and $J(c) = I \cup \{t\}$ from Appendix A.5.1. Assume

Assumption A.40. Finally, assume there exists $d_\alpha(\theta) < \infty$ such that for every $J \subseteq \{1, \dots, n\}$ and each $u \in \{t^-, t^+\}$,

$$D_\alpha(\mathcal{L}_H(Z_J) \parallel \mathcal{L}_u(Z_J)) \leq |J| d_\alpha(\theta)$$

Then the truncated-report Rényi divergence satisfies

$$D_\alpha(\mathcal{L}_H(\tilde{R}^{(n)}) \parallel \mathcal{L}_0(\tilde{R}^{(n)})) \leq (m+1)d_\alpha(\theta) + \frac{1}{\alpha-1} \log \left(\sum_{k=0}^{m+1} \binom{n+1}{k} \right) + \log \frac{1}{\pi_{\min}^{0,\pm}} + C_\alpha$$

where $C_\alpha < \infty$ depends only on α . In particular, ordered sandwiching and two-sided transcript events contribute only an $O_\alpha(1) + \log(1/\pi_{\min}^{0,\pm})$ additive term, uniformly in the truncation horizon n .

Proof. The proof follows the disjoint-component expansion in Proposition A.39, but bounds the selection factor directly using the sandwiching condition (14) and Assumption A.40.

The component expansion expresses $\mathbb{E}_0[(d\mathcal{L}_H/d\mathcal{L}_0)^\alpha]$ as a sum over transcript components. For each component, Assumption A.40 and (14) compare the component's selection probability under H to its probabilities under the two bracketing non-high types t^- and t^+ . This yields a bound on $(w_H^{(c)})^\alpha (w_0^{(c)})^{1-\alpha}$ in terms of posterior non-high mixture weights, and hence in terms of $\pi_{\min}^{0,\pm}$. The result is a constant extra-information-from-selection term $\log(1/\pi_{\min}^{0,\pm})$, independent of the horizon.

By data processing,

$$D_\alpha(\mathcal{L}_H(\tilde{R}^{(n)}) \parallel \mathcal{L}_0(\tilde{R}^{(n)})) \leq D_\alpha(\mathcal{L}_H(\hat{R}^{(n)}) \parallel \mathcal{L}_0(\hat{R}^{(n)}))$$

Let $P \equiv \mathcal{L}_H(\hat{R}^{(n)})$ and $Q \equiv \mathcal{L}_0(\hat{R}^{(n)})$ and write $L \equiv dP/dQ$. Then

$$\exp\left((\alpha-1)D_\alpha(\mathcal{L}_H(\hat{R}^{(n)}) \parallel \mathcal{L}_0(\hat{R}^{(n)}))\right) = \mathbb{E}_Q[L^\alpha]$$

Since components are disjoint,

$$\mathbb{E}_Q[L^\alpha] = \sum_c \mathbb{E}_Q \left[L^\alpha \mathbf{1}\{\hat{R}^{(n)} \text{ is in component } c\} \right]$$

Fix a component $c \equiv (t, I)$ with $t \leq n$ and write $J \equiv J(c)$. Write $w_H^{(c)}(z_J) \equiv \mathbb{P}_H((\tau, I) = c \mid Z_J = z_J)$, and for each non-high type $u \in \mathcal{T}_0$ write $w_u^{(c)}(z_J) \equiv \mathbb{P}_u((\tau, I) = c \mid Z_J = z_J)$.

By Assumption A.40, for $(w_0^{(c)}(z_J) \mathcal{L}_0(Z_J)(dz_J))$ -a.e. z_J there exist a down-set $D \subseteq \mathbb{R}^{|-J|}$ and an up-set $U \subseteq \mathbb{R}^{|-J|}$ such that

$$\{(\tau, I) = c\} \equiv \{Z_{-J} \in D \cup U\} \quad \text{conditional on } Z_J = z_J$$

Since U^c is a down-set, (14) implies the corresponding up-set ordering. Hence

$$\begin{aligned} w_H^{(c)}(z_J) &= \mathbb{P}_H(Z_{-J} \in D \cup U \mid Z_J = z_J) \\ &\leq \mathbb{P}_{t^-}(Z_{-J} \in D \mid Z_J = z_J) + \mathbb{P}_{t^+}(Z_{-J} \in U \mid Z_J = z_J) \\ &\leq w_{t^-}^{(c)}(z_J) + w_{t^+}^{(c)}(z_J) \end{aligned}$$

Write $\pi_u(z_J) \equiv \mathbb{P}(T = u \mid T \in \mathcal{T}_0, Z_J = z_J)$ for posterior non-high mixture weights, so $w_0^{(c)}(z_J) = \sum_{u \in \mathcal{T}_0} \pi_u(z_J) w_u^{(c)}(z_J)$. Therefore,

$$w_0^{(c)}(z_J) \geq \min\{\pi_{t^-}(z_J), \pi_{t^+}(z_J)\} (w_{t^-}^{(c)}(z_J) + w_{t^+}^{(c)}(z_J)) \geq \min\{\pi_{t^-}(z_J), \pi_{t^+}(z_J)\} w_H^{(c)}(z_J)$$

Thus

$$\left(\frac{w_H^{(c)}(z_J)}{w_0^{(c)}(z_J)}\right)^\alpha w_0^{(c)}(z_J) = \left(\frac{w_H^{(c)}(z_J)}{w_0^{(c)}(z_J)}\right)^{\alpha-1} w_H^{(c)}(z_J) \leq \min\{\pi_{t^-}(z_J), \pi_{t^+}(z_J)\}^{1-\alpha}$$

Let $\ell_{u,J}(z_J) \equiv d\mathcal{L}_u(Z_J)/d\mathcal{L}_0(Z_J)(z_J)$ denote type- u likelihood ratios against the non-high mixture. Since $\mathcal{L}_0(Z_J) = \sum_{u \in \mathcal{T}_0} \pi_u^0 \mathcal{L}_u(Z_J)$, Bayes' rule gives

$$\pi_u(z_J) = \pi_u^0 \ell_{u,J}(z_J) \quad \text{for } \mathcal{L}_0(Z_J)\text{-a.e. } z_J$$

Hence

$$\min\{\pi_{t^-}(z_J), \pi_{t^+}(z_J)\}^{1-\alpha} \leq (\pi_{\min}^{0,\pm})^{1-\alpha} \sum_{u \in \{t^-, t^+\}} \ell_{u,J}(z_J)^{1-\alpha}$$

Write $\ell_{H,J}(z_J) \equiv d\mathcal{L}_H(Z_J)/d\mathcal{L}_0(Z_J)(z_J)$. On component c the likelihood ratio factors as $L = \ell_{H,J}(Z_J) \cdot (w_H^{(c)}(Z_J)/w_0^{(c)}(Z_J))$. Therefore,

$$\begin{aligned} \mathbb{E}_Q \left[L^\alpha \mathbf{1}\{\widehat{R}^{(n)} \text{ is in component } c\} \right] &= \int \ell_{H,J}(z_J)^\alpha \left(\frac{w_H^{(c)}(z_J)}{w_0^{(c)}(z_J)}\right)^\alpha w_0^{(c)}(z_J) \mathcal{L}_0(Z_J)(dz_J) \\ &\leq (\pi_{\min}^{0,\pm})^{1-\alpha} \sum_{u \in \{t^-, t^+\}} \int \ell_{H,J}(z_J)^\alpha \ell_{u,J}(z_J)^{1-\alpha} \mathcal{L}_0(Z_J)(dz_J) \end{aligned}$$

Since $d\mathcal{L}_H(Z_J)/d\mathcal{L}_u(Z_J) = \ell_{H,J}/\ell_{u,J}$, each term equals

$$\int \ell_{H,J}^\alpha \ell_{u,J}^{1-\alpha} d\mathcal{L}_0 = \int \left(\frac{d\mathcal{L}_H(Z_J)}{d\mathcal{L}_u(Z_J)}\right)^\alpha d\mathcal{L}_u(Z_J) = \exp\left((\alpha-1)D_\alpha(\mathcal{L}_H(Z_J)\|\mathcal{L}_u(Z_J))\right)$$

By the marginal assumption and $|J| \leq m+1$,

$$\exp\left((\alpha-1)D_\alpha(\mathcal{L}_H(Z_J)\|\mathcal{L}_u(Z_J))\right) \leq \exp\left((\alpha-1)(m+1)d_\alpha(\theta)\right)$$

for $u \in \{t^-, t^+\}$. Therefore each component satisfies

$$\mathbb{E}_Q \left[L^\alpha \mathbf{1}\{\widehat{R}^{(n)} \text{ is in component } c\} \right] \leq 2(\pi_{\min}^{0,\pm})^{1-\alpha} \exp\left((\alpha-1)(m+1)d_\alpha(\theta)\right)$$

Counting components as in Proposition A.39 and taking logs yields the stated bound, with the factor of 2 absorbed into C_α . \square

Corollary A.42. Fix $\alpha > 1$ and integers $n \geq 1$ and $m \geq 0$. Work in the setting of Proposition A.41, but replace Assumption A.40 with the stronger requirement that for every transcript component c and for $(w_0^{(c)}(z_J) \mathcal{L}_0(Z_{J(c)})(dz_J))$ -a.e. z_J , the conditional selection event $\{(\tau, I) = c\}$ is a down-set in the omitted coordinates. Assume there exists a non-high type $t^- \in \mathcal{T}_0$ with $\pi_{t^-}^0 > 0$ such that for every index set $J \subseteq \{1, \dots, n\}$, every z_J , and every down-set $D \subseteq \mathbb{R}^{|-J|}$,

$$\mathbb{P}_H(Z_{-J} \in D \mid Z_J = z_J) \leq \mathbb{P}_{t^-}(Z_{-J} \in D \mid Z_J = z_J)$$

Finally, assume there exists $d_\alpha(\theta) < \infty$ such that for every $J \subseteq \{1, \dots, n\}$,

$$D_\alpha(\mathcal{L}_H(Z_J) \|\mathcal{L}_{t^-}(Z_J)) \leq |J| d_\alpha(\theta)$$

Then the truncated-report Rényi divergence bound of Proposition A.41 holds with the selection term $\log(1/\pi_{\min}^{0,\pm})$ replaced by $\log(1/\pi_{t^-}^0)$. An analogous statement holds for up-sets by reversing the ordering and inequalities.

Proof. Follow the proof of Proposition A.41. Under the down-set transcript condition and the one-sided ordering, Step 2 gives $w_H^{(c)}(z_J) \leq w_{t^-}^{(c)}(z_J)$ and hence $w_0^{(c)}(z_J) \geq \pi_{t^-}(z_J) w_H^{(c)}(z_J)$, yielding the factor $\pi_{t^-}(z_J)^{1-\alpha}$ in place of $\min\{\pi_{t^-}(z_J), \pi_{t^+}(z_J)\}^{1-\alpha}$. Bayes' rule gives $\pi_{t^-}(z_J) = \pi_{t^-}^0 \ell_{t^-,J}(z_J)$ for $\mathcal{L}_0(Z_J)$ -a.e. z_J , so the component integral becomes $\exp((\alpha - 1)D_\alpha(\mathcal{L}_H(Z_J) \parallel \mathcal{L}_{t^-}(Z_J)))$. The marginal Rényi bound and the component count proceed exactly as before. \square

Proposition A.41 provides a verifiable route for handling the endogenous selection term in Assumption A.38. Under two-sided transcript events and ordered sandwiching, it bounds the truncated-report Rényi divergence with a *selection penalty* that is constant in the truncation horizon n (namely $\log(1/\pi_{\min}^{0,\pm}) + O_\alpha(1)$), without needing to control $\Lambda_{\text{sel}}(n, k)$ directly. Since $\kappa_{\text{eff}} \asymp n_{\text{eff}}(\theta)/\gamma \rightarrow \infty$ as $\gamma \downarrow 0$ when $n_{\text{eff}}(\theta) > 0$, this contribution is negligible on the κ_{eff} scale in Theorem 4.5.

A.5.3 Proof of Theorem 4.5

Theorem 4.5 (restated). Fix θ with $n_{\text{eff}}(\theta) > 0$ and consider $\gamma \downarrow 0$. Let δ^γ be any sequence of policies and let (q_H^γ, q_0^γ) be the induced operating points under the selected best response. Assume short disclosure: $|R| \leq m(\gamma)$ a.s. with $m(\gamma) = o(1/\gamma)$. Maintain Assumption A.6 for some $\alpha > 1$, and assume selection control as in Assumption A.38. Then for every $c_H > 0$, if $\liminf_{\gamma \downarrow 0} q_H^\gamma \geq c_H$ then $-\log q_0^\gamma = o(\kappa_{\text{eff}})$ (equivalently $q_0^\gamma = \exp(-o(\kappa_{\text{eff}}))$).

Proof. Fix $\alpha > 1$ and a policy δ^γ satisfying the assumptions of Theorem 4.5. Maintain Assumption A.6 and Assumption A.38. Let τ^* be the selected best-response stopping time and let R be the induced report.

Fix $c_H > 0$ and suppose $\liminf_{\gamma \downarrow 0} q_H(\delta^\gamma) \geq c_H$. Then for all sufficiently small γ we have $q_H(\delta^\gamma) \geq c_H/2$. Define the truncation horizon

$$\bar{n}_\gamma \equiv \left\lceil \frac{4}{\pi_H c_H \gamma} \right\rceil$$

and the truncated report $\tilde{R}^{(\bar{n}_\gamma)}$ as in Appendix A.5.1.

Define the bounded statistic

$$f \equiv \delta^\gamma(R) \mathbf{1}\{\tau^* \leq \bar{n}_\gamma\} \in [0, 1]$$

Since $\tilde{R}^{(\bar{n}_\gamma)}$ equals R on $\{\tau^* \leq \bar{n}_\gamma\}$ and equals \perp otherwise, f is measurable with respect to $\tilde{R}^{(\bar{n}_\gamma)}$.

Applying Lemma A.30 with $P = \mathcal{L}_H(\tilde{R}^{(\bar{n}_\gamma)})$ and $Q = \mathcal{L}_0(\tilde{R}^{(\bar{n}_\gamma)})$ yields

$$\mathbb{E}_0[f] \geq \mathbb{E}_H[f]^{\alpha/(\alpha-1)} \exp\left(-D_\alpha(\mathcal{L}_H(\tilde{R}^{(\bar{n}_\gamma)}) \parallel \mathcal{L}_0(\tilde{R}^{(\bar{n}_\gamma)}))\right)$$

Since $q_0(\delta^\gamma) = \mathbb{E}_0[\delta^\gamma(R)]$ and $\delta^\gamma(R) \geq 0$, we have $q_0(\delta^\gamma) \geq \mathbb{E}_0[f]$.

A constant lower bound on $\mathbb{E}_H[f]$:

$$\mathbb{E}_H[f] = \mathbb{E}_H[\delta^\gamma(R) \mathbf{1}\{\tau^* \leq \bar{n}_\gamma\}] \geq \mathbb{E}_H[\delta^\gamma(R)] - \mathbb{P}_H(\tau^* > \bar{n}_\gamma) = q_H(\delta^\gamma) - \mathbb{P}_H(\tau^* > \bar{n}_\gamma)$$

Lemma A.13 implies $\mathbb{E}_H[\tau^*] \leq 1/(\pi_H \gamma)$, hence Markov's inequality gives

$$\mathbb{P}_H(\tau^* > \bar{n}_\gamma) \leq \frac{1}{\pi_H \gamma \bar{n}_\gamma} \leq \frac{c_H}{4}$$

Therefore for all sufficiently small γ ,

$$\mathbb{E}_H[f] \geq \frac{c_H}{2} - \frac{c_H}{4} = \frac{c_H}{4}$$

Proposition A.39 gives

$$\begin{aligned} D_\alpha\left(\mathcal{L}_H(\tilde{R}^{(\bar{n}_\gamma)}) \parallel \mathcal{L}_0(\tilde{R}^{(\bar{n}_\gamma)})\right) &\leq (m(\gamma) + 1)d_\alpha(\theta) + \frac{1}{\alpha - 1} \log\left(\sum_{k=0}^{m(\gamma)+1} \binom{\bar{n}_\gamma + 1}{k}\right) \\ &\quad + \Lambda_{\text{sel}}(\bar{n}_\gamma, m(\gamma) + 1) \end{aligned}$$

Since $\bar{n}_\gamma \asymp 1/\gamma$ and $m(\gamma) = o(1/\gamma)$, the first term is $o(\kappa_{\text{eff}})$. Moreover, for all sufficiently small γ we have $m(\gamma) + 1 \leq (\bar{n}_\gamma + 1)/2$, so

$$\sum_{k=0}^{m(\gamma)+1} \binom{\bar{n}_\gamma + 1}{k} \leq \exp\left((m(\gamma) + 1) \log\left(\frac{e(\bar{n}_\gamma + 1)}{m(\gamma) + 1}\right)\right)$$

and the combinatorial term is also $o(\kappa_{\text{eff}})$. Finally, since $\bar{n}_\gamma \leq n_\gamma$ for all sufficiently small γ and $n \mapsto \Lambda_{\text{sel}}(n, k)$ is nondecreasing, Assumption A.38 implies

$$\Lambda_{\text{sel}}(\bar{n}_\gamma, m(\gamma) + 1) = o(\kappa_{\text{eff}})$$

Thus

$$D_\alpha\left(\mathcal{L}_H(\tilde{R}^{(\bar{n}_\gamma)}) \parallel \mathcal{L}_0(\tilde{R}^{(\bar{n}_\gamma)})\right) = o(\kappa_{\text{eff}})$$

Combining the displays yields

$$q_0(\delta^\gamma) \geq \left(\frac{c_H}{4}\right)^{\alpha/(\alpha-1)} \exp(-o(\kappa_{\text{eff}})) = \exp(-o(\kappa_{\text{eff}}))$$

equivalently $-\log q_0(\delta^\gamma) = o(\kappa_{\text{eff}})$. □

Corollary A.43. *Fix θ with $n_{\text{eff}}(\theta) > 0$ and consider a sequence of environments with $\gamma \downarrow 0$. Let δ^γ be any sequence of acceptance policies such that under the selected best response $|R| \leq m(\gamma)$ almost surely. Assume the extra information from selection is controlled in the sense of Appendix A.5.*

Fix any constant $c_H > 0$ with $\liminf_{\gamma \downarrow 0} q_H(\delta^\gamma) \geq c_H$ and define

$$\bar{n}_\gamma \equiv \left\lceil \frac{2}{\pi_H c_H \gamma} \right\rceil$$

If along the sequence $q_0(\delta^\gamma) = \exp(-\Theta(\kappa_{\text{eff}}))$, then necessarily $m(\gamma) = \Omega(1/\gamma)$.

