THE INFORMATION COST OF EXPOSURE-WEIGHTED RISK CERTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A risk certificate supports deployment only if it can approve a safe policy with the available data. This requirement is difficult to assess when inputs generate different numbers of predictions