Proof. Suppose for contradiction that along a subsequence $m(\gamma) = o(1/\gamma)$. Fix $c_H > 0$ and recall $\bar{n}_\gamma = \lceil 2/(\pi_H c_H \gamma) \rceil$ from Corollary A.43. Let τ^* denote the selected best-response stopping time under δ^γ and define

$$f \equiv \delta^\gamma(R) \mathbf{1}\{\tau^* \leq \bar{n}_\gamma\} \in [0, 1]$$

which is measurable with respect to the truncated report $\tilde{R}^{(\bar{n}_\gamma)}$. By Lemma A.13 and Markov's inequality,

$$\mathbb{P}_H(\tau^* > \bar{n}_\gamma) \leq \frac{\mathbb{E}_H[\tau^*]}{\bar{n}_\gamma} \leq \frac{1}{\pi_H \gamma \bar{n}_\gamma} \leq \frac{c_H}{2}$$

Therefore $\mathbb{E}_H[f] \geq q_H(\delta^\gamma) - \mathbb{P}_H(\tau^* > \bar{n}_\gamma) \geq c_H/2$.

Applying the Rényi expectation bound (Lemma A.30) with $P = \mathcal{L}_H(\tilde{R}^{(\bar{n}_\gamma)})$, $Q = \mathcal{L}_0(\tilde{R}^{(\bar{n}_\gamma)})$, and f yields

$$\mathbb{E}_0[f] \geq \mathbb{E}_H[f]^{\alpha/(\alpha-1)} \exp\left(-D_\alpha(\mathcal{L}_H(\tilde{R}^{(\bar{n}_\gamma)}) \parallel \mathcal{L}_0(\tilde{R}^{(\bar{n}_\gamma)}))\right)$$

Since $q_0(\delta^\gamma) = \mathbb{E}_0[\delta^\gamma(R)] \geq \mathbb{E}_0[f]$ and $\mathbb{E}_H[f] \geq c_H/2$, it remains to bound the divergence term.

Under the sufficient truncated-report budget in Appendix A.5.1 (Proposition A.39),

$$\begin{aligned} D_\alpha\left(\mathcal{L}_H(\tilde{R}^{(\bar{n}_\gamma)}) \parallel \mathcal{L}_0(\tilde{R}^{(\bar{n}_\gamma)})\right) &\leq (m(\gamma) + 1)d_\alpha(\theta) + \frac{1}{\alpha - 1} \log\left(\sum_{k=0}^{m(\gamma)+1} \binom{\bar{n}_\gamma + 1}{k}\right) \\ &\quad + \Lambda_{\text{sel}}(\bar{n}_\gamma, m(\gamma) + 1) \end{aligned}$$

Since $\bar{n}_\gamma \asymp 1/\gamma$ and $m(\gamma) = o(1/\gamma)$, the first term is $o(\kappa_{\text{eff}})$. Moreover, for all sufficiently small γ we have $m(\gamma) + 1 \leq (\bar{n}_\gamma + 1)/2$, so

$$\sum_{k=0}^{m(\gamma)+1} \binom{\bar{n}_\gamma + 1}{k} \leq \exp\left((m(\gamma) + 1) \log\left(\frac{e(\bar{n}_\gamma + 1)}{m(\gamma) + 1}\right)\right)$$

and the combinatorial term is also $o(\kappa_{\text{eff}})$. Finally, since $\bar{n}_\gamma \leq n_\gamma$ for all sufficiently small γ and $n \mapsto \Lambda_{\text{sel}}(n, k)$ is nondecreasing, Assumption A.38 implies

$$\Lambda_{\text{sel}}(\bar{n}_\gamma, m(\gamma) + 1) = o(\kappa_{\text{eff}})$$

Therefore,

$$q_0(\delta^\gamma) \geq \left(\frac{c_H}{2}\right)^{\alpha/(\alpha-1)} \exp(-o(\kappa_{\text{eff}})) = \exp(-o(\kappa_{\text{eff}}))$$

contradicting $q_0(\delta^\gamma) = \exp(-\Theta(\kappa_{\text{eff}}))$. □

A.6 Gaussian AR(1) computations

This appendix collects Gaussian AR(1) calculations used in the paper. These calculations support the running example in Sections 3 and 5 and provide the main verification routes for the auxiliary conditions used in the short-report and robustness-check arguments (Appendices A.5 and A.3). It records a belief-state route to controlling the extra information created by selection in the ordered three-type setting and verifies that the posterior tail term in the robustness-check bound (12) is exponentially small. Appendix A.1.4 records a primitive sufficient condition for the Bayes-factor tail assumptions in Assumptions A.6 and A.8 in contractive location AR(1) models. Here we verify its translate inequalities for the Gaussian running example (Lemma A.47) and also record a sharper posterior-tail bound based on the Gaussian belief state (Lemma A.49). Finally, it records the spectral-gap computation that underlies the effective-sample-size rate $n_{\text{eff}}(\theta) = 1 - \phi$ in this example (Lemma A.44).

A.6.1 A staged Gaussian example

We work in the Gaussian AR(1) mean-shift model of Example ???. Fix $\phi \in [0, 1)$ and a finite type set \mathcal{T} with type-specific means $(\mu_t)_{t \in \mathcal{T}}$. Conditional on type $T = t$, the latent score $(Z_n)_{n \geq 1}$ follows the stationary Gaussian AR(1)

$$Z_{n+1} = \phi Z_n + (1 - \phi)\mu_t + \varepsilon_{n+1} \quad \varepsilon_{n+1} \sim \mathcal{N}(0, 1 - \phi^2) \text{ i.i.d.}$$

so $Z_n \sim \mathcal{N}(\mu_t, 1)$ marginally, and the researcher observes one-sided p -values $P_n = 1 - \Phi(Z_n)$.

We often specialize to the ordered three-type case $\mathcal{T} = \{N, H, L\}$ with $\mathcal{H} = \{H\}$ and means $\mu_N < \mu_H < \mu_L$. In this ordering, extremely small p -values can be more indicative of L than H , so the one-test likelihood-ratio region can take a “window” form rather than a one-sided threshold.

Lemma A.44. *Fix $\phi \in [0, 1)$ and a type $t \in \mathcal{T}$ in Example ???. Under t , the stationary AR(1) chain (Z_n) is reversible with respect to its stationary law $\mathcal{N}(\mu_t, 1)$. Moreover, its L^2 spectral gap equals*

$$\text{gap}_{\text{ARI}}(\phi) = 1 - \phi$$

Proof. Let $X_n \equiv Z_n - \mu_t$. Then X_n is stationary $\mathcal{N}(0, 1)$ and satisfies

$$X_{n+1} = \phi X_n + \sqrt{1 - \phi^2} \xi_{n+1}, \quad \xi_{n+1} \sim \mathcal{N}(0, 1)$$

Thus (X_n, X_{n+1}) is bivariate normal with mean 0, unit variances, and correlation ϕ , hence its joint density is symmetric in (x, y) , implying detailed balance and reversibility.

For the spectral gap: the transition operator P acts on $L^2(\mathcal{N}(0, 1))$ by

$$Pf(x) = \mathbb{E}[f(\phi x + \sqrt{1 - \phi^2} \xi)]$$

The probabilists’ Hermite polynomials $(H_k)_{k \geq 0}$ form an orthogonal basis of $L^2(\mathcal{N}(0, 1))$. A standard property of Gaussian regression gives

$$\mathbb{E}[H_k(X_{n+1}) \mid X_n = x] = \phi^k H_k(x)$$

so each H_k is an eigenfunction with eigenvalue ϕ^k . The largest eigenvalue is 1 (for constants), and the second largest (in absolute value) is ϕ (for $k = 1$). Thus the L^2 spectral gap is $1 - \phi$. \square

In the Gaussian AR(1) mean-shift example (Example ??), the belief state for the researcher’s POMDP collapses to the one-dimensional sufficient statistic (n, s_n) ; see Appendix A.2, Section A.2.3. This scalar structure is a natural route to bounding extra information from selection in concrete model/policy classes: the only way the researcher can correlate endogenous stopping and selective disclosure with the unreported past is through the belief trajectory, and in Gaussian AR(1) that trajectory is Markov in a scalar statistic. The following subsections verify the auxiliary conditions used in the short-report and robustness-check arguments in the ordered three-type case. We first construct a calibrated one-test witness window (Lemma A.45), then record an exact monotone-shift property of Gaussian AR(1) (Proposition A.46) and use it to verify selection control and the marginal Rényi bound needed for the short-report budget (Proposition A.48). We finally record an explicit posterior-tail bound (Lemma A.49). The key structural requirement is that the endogenous transcript events are two-sided monotone (Definition A.11) in omitted coordinates conditional on the disclosed coordinates; this covers the common case where continuation is triggered by being in either tail of a scalar belief statistic.

A.6.2 A one-test Gaussian witness window

Lemma A.45. *Work in Example ???. Assume $\mathcal{H} = \{H\}$ is a singleton and denote the high-type mean by μ_H . Let $\Delta_{\min} \equiv \min_{t \in \mathcal{T}_0} |\mu_t - \mu_H|$ and assume $\Delta_{\min} > 0$. Define $\delta \equiv \Delta_{\min}/4$ and the (two-sided) witness window*

$$B_0 \equiv \{p \in (0, 1) : \Phi^{-1}(1 - p) \in [\mu_H - \delta, \mu_H + \delta]\}$$

Then Assumption A.2 holds with this B_0 , witness margin $\ell_0 = \Delta_{\min}^2/4$, and high one-test significance probability

$$p_H(B_0) = 2\Phi(\delta) - 1$$

Proof. Fix a non-high type with mean $\mu \neq \mu_H$. The normal density ratio satisfies

$$\frac{f_{\mu_H}(z)}{f_{\mu}(z)} = \exp\left((\mu_H - \mu)\left(z - \frac{\mu_H + \mu}{2}\right)\right)$$

On $z \in [\mu_H - \delta, \mu_H + \delta]$, the exponent is bounded below by $|\mu_H - \mu|(|\mu_H - \mu|/2 - \delta)$, hence by $\Delta_{\min}(\Delta_{\min}/2 - \Delta_{\min}/4) = \Delta_{\min}^2/4$. Mapping z to $p = 1 - \Phi(z)$ yields the stated bound on B_0 . Finally, under the high type $Z_1 \sim \mathcal{N}(\mu_H, 1)$, so $\mathbb{P}_H(P_1 \in B_0) = \mathbb{P}(|Z_1 - \mu_H| \leq \delta) = 2\Phi(\delta) - 1$. \square

A.6.3 Exact attractiveness and down-set ordering

Proposition A.46. *Work in Example ?? with fixed $\phi \in [0, 1)$. Fix two types $t_1, t_0 \in \mathcal{T}$ with $\mu_{t_1} > \mu_{t_0}$. Let \mathbb{P}_t denote the law conditional on $T = t$. Then for every $n \geq 1$ and every index set $J \subseteq \{1, \dots, n\}$, conditional on $Z_J = z_J$ the vector Z_{-J} has the same conditional covariance under t_1 and t_0 , and its conditional mean satisfies*

$$\mathbb{E}_{t_1}[Z_{-J} \mid Z_J = z_J] - \mathbb{E}_{t_0}[Z_{-J} \mid Z_J = z_J] \geq 0 \quad \text{coordinatewise for every } z_J$$

Consequently, for every down-set $D \subseteq \mathbb{R}^{|-J|}$ (Definition A.10),

$$\mathbb{P}_{t_1}(Z_{-J} \in D \mid Z_J = z_J) \leq \mathbb{P}_{t_0}(Z_{-J} \in D \mid Z_J = z_J) \quad \text{for all } z_J$$

Proof. Fix n and write Σ_n for the AR(1) correlation matrix with $(\Sigma_n)_{ij} = \phi^{|i-j|}$. Under type t , $Z_{1:n}$ is multivariate normal with mean $\mu_t \mathbf{1}_n$ and covariance Σ_n . Hence, conditional on $Z_J = z_J$, the conditional covariance of Z_{-J} is the same across components, and the conditional mean differs by

$$m_{t_1}(z_J) - m_{t_0}(z_J) = (\mu_{t_1} - \mu_{t_0})\left(\mathbf{1}_{-J} - \Sigma_{-J,J}\Sigma_{J,J}^{-1}\mathbf{1}_J\right)$$

where $m_t(z_J) \equiv \mathbb{E}_t[Z_{-J} \mid Z_J = z_J]$. Let $K_n \equiv \Sigma_n^{-1}$. A standard block-matrix identity yields

$$\mathbf{1}_{-J} - \Sigma_{-J,J}\Sigma_{J,J}^{-1}\mathbf{1}_J = K_{-J,-J}^{-1}(K_n \mathbf{1}_n)_{-J}$$

For AR(1) with $\phi \in [0, 1)$, K_n is the tridiagonal precision matrix

$$K_n = \frac{1}{1 - \phi^2} \begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 \\ -\phi & 1 + \phi^2 & -\phi & \ddots & \vdots \\ 0 & -\phi & 1 + \phi^2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\phi \\ 0 & \cdots & 0 & -\phi & 1 \end{pmatrix}$$

Its row sums are nonnegative, so $K_n \mathbf{1}_n \geq 0$ entrywise. Moreover, K_n is symmetric positive definite with nonpositive off-diagonal entries and is strictly diagonally dominant, hence it is a (symmetric) nonsingular M -matrix. In particular, every principal submatrix (including $K_{-J,-J}$) has a nonnegative inverse:

$$K_{-J,-J}^{-1} \geq 0 \quad \text{entrywise}$$

Therefore $K_{-J,-J}^{-1}(K_n \mathbf{1}_n)_{-J} \geq 0$ entrywise, and since $\mu_{t_1} - \mu_{t_0} > 0$ we have $m_{t_1}(z_J) - m_{t_0}(z_J) \geq 0$ coordinatewise for all z_J .

Finally, if X is a random vector and $c \geq 0$ is coordinatewise nonnegative, then for any down-set D we have $\{X + c \in D\} \subseteq \{X \in D\}$. Applying this with $X \equiv Z_{-J} - m_{t_0}(z_J)$ and $c \equiv m_{t_1}(z_J) - m_{t_0}(z_J)$ yields the down-set probability inequality. \square

A.6.4 A marginal Rényi bound

Lemma A.47. *Work in Example ?? with fixed $\phi \in [0, 1)$. Fix two types $t_1, t_0 \in \mathcal{T}$ with mean gap $\Delta \equiv \mu_{t_1} - \mu_{t_0}$ and fix $\alpha > 1$. Then for every finite index set $J = \{j_1 < \dots < j_k\} \subset \mathbb{N}$,*

$$D_\alpha(\mathcal{L}_{t_1}(Z_J) \parallel \mathcal{L}_{t_0}(Z_J)) \leq k \cdot \frac{\alpha}{2} \Delta^2 \cdot \frac{1 + \phi}{1 - \phi}$$

Proof. In the Gaussian AR(1) model, $\varepsilon_1 \sim \mathcal{N}(0, 1 - \phi^2)$. For Gaussian translates, the Rényi divergence satisfies

$$D_\alpha(\mathcal{N}(u, \sigma^2) \parallel \mathcal{N}(v, \sigma^2)) = \frac{\alpha}{2\sigma^2} (u - v)^2.$$

Thus Proposition A.9 applies with $v_\alpha(\theta) = 1 - \phi^2$, yielding

$$D_\alpha(\mathcal{L}_{t_1}(Z_J) \parallel \mathcal{L}_{t_0}(Z_J)) \leq |J| \cdot \frac{\alpha}{2(1 - \phi^2)} \Delta^2.$$

Since $(1 - \phi^2)^{-1} \leq (1 + \phi)/(1 - \phi)$, this implies the stated bound. \square

A.6.5 Verification of the selection-control term for short reports

This subsection verifies Assumption A.38 in the ordered three-type Gaussian AR(1) example by checking the ordered-sandwiching sufficient condition in Proposition A.41. The transcript-event condition (Assumption A.40) is natural in Gaussian AR(1) because the belief state is one-dimensional and continuation rules are typically window-based in a scalar belief statistic. The ordered-sandwiching inequality (14) follows from the exact attractiveness property of Gaussian AR(1) (Proposition A.46).

Proposition A.48. *Work in Example ?? with three types $\mathcal{T} = \{N, L, H\}$, $\mathcal{H} = \{H\}$, and means $\mu_N < \mu_H < \mu_L$, with fixed $\phi \in [0, 1)$. Fix $\alpha > 1$ and a horizon $n \geq 1$. Write $\pi_N^0 \equiv \pi_N/\pi_0$ and $\pi_L^0 \equiv \pi_L/\pi_0$ for the prior weights conditional on $T \in \mathcal{T}_0 = \{N, L\}$, and let $\pi_{\min}^0 \equiv \min\{\pi_N^0, \pi_L^0\}$.*

Assume an equilibrium in which $|R| \leq m$ almost surely for some $m \geq 0$. Consider the augmented truncated transcript $\hat{R}^{(n)}$ from Appendix A.5.1 with components $c \equiv (t, I)$ and $J(c) = I \cup \{t\}$. Assume that Assumption A.40 holds for this equilibrium.

Then the truncated-report Rényi budget of Theorem 4.5 holds with

$$d_\alpha(\theta) \equiv \frac{\alpha}{2} \max\{(\mu_H - \mu_N)^2, (\mu_H - \mu_L)^2\} \cdot \frac{1 + \phi}{1 - \phi}$$

up to an additive constant C_α depending only on α . Moreover, the extra-information-from-selection term in the theorem's budget can be taken to be the constant

$$\Lambda_{3\text{type}} \equiv \log \frac{1}{\pi_{\min}^0}$$

In particular, $\Lambda_{3\text{type}}$ is constant in n , so selection contributes only an $O(1)$ additive term on the κ_{eff} scale as $\gamma \downarrow 0$.

Proof. By Lemma A.47, for each $u \in \{N, L\}$ and each finite index set J ,

$$D_\alpha(\mathcal{L}_H(Z_J) \parallel \mathcal{L}_u(Z_J)) \leq |J| \cdot \frac{\alpha}{2} (\mu_H - \mu_u)^2 \cdot \frac{1 + \phi}{1 - \phi}$$

Thus the marginal condition in Proposition A.41 holds with

$$d_\alpha(\theta) \equiv \frac{\alpha}{2} \max\{(\mu_H - \mu_N)^2, (\mu_H - \mu_L)^2\} \cdot \frac{1 + \phi}{1 - \phi}$$

Proposition A.46 verifies the conditional down-set ordering required by Proposition A.41 (since $\mu_N < \mu_H < \mu_L$). In particular, for every J , every z_J , and every down-set D ,

$$\mathbb{P}_L(Z_{-J} \in D \mid Z_J = z_J) \leq \mathbb{P}_H(Z_{-J} \in D \mid Z_J = z_J) \leq \mathbb{P}_N(Z_{-J} \in D \mid Z_J = z_J)$$

which is (14) with $t^- = N$ and $t^+ = L$. Applying Proposition A.41 then yields the truncated-report budget with selection term $\Lambda_{\text{3type}} = \log(1/\pi_{\min}^0)$, up to an additive constant depending only on α . \square

A.6.6 Posterior tails

Lemma A.49. *Work in Example ?? with three types $\mathcal{T} = \{N, H, L\}$ and means $\mu_N < \mu_H < \mu_L$, and fix $\phi \in [0, 1)$. Let $\pi_n(H) = \mathbb{P}(T \in \mathcal{H} \mid \mathcal{F}_n)$ be the posterior probability from Lemma A.24. Fix $\alpha > 1$ and write $\lambda \equiv (\alpha - 1)/\alpha$. Define the minimum mean gap to a non-high type by*

$$\Delta_{\min} \equiv \min\{\mu_H - \mu_N, \mu_L - \mu_H\}$$

Then for every $n \geq 1$ and every $\varepsilon \in (0, 1)$,

$$\mathbb{P}_0(\pi_n(H) > \varepsilon) \leq C_{\pi, \alpha} \varepsilon^{-\lambda} \exp\left(-\frac{\lambda}{2\alpha} \Delta_{\min}^2 b_n\right)$$

where $C_{\pi, \alpha}$ is as in Assumption A.8 and $b_n = \mathbf{1}^\top \Sigma_\phi^{-1} \mathbf{1}$ is as in the discussion of the Gaussian AR(1) belief state (A.2.3). In particular, since $b_n \geq (1 - \phi)(n - 1)/(1 + \phi)$, Assumption A.8 holds in this example with $n_{\text{eff}}(\theta) = 1 - \phi$ and

$$C_\alpha(\theta) = \frac{\Delta_{\min}^2}{2\alpha(1 + \phi)}$$

Consequently, if $n = m(\gamma; c) = \lceil c/\gamma \rceil$ and $\varepsilon = \gamma/(2a)$ with $a \in [\rho, 1]$ and $\log(1/\gamma) = o(\kappa_{\text{eff}})$, then

$$\mathbb{P}_0\left(\pi_{m(\gamma; c)}(H) > \frac{\gamma}{2a}\right) = \exp(-\Omega(\kappa_{\text{eff}}))$$

Proof. Fix $n \geq 1$ and write $p_u(z_{1:n})$ for the joint density of $Z_{1:n}$ under type $u \in \{N, H, L\}$. By Bayes' rule,

$$\pi_n(H) = \frac{\pi_H p_H(Z_{1:n})}{\pi_H p_H(Z_{1:n}) + \pi_N p_N(Z_{1:n}) + \pi_L p_L(Z_{1:n})}.$$

Hence for each $u \in \{N, L\}$,

$$\pi_n(H) \leq \frac{\pi_H p_H(Z_{1:n})}{\pi_u p_u(Z_{1:n})} = \frac{\pi_H}{\pi_u} \cdot \frac{p_H(Z_{1:n})}{p_u(Z_{1:n})},$$

so

$$\{\pi_n(H) > \varepsilon\} \subseteq \left\{ \frac{p_H(Z_{1:n})}{p_u(Z_{1:n})} > \frac{\pi_u}{\pi_H} \varepsilon \right\}.$$

With $\lambda = (\alpha - 1)/\alpha \in (0, 1)$, Markov's inequality gives

$$\mathbb{P}_u(\pi_n(H) > \varepsilon) \leq \left(\frac{\pi_H}{\pi_u \varepsilon}\right)^\lambda \mathbb{E}_u \left[\left(\frac{p_H(Z_{1:n})}{p_u(Z_{1:n})} \right)^\lambda \right].$$

We control the λ -moment of the Bayes factor by factoring it into a marginal term and one-step transition terms. Write $\Delta_u \equiv |\mu_H - \mu_u|$. Under type u , the marginal law is $Z_1 \sim \mathcal{N}(\mu_u, 1)$ and the one-step transition is

$$Z_{t+1} | Z_t \sim \mathcal{N}(\phi Z_t + (1 - \phi)\mu_u, 1 - \phi^2) \quad (t \geq 1).$$

Therefore

$$\frac{p_H(Z_{1:n})}{p_u(Z_{1:n})} = \frac{p_H(Z_1)}{p_u(Z_1)} \cdot \prod_{t=1}^{n-1} \frac{p_H(Z_{t+1} | Z_t)}{p_u(Z_{t+1} | Z_t)}.$$

For equal-variance Gaussian translates, the Chernoff coefficient satisfies

$$\mathbb{E}_{X \sim \mathcal{N}(v, \sigma^2)} \left[\left(\frac{d\mathcal{N}(u, \sigma^2)}{d\mathcal{N}(v, \sigma^2)}(X) \right)^\lambda \right] = \exp\left(-\frac{1}{2}\lambda(1-\lambda)\frac{(u-v)^2}{\sigma^2}\right).$$

Applying this to the marginal term ($\sigma^2 = 1$) gives

$$\mathbb{E}_u \left[\left(\frac{p_H(Z_1)}{p_u(Z_1)} \right)^\lambda \right] = \exp\left(-\frac{\lambda}{2\alpha} \Delta_u^2\right),$$

and applying it to each transition term ($\sigma^2 = 1 - \phi^2$ and mean shift $(1 - \phi)\Delta_u$) gives, for every t ,

$$\mathbb{E}_u \left[\left(\frac{p_H(Z_{t+1} | Z_t)}{p_u(Z_{t+1} | Z_t)} \right)^\lambda \right] = \exp\left(-\frac{\lambda}{2\alpha} \Delta_u^2 \cdot \frac{1 - \phi}{1 + \phi}\right).$$

Moreover, the displayed transition moment is constant (it does not depend on Z_t), so iterating conditional expectations yields

$$\mathbb{E}_u \left[\left(\frac{p_H(Z_{1:n})}{p_u(Z_{1:n})} \right)^\lambda \right] = \exp\left(-\frac{\lambda}{2\alpha} \Delta_u^2 \left(1 + (n-1)\frac{1-\phi}{1+\phi}\right)\right).$$

The AR(1) identity $b_n = \frac{n(1-\phi)+2\phi}{1+\phi} = 1 + (n-1)\frac{1-\phi}{1+\phi}$ (see (A.2.3)) gives

$$\mathbb{E}_u \left[\left(\frac{p_H(Z_{1:n})}{p_u(Z_{1:n})} \right)^\lambda \right] = \exp\left(-\frac{\lambda}{2\alpha} \Delta_u^2 b_n\right).$$

Substitute into the Markov bound to obtain, for $u \in \{N, L\}$,

$$\mathbb{P}_u(\pi_n(H) > \varepsilon) \leq \left(\frac{\pi_H}{\pi_u \varepsilon}\right)^\lambda \exp\left(-\frac{\lambda}{2\alpha} \Delta_u^2 b_n\right).$$

Since \mathbb{P}_0 is a mixture over $\{N, L\}$, a union bound yields

$$\mathbb{P}_0(\pi_n(H) > \varepsilon) \leq \mathbb{P}_N(\pi_n(H) > \varepsilon) + \mathbb{P}_L(\pi_n(H) > \varepsilon).$$

Using $\Delta_u \geq \Delta_{\min}$ and $\pi_u \geq \pi_0 \pi_{\min}^0$ for $u \in \{N, L\}$ yields the displayed bound in the lemma statement.

Finally, the explicit formula for b_n implies $b_n \geq (1 - \phi)(n - 1)/(1 + \phi)$. With $n = m(\gamma; c) = \lceil c/\gamma \rceil$, this gives $b_n = \Omega(\kappa_{\text{eff}})$. If $\varepsilon = \gamma/(2a)$ with $a \in [\rho, 1]$ and $\log(1/\gamma) = o(\kappa_{\text{eff}})$, then $\varepsilon^{-\lambda} = (2a/\gamma)^\lambda = \exp(o(\kappa_{\text{eff}}))$, and the exponential term is $\exp(-\Omega(\kappa_{\text{eff}}))$. \square

A.7 Generalizations

Theorems 4.3, 4.5, and 5.2 are stated under Assumption A.1 because it is a convenient baseline that directly covers the dependence structures emphasized in the main text. We prove more general versions of these results under weaker, modular conditions that separate (i) information accumulation along the evidence stream, (ii) the informational content of selectively disclosed reports, and (iii) concentration of significant-result counts under dependence. This modular view makes explicit which pieces of the standing assumptions (Appendix A.1.1) and which stopping-time tools (Appendix A.2) are actually used by each argument.

A.7.1 General lower bound beyond Assumption A.1

The proof of Theorem 4.3 uses only two ingredients. First, the acceptance decision is generated from the report by a Markov kernel, so data processing gives

$$\text{KL}(\mathcal{L}_H(A) \parallel \mathcal{L}_0(A)) \leq \text{KL}(\mathcal{L}_H(R) \parallel \mathcal{L}_0(R))$$

Second, the researcher faces per-test cost $\gamma > 0$ and acceptance probabilities lie in $[0, 1]$, so Proposition A.16 yields a universal effort bound $\mathbb{E}[\tau^*(\delta)] \leq 1/\gamma$. Thus the lower bound holds under any environment in which the stopped reported history admits a KL budget of the form

$$\text{KL}(\mathcal{L}_H(R) \parallel \mathcal{L}_0(R)) \leq C_0(\theta) + D(\theta)\mathbb{E}[\tau^*(\delta)]$$

for some finite constants $C_0(\theta)$ and $D(\theta)$ that are uniform over mechanisms. Assumption A.1 is one sufficient condition with $C_0(\theta) \equiv C_\nu(\theta)$ and $D(\theta) \equiv D_{\text{mix}}(\theta)$.

A.7.2 General sublinear impossibility beyond Assumption A.1

Theorem 4.5 isolates the additional ingredient needed to make a short-report impossibility quantitative under dependence and selective disclosure. It combines a mechanism-independent change-of-measure bound (Appendix A.5.3) with a modular short-report (truncated-report) Rényi budget (Appendix A.5.1) of the form

$$D_\alpha(\mathcal{L}_H(\tilde{R}^{(n)}) \parallel \mathcal{L}_0(\tilde{R}^{(n)})) \leq (m+1)d_\alpha(\theta) + \frac{1}{\alpha-1} \log \left(\sum_{k=0}^{m+1} \binom{n+1}{k} \right) + \Lambda_{\text{sel}}(n, m+1) + O(1),$$

where m is an almost-sure bound on the report size, $d_\alpha(\theta)$ is a marginal Rényi rate, and Λ_{sel} is an extra-information-from-selection term defined from the induced transcript. Together with the Rényi expectation inequality (Lemma A.30) and the universal time bound (Lemma A.13), this budget implies that if the truncated-report divergence is $o(\kappa_{\text{eff}})$ then achieving $q_0 = \exp(-\Theta(\kappa_{\text{eff}}))$ forces throughput to collapse ($q_H \rightarrow 0$). In particular, when q_H is bounded away from zero, Corollary A.43 truncates at $\bar{n}_\gamma \asymp 1/\gamma$ and shows that under $m(\gamma) = o(1/\gamma)$ and $\gamma \Lambda_{\text{sel}}(\bar{n}_\gamma, m(\gamma) + 1) \rightarrow 0$, short disclosure cannot support $\exp(-\Theta(\kappa_{\text{eff}}))$ false-positive decay at nonvanishing q_H .

A.7.3 General robustness-check achievability beyond Assumption A.1

Theorem 5.2 requires only that, under every non-high type, the count of significant results in B ,

$$S_n(B) \equiv \sum_{i=1}^n \mathbf{1}\{P_i \in B\}$$

satisfies a large-deviation bound at exponent proportional to an effective sample size index. A sufficient condition is a spectral-gap or mixing-time concentration inequality for bounded additive functionals of the evidence process. Assumption A.1(iii) is one convenient way to package this requirement, but the same conclusion holds under alternative mixing assumptions that deliver an inequality of the form

$$\mathbb{P}_t(S_n(B) \geq rn) \leq \exp(-c n_{\text{eff}}(\theta) n)$$

uniformly over non-high types whenever r exceeds the non-high mean significance rate by a fixed slack. When $n \equiv n_\gamma \asymp 1/\gamma$, this yields a robustness-check upper bound of order $\exp(-\Omega(\kappa_{\text{eff}}))$ with $\kappa_{\text{eff}} = n_{\text{eff}}(\theta)/\gamma$.

B Empirical Appendix

This appendix documents the data construction, estimation, and robustness analyses underlying Section 6. The organizing principle is auditability. We specify every empirical object targeted, every filtering rule defining the verified core, and the exact mapping from raw replication artifacts to the mixture, dependence, and counterfactual inputs used in the paper. The companion repository (<https://github.com/gsekeres/agent-specification-search/>) contains the version-controlled surface definitions, prompts, validators, and per-paper artifacts referenced throughout; the appendix is written so that each step can be verified by inspecting files on disk.

Throughout this appendix we label the three mixture types N (null), M (moderate), and E (extreme). Our baseline mixture fixes $\sigma = 1$ (unit variance). The evidence index is $|t|$ (the absolute t -statistic). The evidence window is $B = [1.96, \infty)$, and the false discovery rate is defined with respect to the null type only: $\text{FDR} = \pi_N Q_N(m) / \bar{Q}(m)$, treating extreme-type papers as true positives. The baseline counterfactual fixes $m^{\text{old}} = 50$ (the median of the author-reported regressions; see Section B.2.7) and calibrates λ from automated-workflow timing data.

B.1 Pipeline and workflow

This section documents the pipeline used to generate the specification-level data underlying Section 6. The organizing principle is mechanical auditability. A reader should be able to determine, from files on disk, the universe of specifications we committed to before results were observed, the numerical outputs produced by executing that universe, and the filtering and labeling steps that define the conservative verified core used in estimation. We therefore separate definition of the universe from its execution and verification. We also separate agentic coordination from numerical computation: language models are used to read replication packages, draft and edit code, and produce structured configuration artifacts, while all estimation and inference are carried out by explicit scripts in Python and R.

B.1.1 Surface definition, execution, and verification

The pipeline consists of a version-controlled definition layer and a paper-specific execution layer. The definition layer lives in `specification_tree/`, `prompts/`, and `scripts/`. The paper-specific layer lives in the extracted replication package directory `data/downloads/extracted/{PAPER_ID}/`. Post-run verification artifacts live in `data/verification/{PAPER_ID}/`. The estimation layer lives in `estimation/`.

A key contract is that the definition layer is fixed within a run. The typed surface definition, prompts, and validators do not evolve as the runner encounters new packages. When we discover a missing but legitimate specification family or a recurring resolution rule, it is incorporated only as an explicit, version-controlled change between runs. This ensures that the information available to the executor is stable within a run and that improvements to coverage are attributable to specific commits.

B.1.2 Typed namespaces and execution identifiers

The typed namespace defines the object space and the meaning of each executed row. Every executed object has a typed identifier `spec.id`. The top-level namespace determines how the object is interpreted and where it is allowed to appear. Estimate-like rows live in `baseline`, `design/*`, and `rc/*`, and these are the only rows permitted in `specification_results.csv`. Inference-only recomputations are typed as `infer/*` and are recorded separately in `inference_results.csv` when

requested. Diagnostics, sensitivity objects, post-processing transforms, and concept changes are typed as `diag/*`, `sens/*`, `post/*`, and `explore/*` and are recorded, when executed, in separate tables rather than being mixed into the estimate table. This typing discipline is enforced by an explicit contract for table schemas and JSON payloads, defined in `specification_tree/CONTRACT.md`.

Each estimate-like row carries a stable identifier `spec_run_id` and a `baseline_group_id` that links the row to a baseline claim object defined pre-run. Each row also records a pointer field, `spec_tree_path`, that references a specific admissible object in the typed namespace. The contract requires that the runner store full model output and audit metadata in `coefficient_vector_json`, using reserved keys that keep the payload mechanically parseable across heterogeneous designs. For successful estimate-like rows these reserved blocks include the full coefficient vector, an inference description for the scalar uncertainty reported in the row, a software-environment block, and a deterministic hash of the pre-run surface used for the run.

Runner scripts import shared output helpers in `scripts/agent_output_utils.py` that implement the reserved-key JSON schema, deterministic surface hashing, and standardized software blocks, reducing paper-by-paper drift in how outputs are recorded.

The contract is not a narrative description; it is a mechanically checked interface. The validator `scripts/validate_agent_outputs.py` checks required columns, namespace rules, existence and validity of `spec_tree_path` pointers, surface membership of `spec_id` patterns at the baseline-group level, and consistency of the recorded surface hash. A normalization script `scripts/normalize_agent_outputs.py` exists to migrate legacy outputs into the reserved-key JSON schema when needed.

B.1.3 Specification surface and revealed search space

The most consequential paper-specific object is the specification surface. The surface is a machine-readable contract written to `SPECIFICATION_SURFACE.json` and an accompanying human-readable summary `SPECIFICATION_SURFACE.md`. The surface is constructed and reviewed before any models are run. It defines baseline claim objects (baseline groups), records the paper’s canonical baseline specifications, and commits ex ante to an executable universe of estimand-preserving variants together with a canonical inference choice for estimate rows. It also encodes the constraints, budgets, and seeded sampling rules that make large search spaces auditable.

Our surface is keyed to the paper’s revealed search space. The revealed surface is the minimum set of forks a reader can confirm the researcher had to navigate given what is disclosed in the main text and interpreted appendices. This is intentionally more conservative than the potential search space of all imaginable variants. The surface can include standardized stress tests beyond what a paper reveals, but those are treated as explicit expansions and are typed and labeled so that the distinction between disclosure-implied forks and protocol-added stress tests remains auditably sharp.

Two surface constraints illustrate the role of the revealed surface. First, control inclusion is a central axis that is often combinatorial. When the surface includes control-subset sampling, it bounds subset sizes by the control-count envelope revealed by the paper’s own main specifications and uses a seed recorded in the surface to generate a reproducible set of draws. Second, many estimators are bundles with multiple components, such as IV or doubly robust and DML procedures. For such bundles, the surface records whether covariate adjustment is linked across components and enforces joint variation when it is linked, preventing the runner from inventing a cross-product space that the paper did not reveal.

B.1.4 Verification layers and guardrails

Verification occurs in layers, with each layer producing an explicit artifact. Before any models are run, a surface builder agent, prompted by `prompts/03_spec_surface_builder.md`, produces a candidate surface. A separate surface verifier, prompted by `prompts/04_spec_surface_verifier.md`, then reviews and may edit `SPECIFICATION_SURFACE.json` before any execution. The verifier records its reasoning and any edits in `SPEC_SURFACE_REVIEW.md`. This stage concentrates discretionary judgment into a single pre-run object that is easy to audit: the approved surface is the definition of the executable universe for that paper.

Execution is then driven entirely by the approved surface. The runner agent, prompted by `prompts/05_spec_searcher.md`, executes only whitelisted specifications and saves its executable script to `scripts/paper_analyses/{PAPER_ID}.py`. It writes `specification_results.csv` and a narrative log `SPECIFICATION_SEARCH.md` to the extracted package directory. Every planned specification appears as a row, including explicit failure rows, so that the realized output can be audited against the pre-run budget. When the surface requests inference variants, the runner writes the corresponding typed side table rather than mixing those recomputations into the estimate table.

After execution, the pipeline enforces the contract mechanically by running `scripts/validate_agent_outputs.py`. This validator checks schema and typing invariants that are easy for humans to miss in large runs. Post-run verification then audits the realized outputs without running new regressions. The post-run verifier, prompted by `prompts/06_post_run_verifier.md`, checks for drift and incoherence and assigns conservative labels that govern downstream estimation. Verification artifacts are written to `data/verification/{PAPER_ID}/`. The key output is `verification_spec_map.csv`, which maps `spec_run_id` to a validity indicator and a core-test indicator, along with a category label and a short justification. The verifier also records baseline specifications in `verification_baselines.json` and writes a narrative `VERIFICATION_REPORT.md` summarizing issues.

The verified core used in all downstream estimation is defined as the subset of estimate-like rows that are valid and labeled as estimand-preserving core tests for a baseline group. This core is intentionally conservative. When the verifier cannot confirm that a row preserves the claim object implied by the surface, it is excluded from the core by default.

B.1.5 Estimation pipeline

The estimation pipeline lives in `estimation/` and is driven by `python estimation/run_all.py`. It ingests per-paper execution outputs and verification maps, builds unified datasets and their verified and verified-core subsets in `estimation/data/`, fits the evidence mixture model and estimates within-paper dependence, and computes counterfactual disclosure requirements under a cost shift, writing machine-readable outputs to `estimation/results/`. For the paired-benchmark exercise in Sample A, the pipeline constructs an auditable mapping from benchmark claims to baseline groups in `estimation/data/i4r_claim_map.csv` and uses that map to define matched reproductions. Finally, the pipeline renders manuscript-ready tables and figures into the manuscript directories.

B.2 Samples, inference, and validation

B.2.1 Sample A: paired replications (independent benchmark)

Sample A is built from the AEA-journal replication exercise in [Brodeur et al. \(2026\)](#). For each paper we target the same canonical claim emphasized in the independent reanalysis protocol and record three claim-level statistics: the published (original-study) canonical claim statistic t_i^{orig} , the

independent re-analysis statistic t_i^{ind} , and our automated baseline reproduction t_i^{auto} for the same claim. All objects are recorded as absolute t -statistics $|t|$ and as the derived evidence indices defined in Section B.2.8.

Mapping from the independent reanalysis protocol to our automated specification surface is nontrivial: each paper’s replication package has a different structure, and the canonical claim may correspond to a specific coefficient in a specific table. We resolve this mapping using a combination of verification baseline group labels (from the post-run verification agent), token-overlap scoring between the independent reanalysis claim description and our baseline group definitions, and manual overrides where the automated mapping is flagged for review. The mapping is recorded in `estimation/data/i4r_claim_map.csv` and is fully auditable.

B.2.2 Sample B: surface-defined specification sets (post-automation regime)

Sample B consists of 103 papers for which we can ingest the public replication package and construct a standardized specification surface. This includes the 40 independent reanalysis papers in Sample A and an additional 63 papers with public data and code. For each paper the automated workflow executes the surface-defined robustness universe around a baseline claim object, spanning method variations, inference choices, sample restrictions, and covariate sets.

The full sample contains 5,793 specifications across 97 papers after filtering invalid rows (mean 59.7 specs/paper). The verified-core subset contains 5,569 specifications across 96 papers (mean 58.0 core specs/paper).

B.2.3 Sample B paper citations

Table 1 lists the 103 papers in Sample B and cites each source paper. The bibliography entries are generated from the AEA metadata record linked to each openICPSR replication package.

Table 1: Sample B papers with bibliography entries. The table lists the 103 AEA-journal papers for which the pipeline constructs a standardized specification surface. Each citation appears in the bibliography through the table entry.

Paper ID	Citation and title	Journal	Year
111185-V1	Rudik (2020), <i>Optimal Climate Policy When Damages are Unknown</i>	AEJ: Policy	2020
112338-V1	Duggan and Morton (2010), <i>The Effect of Medicare Part D on Pharmaceutical Prices and Utilization</i>	AER	2010
112370-V1	Chang (2010), <i>Elections, Capital Flows, and Politico-Economic Equilibria</i>	AER	2010
112415-V1	Reutskaja et al. (2011), <i>Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study</i>	AER	2011
112431-V1	Ferraz and Finan (2011), <i>Electoral Accountability and Corruption: Evidence from the Audits of Local Governments</i>	AER	2011
112444-V1	Reinhart and Rogoff (2011), <i>From Financial Crash to Debt Crisis</i>	AER	2011
112451-V1	Furman and Stern (2011), <i>Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research</i>	AER	2011
112466-V1	Xiong and Yu (2011), <i>The Chinese Warrants Bubble</i>	AER	2011
112474-V1	Dinkelman (2011), <i>The Effects of Rural Electrification on Employment: New Evidence from South Africa</i>	AER	2011
112498-V1	Mazzocco and Saini (2012), <i>Testing Efficient Risk Sharing with Heterogeneous Risk Preferences</i>	AER	2012
112517-V1	Fowle et al. (2012), <i>What Do Emissions Markets Deliver and to Whom? Evidence from Southern California’s NOx Trading Program</i>	AER	2012
112574-V1	Faye and Niehaus (2012), <i>Political Aid Cycles</i>	AER	2012
112587-V1	Einav et al. (2013), <i>Selection on Moral Hazard in Health Insurance</i>	AER	2013
112746-V1	Dubois et al. (2014), <i>Do Prices and Attributes Explain International Differences in Food Purchases?</i>	AER	2014
112749-V1	Hornbeck and Naidu (2014), <i>When the Levee Breaks: Black Migration and Economic Development in the American South</i>	AER	2014

continued on next page

Paper ID	Citation and title	Journal	Year
112756-V1	Martinez-Bravo (2014), <i>The Role of Local Officials in New Democracies: Evidence from Indonesia</i>	AER	2014
112785-V1	Hlathswayo and Spence (2014), <i>Demand and Defective Growth Patterns: The Role of The Tradable and Non-Tradable Sectors in an Open Economy</i>	AER	2014
112791-V1	Baicker et al. (2014), <i>The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment</i>	AER	2014
112805-V1	Deming (2014), <i>Using School Choice Lotteries to Test Measures of School Effectiveness</i>	AER	2014
112815-V1	Hoxby (2014), <i>The Economics of Online Postsecondary Education: MOOCs, Nonselective Education, and Highly Selective Education</i>	AER	2014
112840-V1	Azzimonti et al. (2014), <i>Financial Globalization, Inequality, and the Rising Public Debt</i>	AER	2014
112853-V1	Manuelli and Seshadri (2014), <i>Human Capital and the Wealth of Nations</i>	AER	2014
112908-V1	Gowrisankaran et al. (2015), <i>Mergers When Prices Are Negotiated: Evidence from the Hospital Industry</i>	AER	2015
113046-V1	Romer and Romer (2017), <i>New Evidence on the Aftermath of Financial Crises in Advanced Countries</i>	AER	2017
113066-V1	Steinwender (2018), <i>Real Effects of Information Frictions: When the States and the Kingdom Became United</i>	AER	2018
113109-V1	Charles et al. (2018), <i>Housing Booms and Busts, Labor Market Opportunities, and College Attendance</i>	AER	2018
113182-V1	Condra et al. (2018), <i>The Logic of Insurgent Electoral Violence</i>	AER	2018
113229-V1	Bohnet et al. (2008), <i>Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States</i>	AER	2008
113366-V1	Bajari et al. (2015), <i>Machine Learning Methods for Demand Estimation</i>	AER	2015
113407-V1	Bitler and Hoynes (2015), <i>Living Arrangements, Doubling Up, and the Great Recession: Was This Time Different?</i>	AER	2015
113500-V1	Babcock et al. (2017), <i>Gender Differences in the Allocation of Low-Promotability Tasks: The Role of Backlash</i>	AER	2017
113513-V1	Stock (2017), <i>Trends in Economics and Other Undergraduate Majors</i>	AER	2017
113517-V1	Moscarini and Postel-Vinay (2017), <i>The Relative Power of Employment-to-Employment Reallocation and Unemployment Exits in Predicting Wage Growth</i>	AER	2017
113561-V1	Fong and Luttmer (2009), <i>What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty</i>	AEJ: Applied	2009
113566-V1	Jacob and Lefgren (2009), <i>The Effect of Grade Retention on High School Completion</i>	AEJ: Applied	2009
113577-V1	Jackson and Bruegmann (2009), <i>Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers</i>	AEJ: Applied	2009
113592-V1	Fack and Grenet (2015), <i>Improving College Access and Success for Low-Income Students: Evidence from a Large Need-Based Grant Program</i>	AEJ: Applied	2015
113630-V1	Nyqvist et al. (2017), <i>Experimental Evidence on the Long-Run Impact of Community-Based Monitoring</i>	AEJ: Applied	2017
113684-V1	Miller (2017), <i>The Persistent Effect of Temporary Affirmative Action</i>	AEJ: Applied	2017
113744-V1	Wisdom et al. (2010), <i>Promoting Healthy Choices: Information versus Convenience</i>	AEJ: Applied	2010
113812-V1	Stange (2012), <i>An Empirical Investigation of the Option Value of College Enrollment</i>	AEJ: Applied	2012
113893-V1	DellaVigna et al. (2014), <i>Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia</i>	AEJ: Applied	2014
114098-V1	Eden (2016), <i>Excessive Financing Costs in a Representative Agent Framework</i>	AEJ: Macro	2016
114295-V1	Charnavoki and Dolado (2014), <i>The Effects of Global Shocks on Small Commodity-Exporting Economies: Lessons from Canada</i>	AEJ: Macro	2014
114542-V1	Cattaneo et al. (2009), <i>Housing, Health, and Happiness</i>	AEJ: Policy	2009
114633-V1	Lawson (2017), <i>Liquidity Constraints, Fiscal Externalities and Optimal Tuition Subsidies</i>	AEJ: Policy	2017
114747-V1	David et al. (2010), <i>The Effects of Pharmaceutical Marketing and Promotion on Adverse Drug Events and Regulation</i>	AEJ: Policy	2010
114759-V1	Camacho and Conover (2011), <i>Manipulation of Social Program Eligibility</i>	AEJ: Policy	2011
114828-V1	Grosfeld et al. (2013), <i>Persistent Antimarket Culture: A Legacy of the Pale of Settlement after the Holocaust</i>	AEJ: Policy	2013
114833-V1	Allcott (2013), <i>The Welfare Effects of Misperceived Product Costs: Data and Calibrations from the Automobile Market</i>	AEJ: Policy	2013

continued on next page

Paper ID	Citation and title	Journal	Year
114875-V1	Bronzini and Iachini (2014) , <i>Are Incentives for R&D Effective? Evidence from a Regression Discontinuity Approach</i>	AEJ: Policy	2014
114879-V1	Davis et al. (2014) , <i>Cash for Coolers: Evaluating a Large-Scale Appliance Replacement Program in Mexico</i>	AEJ: Policy	2014
116031-V1	Neary (2004) , <i>Rationalizing the Penn World Table: True Multilateral Indices for International Comparisons of Real Income</i>	AER	2004
116063-V1	Dafny (2005) , <i>How Do Hospitals Respond to Price Changes?</i>	AER	2005
116139-V1	Kosfeld and Rustagi (2015) , <i>Leader Punishment and Cooperation in Groups: Experimental Field Evidence from Commons Management in Ethiopia</i>	AER	2015
116164-V1	Lockwood (2018) , <i>Incidental Bequests and the Choice to Self-Insure Late-Life Risks</i>	AER	2018
116167-V1	Durante et al. (2019) , <i>The Political Legacy of Entertainment TV</i>	AER	2019
116224-V1	Kuruscu (2006) , <i>Training and Lifetime Income</i>	AER	2006
116248-V1	Costa-Gomes and Crawford (2006) , <i>Cognition and Behavior in Two-Person Guessing Games: An Experimental Study</i>	AER	2006
116280-V1	Zanjani (2007) , <i>Regulation, Capital, and the Evolution of Organizational Form in US Life Insurance</i>	AER	2007
116531-V1	Marx and Turner (2019) , <i>Student Loan Nudges: Experimental Evidence on Borrowing and Educational Attainment</i>	AEJ: Policy	2019
120078-V1	Laouénan and Rathelot (2022) , <i>Can Information Reduce Ethnic Discrimination? Evidence from Airbnb</i>	AEJ: Applied	2022
120483-V1	Esposito (2022) , <i>The Side Effects of Immunity: Malaria and African Slavery in the United States</i>	AEJ: Applied	2022
120568-V1	Pries and Rogerson (2022) , <i>Declining Worker Turnover: The Role of Short-Duration Employment Spells</i>	AEJ: Macro	2022
125321-V1	Greenstone et al. (2022) , <i>Can Technology Solve the Principal-Agent Problem? Evidence from China's War on Air Pollution</i>	AER: Insights	2022
125821-V1	Baron (2022) , <i>School Spending and Student Outcomes: Evidence from Revenue Limit Elections in Wisconsin</i>	AEJ: Policy	2022
126722-V1	Lopez et al. (2022) , <i>Does Patient Demand Contribute to the Overuse of Prescription Drugs?</i>	AEJ: Applied	2022
128143-V1	Douenne and Fabre (2022) , <i>Yellow Vests, Pessimistic Beliefs, and Carbon Tax Aversion</i>	AEJ: Policy	2022
128521-V1	Arthi et al. (2022) , <i>Recessions, Mortality, and Migration Bias: Evidence from the Lancashire Cotton Famine</i>	AEJ: Applied	2022
130141-V1	Görtz et al. (2022) , <i>News Shocks under Financial Frictions</i>	AEJ: Macro	2022
130784-V1	Wilson (2022) , <i>Child Marriage Bans and Female Schooling and Labor Market Outcomes: Evidence from Natural Experiments in 17 Low- and Middle-Income Countries</i>	AEJ: Policy	2022
131981-V1	Altindag et al. (2022) , <i>Mental Health Costs of Lockdowns: Evidence from Age-Specific Curfews in Turkey</i>	AEJ: Applied	2022
133501-V1	Huh and Reif (2021) , <i>Teenage Driving, Mortality, and Risky Behaviors</i>	AER: Insights	2021
134041-V1	Settele (2022) , <i>How Do Beliefs about the Gender Wage Gap Affect the Demand for Public Policy?</i>	AEJ: Policy	2022
134622-V1	Azoulay et al. (2022) , <i>Immigration and Entrepreneurship in the United States</i>	AER: Insights	2022
136741-V1	Williams (2022) , <i>Historical Lynchings and the Contemporary Voting Behavior of Blacks</i>	AEJ: Applied	2022
138401-V1	Atwood (2022) , <i>The Long-Term Effects of Measles Vaccination on Earnings and Employment</i>	AEJ: Policy	2022
138922-V1	Marcus et al. (2022) , <i>The Long-Run Effects of Sports Club Vouchers for Primary School Children</i>	AEJ: Policy	2022
139262-V1	Drobner (2022) , <i>Motivated Beliefs and Anticipation of Uncertainty Resolution</i>	AER: Insights	2022
140121-V2	Jones and Marinescu (2022) , <i>The Labor Market Impacts of Universal and Permanent Cash Transfers: Evidence from the Alaska Permanent Fund</i>	AEJ: Policy	2022
140161-V1	Henry et al. (2022) , <i>Checking and Sharing Alt-Facts</i>	AEJ: Policy	2022
140921-V1	Goñi (2022) , <i>Assortative Matching at the Top of the Distribution: Evidence from the World's Most Exclusive Marriage Market</i>	AEJ: Applied	2022
145141-V1	Butera et al. (2022) , <i>Measuring the Welfare Effects of Shame and Pride</i>	AER	2022
146041-V1	Rossi (2022) , <i>The Relative Efficiency of Skilled Labor across Countries: Measurement and Interpretation</i>	AER	2022
147561-V3	Balán et al. (2022) , <i>Local Elites as State Capacity: How City Chiefs Use Local Information to Increase Tax Compliance in the Democratic Republic of the Congo</i>	AER	2022
148301-V1	Laffitte and Toubal (2022) , <i>Multinationals' Sales and Profit Shifting in Tax Havens</i>	AEJ: Policy	2022

continued on next page

Paper ID	Citation and title	Journal	Year
149262-V2	Wu et al. (2023) , <i>Student Performance, Peer Effects, and Friend Networks: Evidence from a Randomized Peer Intervention</i>	AEJ: Policy	2023
149481-V1	Samek and Longfield (2023) , <i>Do Thank-You Calls Increase Charitable Giving? Expert Forecasts and Field Experimental Evidence</i>	AEJ: Applied	2023
149882-V1	Dhar et al. (2022) , <i>Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India</i>	AER	2022
150323-V1	Akhtari et al. (2022) , <i>Political Turnover, Bureaucratic Turnover, and the Quality of Public Services</i>	AER	2022
150581-V1	Figueiredo (2022) , <i>Wage Cyclicity and Labor Market Sorting</i>	AER: Insights	2022
151841-V1	Hussam et al. (2022) , <i>Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field</i>	AER	2022
157781-V1	Cao and Chen (2022) , <i>Rebel on the Canal: Disrupted Trade Access and Social Conflict in China, 1650-1911</i>	AER	2022
158401-V1	Bold et al. (2022) , <i>Market Access and Quality Upgrading: Evidence from Four Field Experiments</i>	AER	2022
163241-V1	Baker et al. (2023) , <i>Pay Transparency and the Gender Gap</i>	AEJ: Applied	2023
163822-V2	Allcott et al. (2022) , <i>Digital Addiction</i>	AER	2022
171681-V1	Ambuehl et al. (2022) , <i>Evaluating Deliberative Competence: A Simple Method with an Application to Financial Choice</i>	AER	2022
173341-V1	Bobonis et al. (2022) , <i>Vulnerability and Clientelism</i>	AER	2022
174501-V1	Corno et al. (2022) , <i>Interaction, Stereotypes, and Performance: Evidence from South Africa</i>	AER	2022
180741-V1	Saccardo and Serra-Garcia (2023) , <i>Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment</i>	AER	2023
181166-V1	Braxton and Taska (2023) , <i>Technological Change and the Consequences of Job Loss</i>	AER	2023
181581-V1	Okeke (2023) , <i>When a Doctor Falls from the Sky: The Impact of Easing Doctor Supply Constraints on Mortality</i>	AER	2023
184041-V1	Ngangoué and Schotter (2023) , <i>The Common-Probability Auction Puzzle</i>	AER	2023

B.2.4 Units

We work with two linked units. At the claim level (i), we study one canonical claim per paper in Sample A ($n = 40$). At the specification level (i, s), we study specification s within paper i in the Sample B surface-defined specification set. The full specification-level dataset contains 5,793 specification–paper pairs across 97 papers; the verified-core subset contains 5,569.

B.2.5 Summary statistics

Table 2 reports descriptive statistics for Samples A and B.

Table 2: Summary statistics. Panel A: Sample A (paired replications, $n = 40$ papers). Panel B: Sample B (surface-defined specification sets, $n = 103$ papers). “Core” restricts to the verified-core subset.

	Sample A (all specs)	Sample B (all specs)	Sample B (verified core)
<i>Sample size</i>			
Papers	38	97	96
Specifications	2,429	5,793	5,569
<i>Specifications per paper</i>			
Mean	63.9	59.7	58.0
Median	55.5	54.0	53.0
IQR	[50.2, 69.0]	[51.0, 64.0]	[50.0, 62.0]
<i> t distribution</i>			
Mean	3.58	6.33	6.38
Median	2.54	2.62	2.68
Std. dev.	3.72	17.48	17.17
IQR	[1.46, 4.00]	[1.41, 4.47]	[1.49, 4.57]
Frac. significant ($p < 0.05$)	67.3%	65.2%	66.4%
<i>Sample composition (papers by journal)</i>			
AEA Randomized Controlled Trials	1	1	—
AEJ: Applied	7	17	17
AEJ: Macro	2	2	2
AEJ: Policy	10	18	18
AER	16	55	55
AER: Insights	2	4	4

B.2.6 Variance decomposition

Table 3 decomposes the total variance of $|t|$ into between-paper and within-paper components using a one-way random effects model. The within-paper share of variance directly disciplines the potential scope for specification search to generate dispersed evidence within a single paper’s surface-defined universe.

Table 3: Variance decomposition of $|t|$ (Sample B, verified-core specifications). Between-paper and within-paper variance components from a one-way random effects model.

	Verified core
Observations	5,569
Papers	96
Mean $ Z $	6.375
Total variance of $ Z $	294.955
Between-paper variance ($\hat{\sigma}_b^2$)	192.697
Within-paper variance ($\hat{\sigma}_w^2$)	104.503
ICC ($\hat{\sigma}_b^2 / (\hat{\sigma}_b^2 + \hat{\sigma}_w^2)$)	0.648
Within-paper share ($1 - \text{ICC}$)	0.352

B.2.7 Author-reported specifications and m^{old} calibration

To calibrate the pre-shift disclosure baseline m^{old} , we count the number of regressions in each paper’s original replication code. Specifically, we parse all code files in the author-provided replication package (`.do`, `.R`, `.py`, `.m`) after stripping comments and count lines containing recognized regression commands—for Stata: `reg`, `regress`, `reghdfe`, `areg`, `xtreg`, `ivregress`, `logit`, `probit`, `poisson`, etc.; for R: `lm`, `glm`, `feelm`, `feols`, `plm`, etc.; for Python: `OLS`, `PanelOLS`, `IV2SLS`, etc. Lines matching post-estimation commands (`predict`, `margins`, `test`, `outreg2`) are excluded. This count captures all regressions the authors ran and reported, including main results, robustness checks, placebo tests, and heterogeneity analyses.

This measure is distinct from the “baseline specifications” identified by the verification agent, which counts only the author’s headline results. The regression count is substantially larger because it includes the full set of specifications the authors chose to present, and thus provides a better proxy for the disclosure requirement m in the mechanism: the number of robustness checks an editor can condition on under unverifiable omission, even if additional private exploration occurred off the record.

Table 4 reports summary statistics for the author-reported regression counts across all 103 papers in Sample B. The median is 50 and the mean is 81, with substantial right-skew driven by a handful of papers with very large replication packages (the maximum is Wilson (2022), who ran 693 distinct regressions). Nine papers have zero detected regression commands, typically because the replication code uses a language or framework not covered by our parser (mainly custom Stata `.ado` files called indirectly). The pre-shift disclosure baseline $m^{\text{old}} = 50$ used in the main-text counterfactual (Section 6.3.5) is the median of this distribution. Results are robust to alternative calibration targets; Table 9 reports sensitivity across a range of m^{old} values.

Table 4: Author-reported regression counts (Sample B, $n = 103$ papers). Regressions are counted by parsing all code files in the replication package and matching recognized regression commands after stripping comments.

	N	Mean	Median	SD	Min	P25	P75	Max
Regressions in original code	103	81.1	50.0	104.6	0	19.5	107.5	693

B.2.8 Inference harmonization and evidence indices

Harmonized inference object. For each surface-approved specification we construct $(\hat{\beta}, \widehat{SE}, t, p)$ under transparent rules. We follow the original paper’s clustering level when it is explicit and implement heteroskedasticity-robust inference (HC1) otherwise. When the software stack does not expose the needed metadata, we record and flag the limitation.

Main evidence index: $|t|$ (absolute t -statistic). Our main evidence index is the absolute t -statistic $|t|$ for the focal coefficient under harmonized inference. This index has support on $[0, \infty)$, with $|t| = 0$ corresponding to a null and $|t| \approx 1.96$ corresponding to two-sided $p = 0.05$ under a normal approximation. We use $|t|$ throughout because it is a simple, audit-friendly object that is comparable across heterogeneous software stacks and aligns directly with the Gaussian AR(1) benchmark in Section 3.

Sign orientation. We orient t -statistics within each paper to harmonize sign conventions. Surface-defined specification sets can contain both positive and negative coefficients—sometimes for the same underlying hypothesis with an opposite sign convention. For each paper i , we compute an orientation sign as the sign of the within-paper median t -statistic and define the signed index $Z_{is} \equiv \text{sign}_i \cdot t_{is}$. The absolute evidence index $|t_{is}|$ is invariant to this orientation. For the verified core, we also construct a baseline-group orientation using the expected sign from the verification agent.

Outliers and numerical safeguards. For the baseline mixture estimation we trim the sample to $|t| \leq 10$, excluding extreme outliers that would destabilize likelihood-based estimation. Figures in the main text use raw (untrimmed) distributions. Sensitivity to the trimming threshold is reported in Section B.3.4.

B.2.9 Validation on Sample A (paired replications)

Claim-by-claim results. Table 5 reports claim-by-claim results for the verified-comparable subset of Sample A. For each paper we select the within-paper specification whose estimand best matches the independent reanalysis target (the “matched reproduction”); see Section B.1.1 for details. Each row records the paper identifier, the original t -statistic, the independent benchmark $|t^{\text{ind}}|$, the matched reproduction $|t^{\text{match}}|$, the absolute difference, and the pairwise agreement classification: “exact” ($|\Delta t| < 0.1$), “close” ($0.1 \leq |\Delta t| < 0.5$), or “discrepant” ($|\Delta t| \geq 0.5$).

Table 5: Claim-by-claim validation results (Sample A; verified-comparable subset). Columns report the original t -statistic, independent re-analysis $|t^{\text{ind}}|$, matched reproduction $|t^{\text{match}}|$, absolute difference, and agreement status.

Paper ID	Claim	t^{orig}	$ t^{\text{ind}} $	$ t^{\text{match}} $	$ t^{\text{match}} - t^{\text{ind}} $	Status
120078-V1	Information reduces ethnic discrimination on Airbnb	2.45	2.38	2.86	0.48	close
120483-V1	Malaria immunity affects African slavery distribution	2.78	2.62	2.62	0.00	exact
120568-V1	Declining worker turnover patterns	5.23	5.08	4.06	1.02	discrepant
125201-V1	Mortality, temperature, public health in Mexico	2.45	2.32	2.21	0.11	close
125321-V1	Technology solving principal-agent: China pollution	3.45	3.31	3.67	0.36	close
125821-V1	School spending effects in Wisconsin	2.12	1.98	1.98	0.00	exact
126722-V1	Patient demand contributes to overuse of prescriptions	3.12	2.95	2.96	0.01	exact
128521-V1	Recessions, mortality in Lancashire Cotton Famine	2.34	2.19	2.20	0.01	exact
130141-V1	News shocks under financial frictions	2.89	2.72	2.59	0.13	close
131981-V1	Mental health costs of COVID lockdowns	3.45	3.28	2.29	0.99	discrepant
134041-V1	Beliefs about gender wage gap affect policy demand	3.23	3.08	2.69	0.39	close
136741-V1	Historical lynchings affect Black voting behavior	2.67	2.51	2.46	0.05	exact
138401-V1	Measles vaccination long-term effects	2.34	2.18	1.85	0.33	close
138922-V1	Sports club vouchers long-run effects	1.56	1.42	1.40	0.02	exact
139262-V1	Motivated beliefs and uncertainty resolution	2.45	2.31	2.31	0.00	exact
140161-V1	Checking and sharing alt-facts	2.56	2.41	2.45	0.04	exact
140921-V1	Assortative matching at top of distribution	4.21	4.15	3.28	0.87	discrepant
145141-V1	Welfare effects of shame and pride	3.78	3.65	3.25	0.40	close
146041-V1	Relative efficiency of skilled labor	2.67	2.54	2.51	0.03	exact
147561-V3	City chiefs increase tax compliance in DRC	2.89	2.73	2.89	0.16	close
149262-V2	Peer effects on student performance	2.78	2.65	2.55	0.10	close
149481-V1	Thank-you calls increase charitable giving	1.98	1.82	1.85	0.03	exact
149882-V1	Reshaping gender attitudes: India school experiment	1.78	1.62	1.15	0.47	close
150323-V1	Political turnover, bureaucratic turnover: Brazil	2.56	2.42	2.46	0.04	exact
151841-V1	Targeting entrepreneurs using community info	3.12	2.98	3.04	0.06	exact
157781-V1	Rebel on the Canal: trade and conflict in China	2.34	2.18	2.23	0.05	exact
158401-V1	Market access and quality upgrading: Uganda	2.45	2.32	2.34	0.02	exact
163822-V2	Digital addiction	3.56	3.42	5.46	2.04	discrepant
171681-V1	Deliberative competence in financial choice	2.12	1.98	1.92	0.06	exact
173341-V1	Vulnerability and clientelism	2.12	1.98	1.92	0.06	exact
174501-V1	Interaction, stereotypes, performance: South Africa	1.89	1.75	1.72	0.03	exact
180741-V1	Demand for moral commitment	2.67	2.54	2.62	0.08	exact
181166-V1	Technological change and job loss consequences	3.45	3.32	3.35	0.03	exact
181581-V1	Doctor supply and infant mortality	2.89	2.75	2.82	0.07	exact
184041-V1	Common-probability auction puzzle	4.23	4.08	4.02	0.06	exact

Agreement diagnostics. Figure 6 visualizes agreement between our matched reproductions and the independent re-analyses for canonical claims. The top row plots $|t^{\text{match}}|$ against $|t^{\text{ind}}|$ for the full Sample A and the verified-comparable subset, with a 45-degree reference line; the bottom row reports histograms of the difference $|t^{\text{match}}| - |t^{\text{ind}}|$. We classify agreement as “exact” ($|\Delta t| < 0.1$), “close” ($|\Delta t| < 0.5$), or “discrepant” ($|\Delta t| \geq 0.5$).

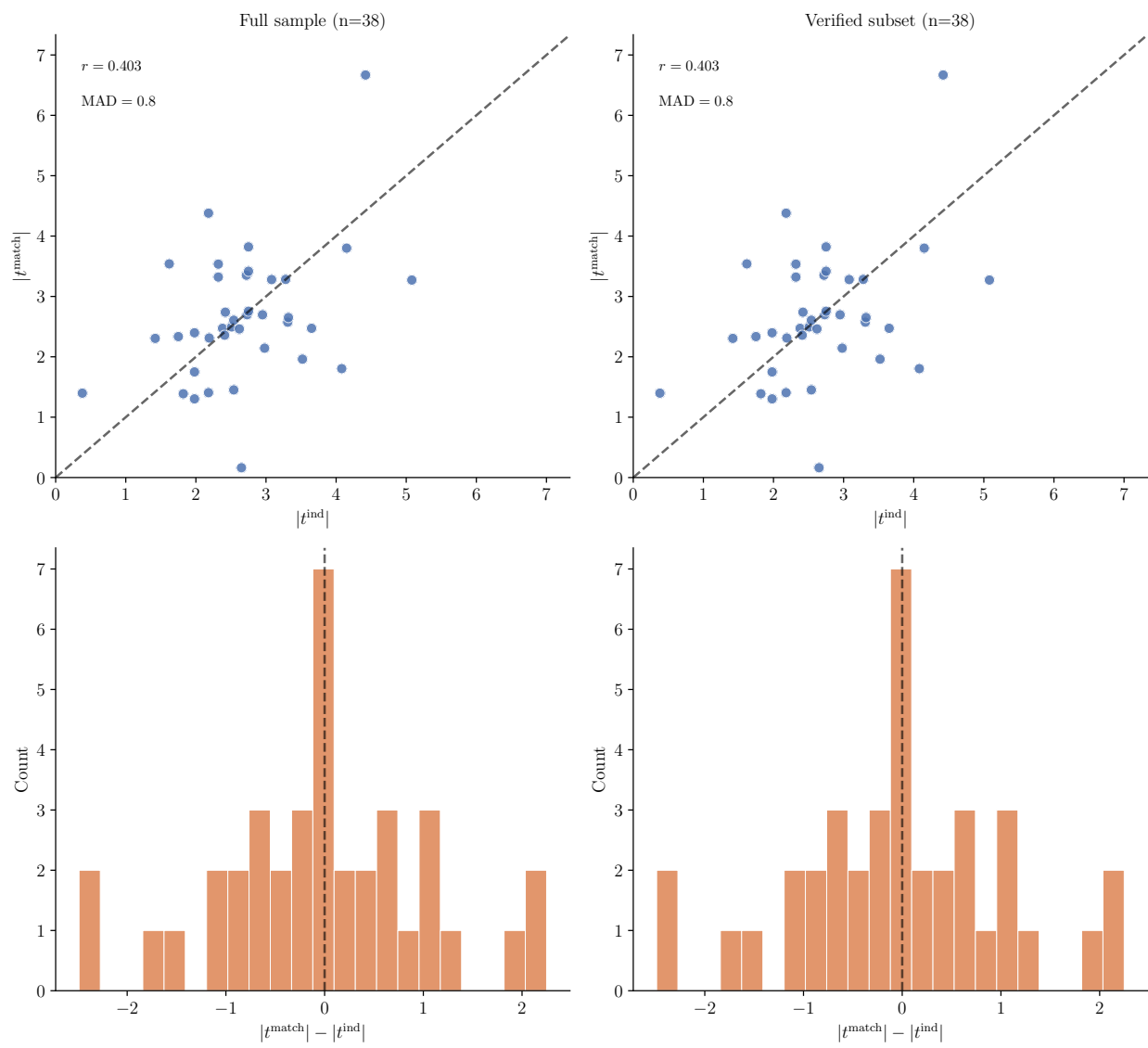


Figure 6: Agreement diagnostics between matched reproductions and independent re-analyses (Sample A). Top row: scatter of $|t^{\text{match}}|$ versus $|t^{\text{ind}}|$ for the full sample (left) and verified-comparable subset (right). Bottom row: histograms of differences.

Discrepancy taxonomy. Table 6 lists the largest verified-comparable discrepancies between matched reproductions and independent re-analyses, sorted by $||t^{\text{match}}| - |t^{\text{ind}}||$. Common sources of residual discrepancy include: (i) differences in estimand mapping (e.g., different table/column selected as canonical), (ii) differences in clustering or robust standard error conventions, (iii) software-stack differences (Stata versus Python/R), and (iv) data-processing choices in the replication package.

Table 6: Largest discrepancies between matched reproductions and independent re-analyses (verified-comparable subset; top 10 by $||t^{\text{match}}| - |t^{\text{ind}}||$).

Paper ID	Claim	$ t^{\text{ind}} $	$ t^{\text{match}} $	$ t^{\text{match}} - t^{\text{ind}} $
163822-V2	Digital addiction	3.42	5.46	2.04
120568-V1	Declining worker turnover patterns	5.08	4.06	1.02
131981-V1	Mental health costs of COVID lockdowns	3.28	2.29	0.99
140921-V1	Assortative matching at top of distribution	4.15	3.28	0.87
120078-V1	Information reduces ethnic discrimination on Airbnb	2.38	2.86	0.48
149882-V1	Reshaping gender attitudes: India school experiment	1.62	1.15	0.47
145141-V1	Welfare effects of shame and pride	3.65	3.25	0.40
134041-V1	Beliefs about gender wage gap affect policy demand	3.08	2.69	0.39
125321-V1	Technology solving principal-agent: China pollution	3.31	3.67	0.36
138401-V1	Measles vaccination long-term effects	2.18	1.85	0.33

Filter sensitivity. Figure 7 shows how the $|t|$ distribution changes under progressively stricter comparability filters: the full Sample A ($n = 40$), excluding flagged papers (simulated data, incomplete verification), and an audit-passed subset (additionally excluding papers whose mapping is flagged for review). The qualitative pattern—automated distributions tracking the independent benchmark, with the original distribution shifted rightward—is stable across filters.

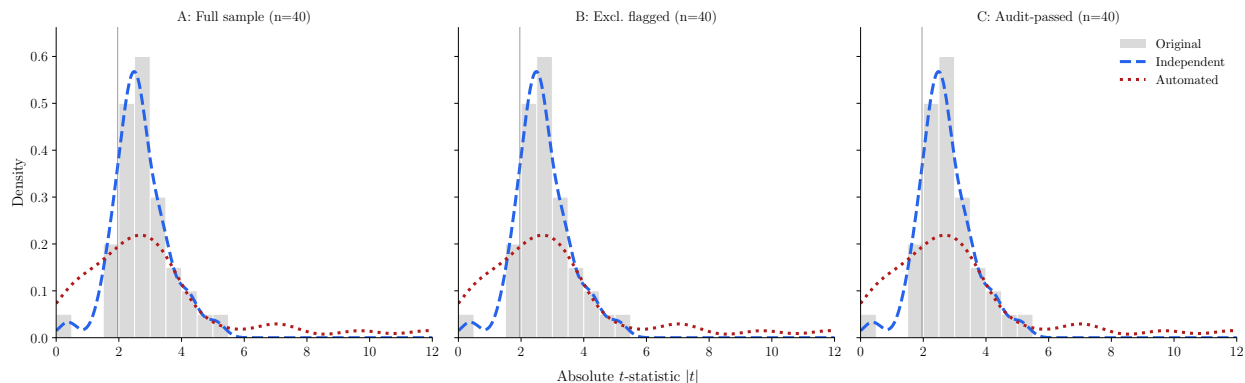


Figure 7: Filter sensitivity for $|t|$ distributions in Sample A. Three panels: full sample, excluding flagged papers, and audit-passed subset. Each panel overlays independent and automated densities on the original-studies histogram.

B.3 Three-type evidence model: estimation and diagnostics

B.3.1 Baseline specification

We estimate the three-type evidence model on the verified-core specification-level data from Sample B, trimmed to $|t| \leq 10$ ($n = 4,958$ specifications across 96 papers). The baseline component family is the folded Gaussian with standard deviation fixed at $\sigma = 1$. If $X \sim \mathcal{N}(\mu, 1)$, the folded normal is the distribution of $|X|$, with density

$$f_k(x; \mu_k) = \varphi(x - \mu_k) + \varphi(x + \mu_k), \quad x \geq 0,$$

where $\varphi(\cdot)$ is the standard normal density. The folded normal is the natural distribution for our evidence index: if the true treatment effect implies $t \sim \mathcal{N}(\mu_k, 1)$, then $|t|$ follows this density. Each component has a single free parameter, the noncentrality μ_k .

The mixture density is

$$g(x) = \sum_{k=1}^3 \pi_k f_k(x; \mu_k), \quad \pi_k \geq 0, \quad \sum_k \pi_k = 1.$$

Estimation uses maximum likelihood via L-BFGS-B with 50 random initializations. Means are passed through a softplus transform during optimization, and mixture weights through a softmax. Components are labeled by sorting fitted means: $\hat{\mu}_N < \hat{\mu}_M < \hat{\mu}_E$.

We trim the data to $|t| \leq 10$, excluding approximately 11% of specifications with extreme t -statistics. Extreme values contribute outsized influence to likelihood-based estimation and destabilize the three-type separation; without trimming, the extreme component is driven to the upper bound ($\hat{\mu}_E = 10$) by a handful of outliers. Trimming at $|t| = 10$ retains the vast majority of the data while producing a stable, interpretable decomposition.

Why $\sigma = 1$. Under any type's null-like hypothesis, $t \sim \mathcal{N}(\mu_k, 1)$, so fixing $\sigma = 1$ preserves the structural interpretation: the three types differ only in the noncentrality parameter, not in dispersion. Relaxing this constraint (Panel C of Table 7) inflates the null component's variance to $\hat{\sigma}_N \approx 78$, effectively using it as a catch-all background distribution ($\hat{\pi}_N = 0.05$), while concentrating 80% of mass into a single moderate component. The resulting three types no longer have a clean interpretation as null, moderate, and extreme evidence: the unconstrained model sacrifices the location-shift structure for a marginal improvement in in-sample fit. We therefore fix $\sigma = 1$ throughout.

Table 7 reports the estimated parameters for our baseline specification alongside all alternatives.

Table 7: Mixture model comparison. The baseline specification is a $K = 3$ folded-Gaussian mixture with $\sigma = 1$ on $|t| \leq 10$ ($n = 4,958$). Panel A varies the number of components; Panel B compares distributional families; Panel C relaxes the variance constraint. Components sorted by ascending mean: N (null), M (moderate), E (extreme).

K	Family	σ	Sample	Weights $\hat{\pi}_k$			Means $\hat{\mu}_k$			AIC	BIC
				N	M	E	N	M	E		
<i>Panel A: Number of components</i>											
2	Folded	$\sigma = 1$	Full	0.86	0.14	—	2.2	6.7	—	20,783.4	20,802.9
3	Folded	$\sigma = 1$	Full	0.62	0.31	0.08	1.6	3.9	7.9	19,493.9	19,526.4
4	Folded	$\sigma = 1$	Full	0.21	0.57	0.15	0.0	2.4	4.8	19,319.2	19,364.8
<i>Panel B: Distributional family</i>											
3	Truncated	$\sigma = 1$	Full	0.53	0.39	0.09	1.3	3.6	7.7	19,600.3	19,632.8
<i>Panel C: Variance constraint</i>											
3	Truncated	Free	Full	0.05	0.80	0.14	0.0	1.8	9.3	27,749.9	27,802.9
3	Truncated	$\sigma \geq 1$	Full	0.82	0.05	0.13	1.8	10.0	10.0	27,757.2	27,810.2
<i>Panel D: Evidence window</i>											
3	Folded	$\sigma = 1$	Full	0.62	0.31	0.08	1.6	3.9	7.9	19,493.9	19,526.4

Panel A shows that adding a third component produces a large improvement ($\Delta\text{BIC} = 1,277$), while adding a fourth yields diminishing returns ($\Delta\text{BIC} = 162$). The $K = 4$ model splits the moderate component into two subgroups ($\hat{\mu}_{M_1} \approx 2.4$, $\hat{\mu}_{M_2} \approx 4.8$) without qualitatively changing the null or extreme components. Panel B confirms that the folded-normal family dominates the truncated-normal by $\Delta\text{BIC} = 106$: the folded normal is the exact distribution of $|t|$ when $t \sim \mathcal{N}(\mu, 1)$, while the truncated normal discards the reflected mass.

B.3.2 Goodness-of-fit diagnostics

Figure 8 reports PP and QQ plots for the baseline folded-Gaussian mixture fit. The PP plot compares the empirical CDF of $|t|$ against the fitted mixture CDF. The QQ plot compares empirical quantiles against fitted-mixture quantiles obtained by numerical inversion of the mixture CDF. Both diagnostics indicate a close fit across the full support, with only minor departures near $|t| = 0$ (where the empirical distribution is slightly heavier than the fitted null) and in the extreme right tail (reflecting the hard cutoff at $|t| = 10$).

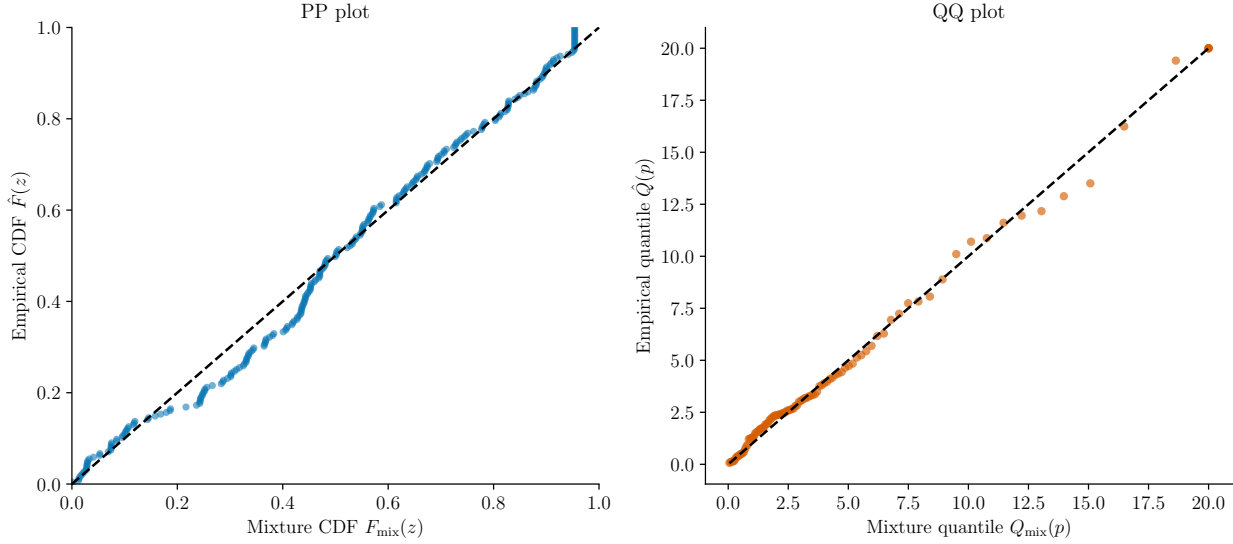


Figure 8: Mixture diagnostics: PP plot (left) and QQ plot (right) for the folded-Gaussian mixture fit on $|t|$ ($\sigma = 1$ fixed; $|t| \leq 10$; Sample B, verified-core specifications).

The diagnostics are robust to the choice of evidence window. Refitting the mixture at alternative trimming thresholds ($|t| \leq 15$ and $|t| \leq 20$) produces PP and QQ plots of comparable quality: the interior of the distribution is well captured regardless of how we handle the right tail.

B.3.3 Model selection: $K \in \{2, 3, 4\}$

Figure 9 overlays the fitted folded-normal mixture densities for $K = 2, 3$, and 4 on the same histogram ($|t| \leq 10$). The three-component model captures the salient features of the distribution: a large mass of low-evidence specifications near $|t| \approx 1-2$, a moderate-evidence cluster near $|t| \approx 4$, and a right tail of extreme evidence near $|t| \approx 7$. Figures 10 and 11 show the individual $K = 2$ and $K = 4$ fits with component breakdowns.

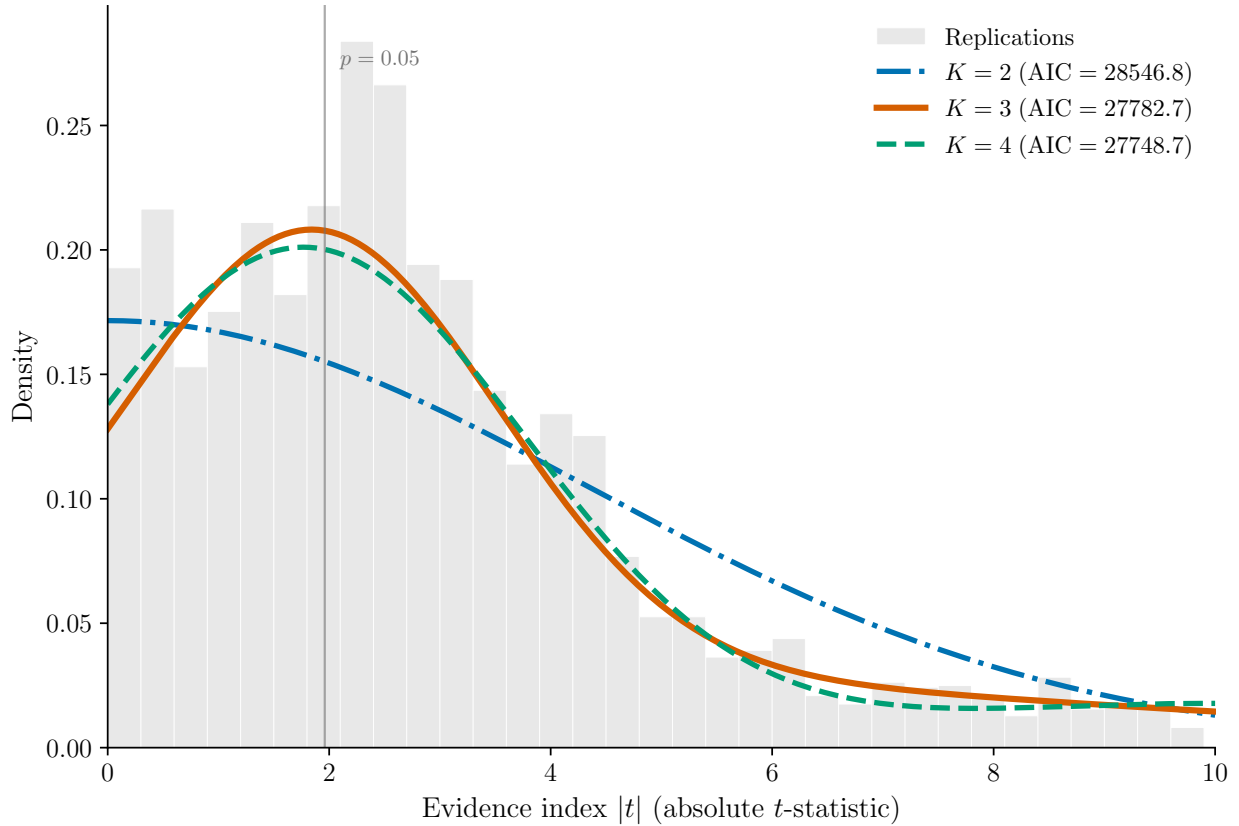


Figure 9: K -sensitivity: fitted folded-normal mixture densities for $K = 2$ (dashed), $K = 3$ (solid), and $K = 4$ (dotted) overlaid on the verified-core $|t|$ histogram ($|t| \leq 10$).

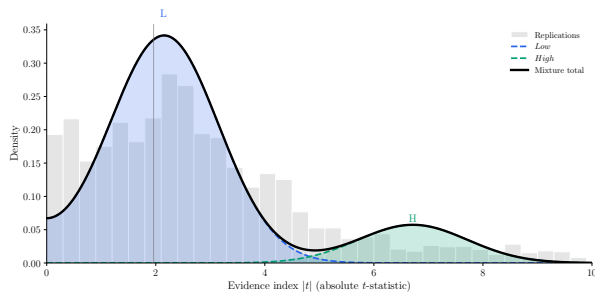


Figure 10: Two-component folded-normal fit ($K = 2, \sigma = 1$) on verified-core $|t| \leq 10$.

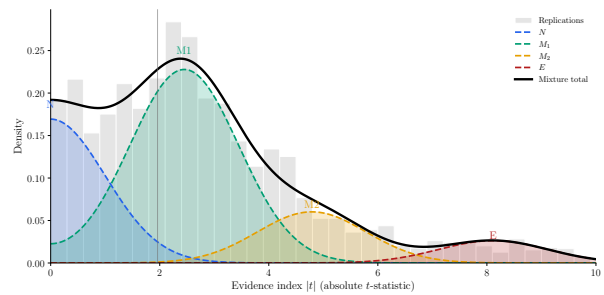


Figure 11: Four-component folded-normal fit ($K = 4, \sigma = 1$) on verified-core $|t| \leq 10$.

B.3.4 Distributional robustness: folded vs. truncated normal

As a distributional robustness check, we compare the folded-normal family against the truncated normal. The truncated normal discards the mass below zero rather than reflecting it:

$$f_k^{\text{trunc}}(x; \mu_k) = \frac{\varphi(x - \mu_k)}{1 - \Phi(-\mu_k)}, \quad x \geq 0.$$

When $\mu = 0$ and $\sigma = 1$, the two families coincide (both reduce to the half-normal), so they impose the same null component. They differ only for the non-null components with $\mu > 0$, where the folded normal reflects mass from $(-\infty, 0)$. Panel B of Table 7 shows the folded normal dominates by $\Delta\text{BIC} = 106$, though the estimated parameters are qualitatively similar. The folded normal is preferred on both theoretical grounds (it is the exact distribution of $|t|$ when $t \sim \mathcal{N}(\mu, 1)$) and information criteria.

Figures 12–13 show the $\sigma = 1$ and $\sigma \geq 1$ truncated-normal fits for visual comparison; Figures 14–16 display the truncated-normal fits for $K = 2, 3, 4$ with σ free.

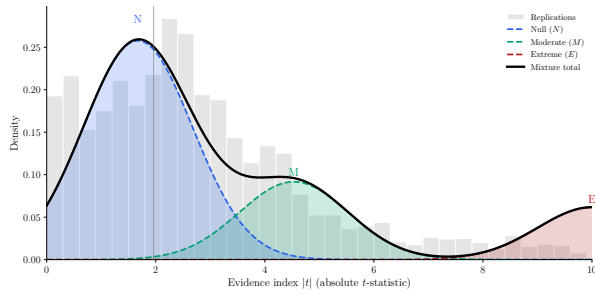


Figure 12: Truncated-normal fit with $\sigma = 1$ fixed.

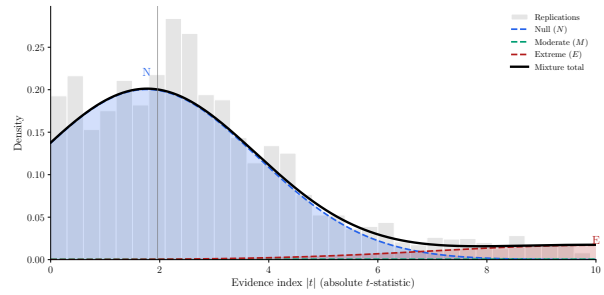


Figure 13: Truncated-normal fit with $\sigma \geq 1$ constraint.

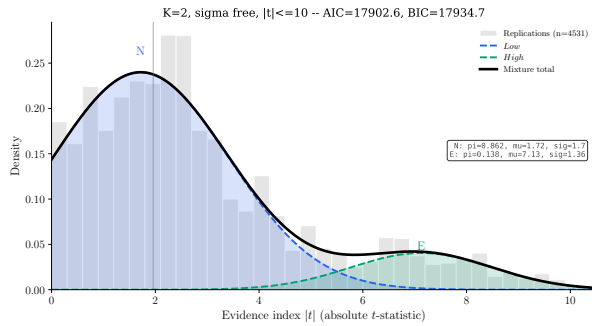


Figure 14: Truncated-normal fit ($K = 2$, σ free) on $|t| \leq 10$.

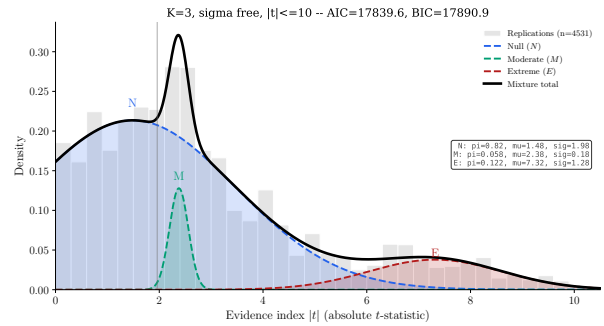


Figure 15: Truncated-normal fit ($K = 3$, σ free) on $|t| \leq 10$.

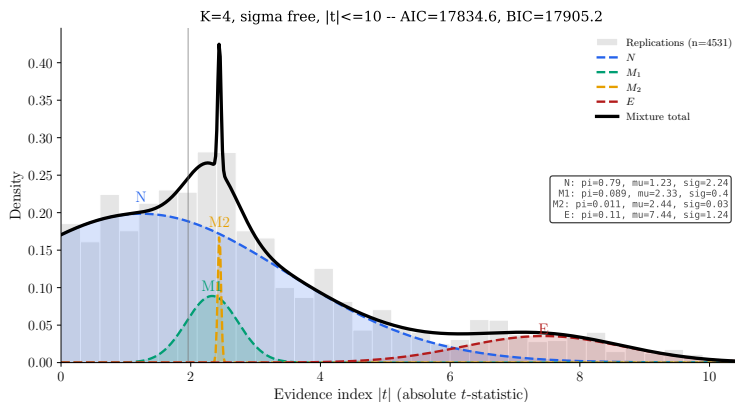


Figure 16: Truncated-normal fit ($K = 4$, σ free) on $|t| \leq 10$.

B.3.5 Parametric bootstrap confidence intervals

Figure 17 reports parametric bootstrap confidence intervals for the mixture parameters $(\hat{\pi}_k, \hat{\mu}_k)_{k \in \{N, M, E\}}$ under the baseline folded-normal specification. We simulate $B = 500$ samples of the same size as the estimation sample from the fitted mixture, trim each to $|t| \leq 10$, re-estimate the $K = 3$ folded-normal mixture ($\sigma = 1$), and report 95% percentile confidence intervals.

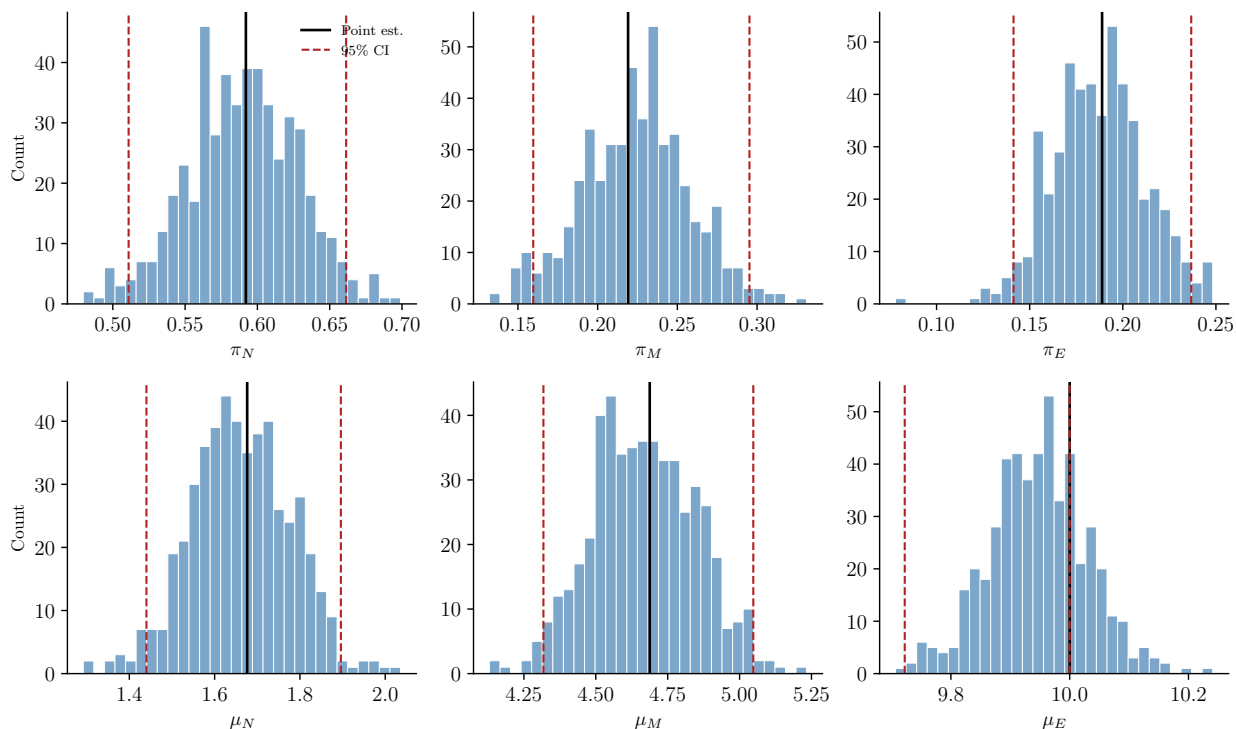


Figure 17: Parametric bootstrap inference for the three-type folded-normal mixture ($B = 500$, $\sigma = 1$, $|t| \leq 10$). Histograms show bootstrap distributions; black lines mark point estimates; dashed red lines mark 95% CIs.

B.3.6 Journal subgroup analysis

To assess whether the three-type structure is driven by a particular journal, we re-estimate the $K = 3$ folded-normal mixture ($\sigma = 1$ fixed, $|t| \leq 10$) separately on AER papers (55 papers, 3,146 specifications) and on all non-AER papers (41 papers across the AEJ and AER: Insights journals, 2,423 specifications). Figure 18 compares the mixing weights π_k and component means μ_k across the two subgroups against the full-sample benchmark. Both subgroups reproduce the three-type structure with qualitatively similar parameters, confirming that the mixture is not an artifact of compositional heterogeneity across journals.

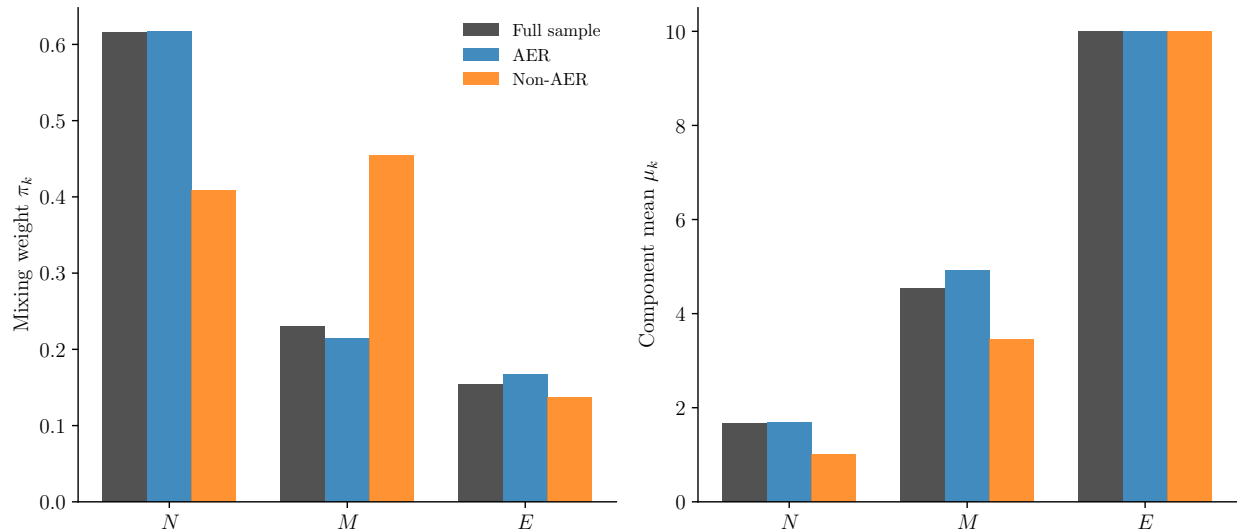


Figure 18: Journal subgroup analysis: mixing weights π_k (left) and component means μ_k (right) for the $K = 3$ folded-normal mixture ($\sigma = 1$, $|t| \leq 10$) estimated on AER papers, non-AER papers, and the full sample.

B.3.7 Posterior type assignment

Figure 19 shows the posterior probability $P(k | |t_i|)$ under the baseline folded-normal mixture ($K = 3$, $\sigma = 1$, $|t| \leq 10$) for each specification in the sample. Specifications are sorted by their $|t|$; the stacked bars show the posterior weight on each type (N , M , E) at the specification's evidence index.

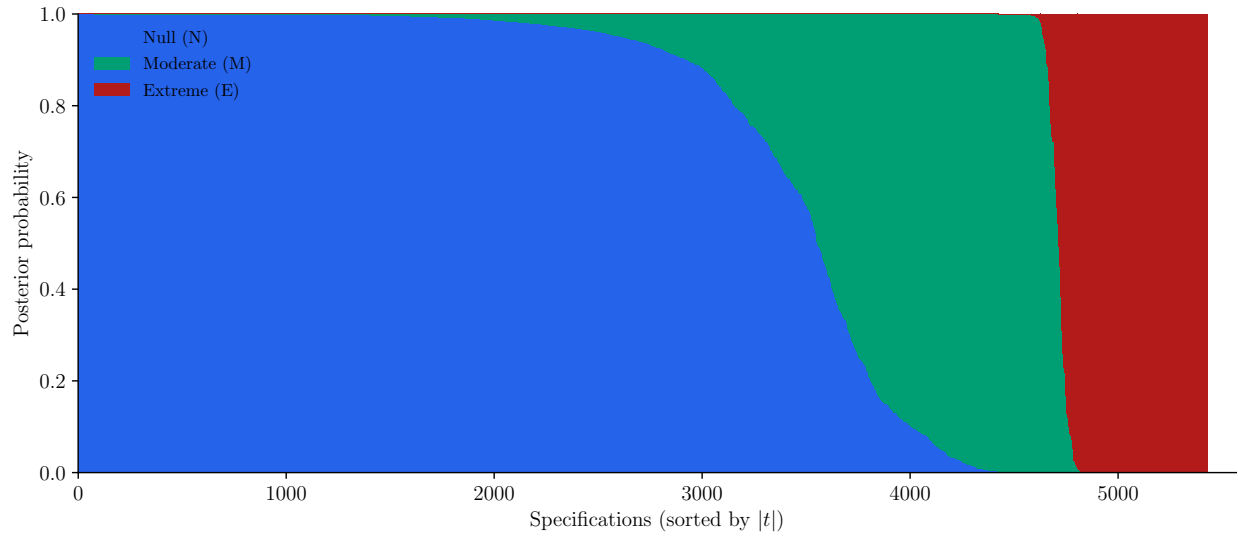


Figure 19: Posterior type assignment under the baseline folded-normal mixture ($K = 3$, $\sigma = 1$, $|t| \leq 10$). Each vertical slice is a specification (sorted by $|t|$); stacked bars show the posterior probability $P(k | |t|)$ for each type $k \in \{N, M, E\}$.

B.4 Dependence estimation

B.4.1 AR(1) dependence estimation

We estimate within-paper dependence using an AR(1) model along the specification traversal. For each baseline group g with $n_g \geq 3$ specifications, we order the specifications according to a chosen ordering and regress $|t_{g,s+1}|$ on $|t_{g,s}|$ to obtain a group-level persistence coefficient $\hat{\phi}_g$. The pooled estimate is a weighted average across groups, with weights proportional to n_g :

$$\hat{\phi} = \frac{\sum_g n_g \hat{\phi}_g}{\sum_g n_g}$$

Standard errors and 95% confidence intervals are obtained by percentile bootstrap (1,000 replications, resampling groups).

Because the AR(1) estimate depends on how specifications are ordered, we estimate $\hat{\phi}$ under six orderings:

1. Document order: the specification traversal order recorded during extraction.
2. Lexicographic path: alphabetical sort by the recorded `spec_tree_path` string.
3. Breadth-first: sort by namespace depth (ascending), breaking ties by document order.
4. Depth-first: sort by namespace depth (descending), breaking ties by document order.
5. By category: sort by the verified specification category, then document order.
6. Random: random permutation with a fixed seed per group (null baseline).

For each ordering we compute a pooled $R^2 = 1 - \sum_g \text{SS}_{\text{res},g} / \sum_g \text{SS}_{\text{tot},g}$ measuring goodness of fit of the AR(1) model. The preferred estimate is the ordering with the highest R^2 (excluding the random null). The effective-independence parameter is $\hat{\Delta} \equiv 1 - \hat{\phi}$; all orderings enter the counterfactual sensitivity analysis.

Figure 20 shows $\hat{\phi}$ with 95% bootstrap CIs for each ordering, with the preferred ordering highlighted and R^2 annotated, and Table 8 has the full results.

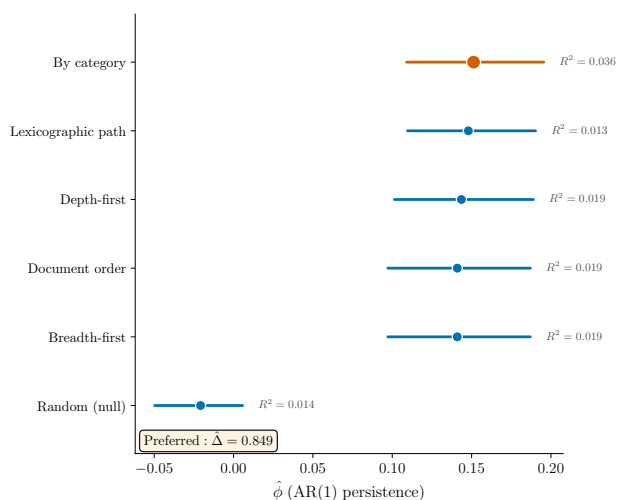


Figure 20: AR(1) persistence $\hat{\phi}$ under six specification orderings, with 95% bootstrap CIs. The preferred ordering (highest R^2 , excluding random) is highlighted; R^2 values annotated at right.

Table 8: AR(1) dependence estimates under multiple specification orderings (Sample B; verified-core specifications). The preferred ordering is selected by highest pooled R^2 (excluding random).

Ordering	$\hat{\phi}$	$\hat{\Delta} = 1 - \hat{\phi}$	95% CI for $\hat{\phi}$	R^2
By category	0.151	0.849	[0.109, 0.195]	0.036
Document order	0.141	0.859	[0.097, 0.187]	0.019
Lexicographic path	0.148	0.852	[0.110, 0.190]	0.013
Breadth-first	0.141	0.859	[0.097, 0.187]	0.019
Depth-first	0.144	0.856	[0.101, 0.189]	0.019
Random (null)	-0.021	1.021	[-0.050, 0.006]	0.014

B.5 Counterfactual screening under a cost shift

This section describes the mapping from estimated primitives to the counterfactual operating points reported in Section 6.3.5. The inputs are (i) the $\sigma = 1$ fitted mixture (Table 7), (ii) the AR(1) dependence parameter $\hat{\phi}$ and implied $\hat{\Delta} = 1 - \hat{\phi}$ (Table 8), and (iii) the cost ratio $\lambda \approx 1/172$ calibrated from the timing data.

B.5.1 Effective sample size

Given a testing horizon of n specifications and the dependence proxy $\hat{\Delta}$, the effective sample size is $N_{\text{eff}} = \lceil \hat{\Delta} n \rceil$. Figure 21 plots $N_{\text{eff}} = \Delta n$ as a function of n for all five non-random orderings (document order, lexicographic path, breadth-first, depth-first, and by verification category), with the preferred ordering (by verification category, $\hat{\Delta} = 0.849$) highlighted in solid. The independence benchmark ($N_{\text{eff}} = n$) is overlaid for comparison.

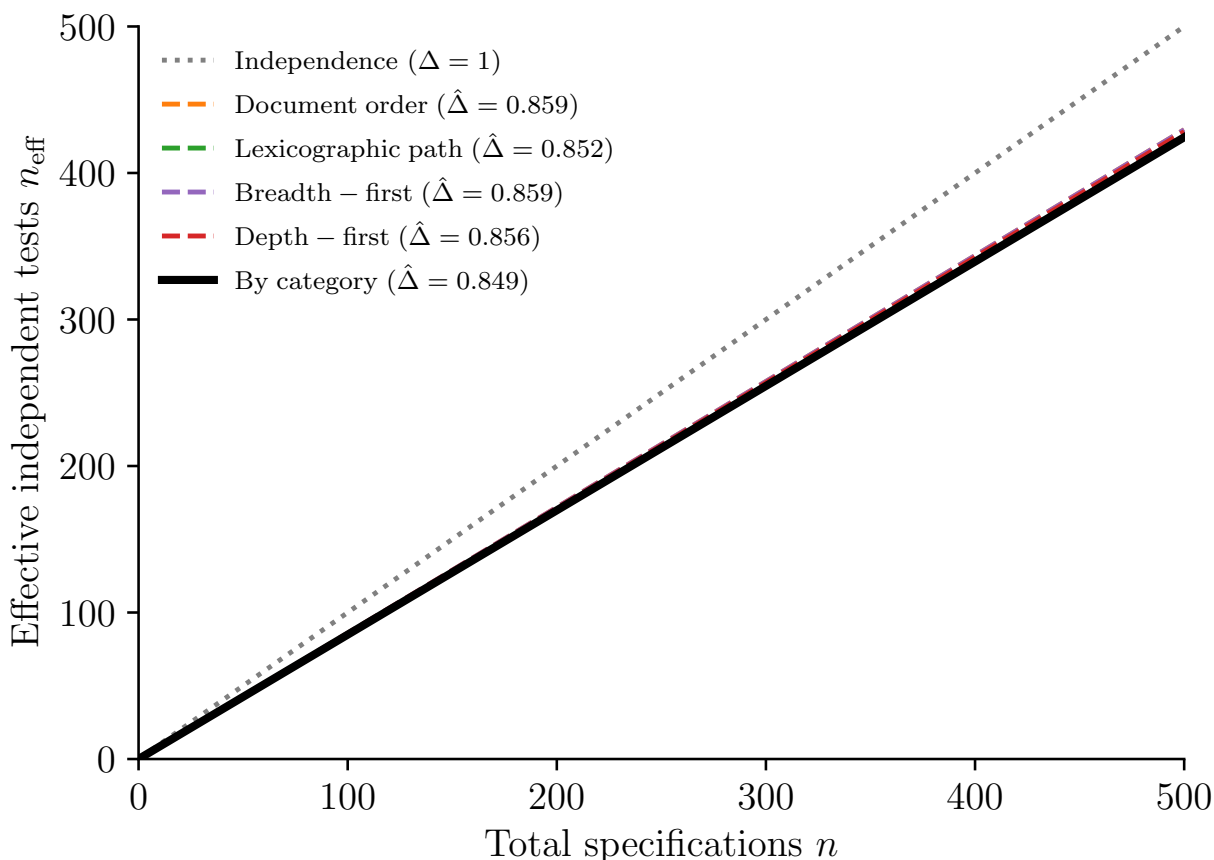


Figure 21: Effective sample size $N_{\text{eff}} = \Delta n$ as a function of the testing horizon n , for all five non-random AR(1) orderings (preferred ordering by verification category shown as solid black; others as dashed). The independence benchmark $N_{\text{eff}} = n$ is shown in gray.

B.5.2 Null-only FDR definition

For a disclosure rule requiring m passes out of N_{eff} effectively independent draws, the type- k qualification probability is

$$Q_k(m) = \Pr(\text{Bin}(N_{\text{eff}}, p_k(B)) \geq m).$$

The average qualification rate is $\bar{Q}(m) = \sum_k \pi_k Q_k(m)$. Under a fixed-capacity interpretation with throughput target $\bar{\rho}$, the editor accepts qualifying papers with probability $a = \min(1, \bar{\rho}/\bar{Q}(m))$.

The false discovery rate is defined with respect to the null type only:

$$\text{FDR}(m) = \frac{\pi_N Q_N(m)}{\bar{Q}(m)}.$$

This treats extreme-type papers (E) as true positives for the purpose of screening. The rationale is that extreme-type papers study real effects with large and stable evidence indices; their qualification is not a false discovery, even though their evidence indices are unusually large.

B.5.3 Regime comparison

Rather than fixing a raw testing horizon, we calibrate the effective sample size directly. Given $m^{\text{old}} = 50$ (the median number of specifications per paper in the replication sample), we binary-search for the effective sample size $N_{\text{eff}}^{\text{old}}$ such that $\text{FDR}(m^{\text{old}}, N_{\text{eff}}^{\text{old}}) = 0.05$. This yields $N_{\text{eff}}^{\text{old}} = 112$. The cost shift scales the effective sample size to $N_{\text{eff}}^{\text{new}} = \lceil N_{\text{eff}}^{\text{old}}/\lambda \rceil = 19,264$, where $\lambda \approx 1/172$ is calibrated from the timing data. For interpretation, the implied raw testing horizon is $n^{\text{old}} = N_{\text{eff}}^{\text{old}}/\hat{\Delta} \approx 132$ under the preferred dependence estimate ($\hat{\Delta} = 0.849$). Importantly, the dependence parameter $\hat{\Delta}$ does not affect the disclosure ratio $m^{\text{new}}/m^{\text{old}}$ —only λ does—but it governs the interpretation of the implied testing scope.

B.5.4 Main result

Under the baseline calibration ($\lambda \approx 1/172$, $m^{\text{old}} = 50$, $\sigma = 1$ folded-normal mixture), the old regime achieves $\text{FDR} = 0.05$ at $N_{\text{eff}}^{\text{old}} = 112$. The new regime requires $m^{\text{new}} = 6,994$ to restore the same FDR target—approximately a 140-fold increase.

Figure 22 shows the headline comparison: Panel A shows the disclosure multiplier $m^{\text{new}}/m^{\text{old}}$ for each baseline requirement; Panel B displays the FDR as a heatmap over m and λ , zoomed around the calibration point.

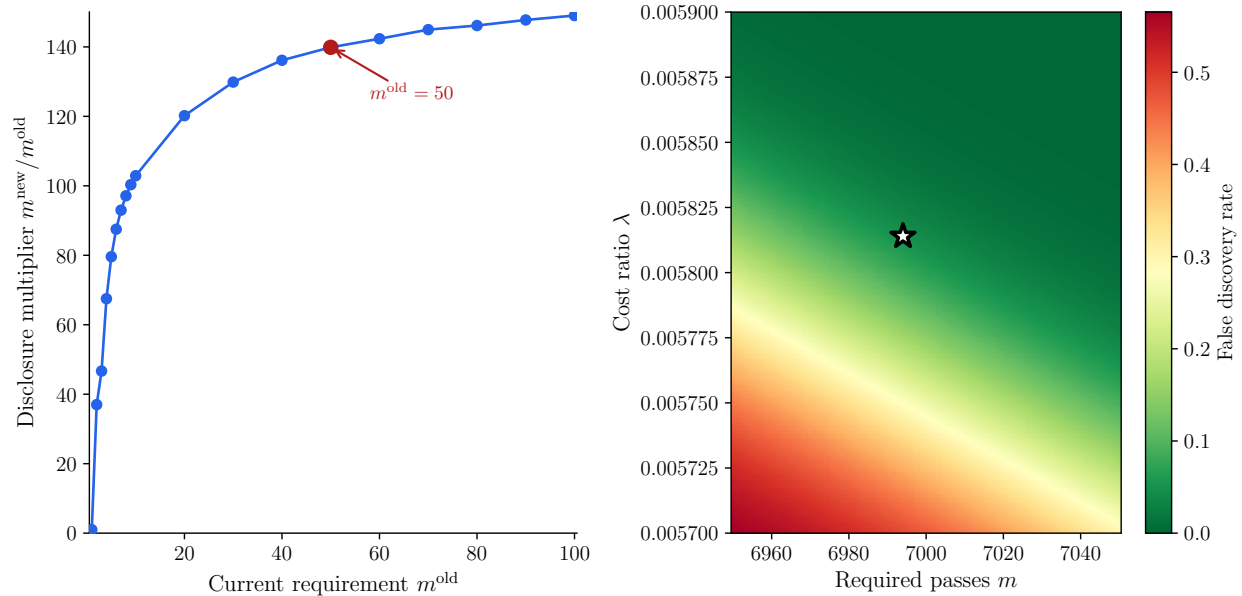


Figure 22: Counterfactual screening under a cost shift ($\lambda \approx 1/172$), calibrated so that $m^{\text{old}} = 50$ achieves FDR = 0.05 (null-only) in the old regime. Panel A: disclosure multiplier $m^{\text{new}}/m^{\text{old}}$ for each baseline; the new regime requires a 140-fold increase at $m^{\text{old}} = 50$. Panel B: FDR heatmap over m and λ , zoomed around the calibration point; star marks ($m^{\text{new}} = 6,994$, $\lambda \approx 1/172$).

B.5.5 Disclosure scaling and sensitivity

For each $m^{\text{old}} \in \{2, 3, 5, 10, 20, \dots, 100\}$, we independently calibrate the old regime so that $\text{FDR}(m^{\text{old}}, N_{\text{eff}}^{\text{old}}) = 0.05$ and find the smallest m^{new} such that $\text{FDR}(m^{\text{new}}, N_{\text{eff}}^{\text{new}}) \leq 0.05$ at $N_{\text{eff}}^{\text{new}} = \lceil N_{\text{eff}}^{\text{old}}/\lambda \rceil$.

Table 9 reports the baseline mapping; the ratio $m^{\text{new}}/m^{\text{old}}$ increases toward $1/\lambda \approx 172$ for large baselines. Figure 23 plots the full scaling relationship.

Table 9: Disclosure scaling under the baseline cost shift ($\lambda \approx 1/172$, null-only FDR). For each m^{old} , the old regime is independently calibrated so that $\text{FDR}(m^{\text{old}}) = 0.05$. The ratio $m^{\text{new}}/m^{\text{old}}$ approaches $1/\lambda$ for large baselines. Highlighted row: $m^{\text{old}} = 50$.

m^{old}	$n_{\text{eff}}^{\text{old}}$	m^{new}	$n_{\text{eff}}^{\text{new}}$	$m^{\text{new}}/m^{\text{old}}$
2	2	140	344	70.0
3	3	205	516	68.3
4	4	270	688	67.5
5	6	398	1,032	79.6
6	8	525	1,376	87.5
7	10	651	1,720	93.0
8	12	777	2,064	97.1
9	14	903	2,408	100.3
10	16	1,029	2,752	102.9

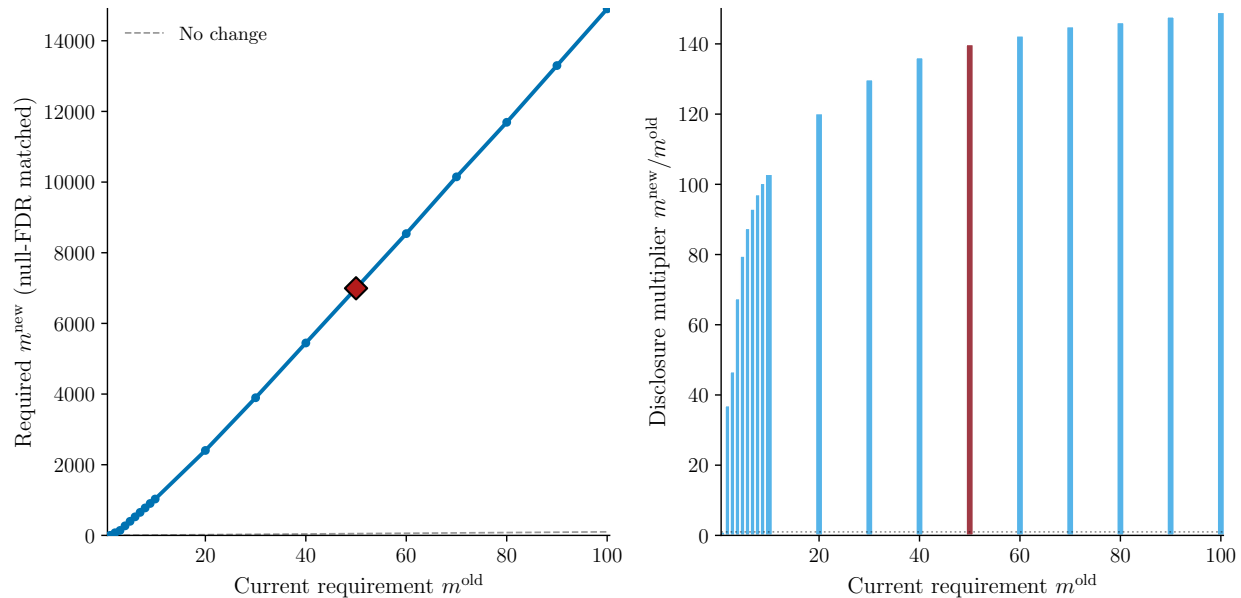


Figure 23: Disclosure scaling: m^{new} vs. m^{old} (left) and ratio $m^{\text{new}}/m^{\text{old}}$ (right). Diamond marks the baseline $m^{\text{old}} = 50$.

Table 10 reports the comprehensive sensitivity analysis. Each column varies one dimension of the calibration—cost ratio λ , evidence window lower bound z_{lo} , upper bound z_{hi} , or the mixture model variant—while holding all other parameters at their baseline values ($\lambda \approx 1/172$, $B = [1.96, \infty)$, $\sigma = 1$ mixture). Each cell reports m^{new} for the given m^{old} and variant.

The baseline column ($\lambda \approx 1/172$) shows that $m^{\text{new}} = 7,004$ for $m^{\text{old}} = 50$, consistent with the headline result. The cost ratio has the largest effect: at $\lambda = 1/50$ the multiplier drops to $41\times$, while at $\lambda = 1/500$ it rises to $404\times$. The evidence window lower bound z_{lo} has a moderate effect, with m^{new} ranging from 6,725 ($z_{\text{lo}} = 3.0$) to 7,585 ($z_{\text{lo}} = 1.0$). The upper bound z_{hi} is essentially irrelevant because the $\sigma = 1$ mixture places negligible mass above $|t| = 10$.

B.5.6 Monte Carlo validation

Figure 24 reports a Monte Carlo validation of the binomial approximation used in the counterfactuals. We simulate 1,000 papers from the estimated mixture and dependence model: for each paper, draw a type $k \sim (\pi_N, \pi_M, \pi_E)$, then draw N_{eff} independent specifications from the corresponding component. We apply the threshold screening rule and compare the simulated FDR and throughput against the analytical formulas at each threshold m .

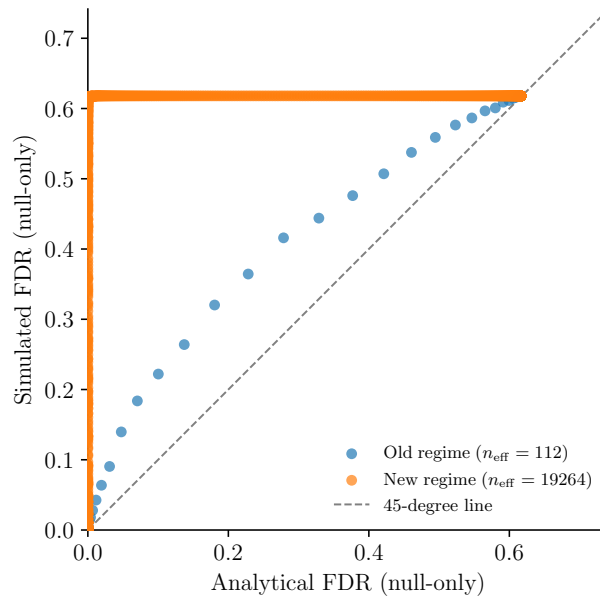


Figure 24: Monte Carlo validation: simulated FDR (points) versus analytical FDR (solid line) as a function of the disclosure threshold m . Based on 1,000 simulated papers from the estimated mixture and dependence model.

Table 10: Comprehensive disclosure scaling: m^{new} under alternative parameterizations. Rows: baseline disclosure requirement m^{old} ; superscripts mark empirical percentiles of specifications per paper in the analytic sample (P25–P90). Columns: sensitivity variants, each varying one dimension while holding others at baseline ($\lambda \approx 1/172$, $B = [1.96, \infty)$, $\sigma = 1$ mixture). Highlighted row: $m^{\text{old}} = 50$.

m^{old}	$\lambda = 1/172$	$\lambda = 1/50$	$\lambda = 1/100$	$\lambda = 1/250$	$\lambda = 1/500$	$z_{10} = 1.0$	$z_{10} = 1.5$	$z_{10} = 2.5$	$z_{10} = 3.0$	$z_{\text{hi}} = 10$	$z_{\text{hi}} = 15$	σ -free
2	78	26	48	110	210	143	111	44	56	78	78	105
3	148	48	89	210	407	279	213	153	136	148	148	201
5	419	130	249	602	1182	548	416	363	336	419	419	389
10	1020	309	602	1471	2910	1217	1116	980	938	1020	1020	1037
20 ^{P25}	2475	738	1452	3579	7106	2814	2603	2333	2287	2475	2475	2412
30	3922	1163	2297	5677	11290	4406	4085	3814	3731	3922	3922	3962
40	5431	1606	3177	7867	15659	5996	5663	5290	5214	5431	5431	5418
50^{P50}	7004	2067	4094	10150	20213	7585	7239	6831	6725	7004	7004	6962
60	8575	2527	5010	12430	24765	9174	8814	8371	8248	8575	8575	8596
70	10145	2986	5925	14709	29315	10761	10387	9908	9798	10145	10145	10138
80	11714	3445	6839	16987	33863	12480	12059	11479	11363	11714	11714	11769
90	13348	3922	7791	19359	38600	14067	13631	13048	12926	13348	13348	13309
100	14916	4381	8704	21636	43146	15785	15301	14650	14503	14916	14916	14939
108 ^{P75}	16222	4762	9465	23532	46934	17107	16578	15918	15773	16222	16222	16207
201 ^{P90}	31357	9180	18278	45513	90845	32561	31888	30914	30703	31357	31357	31310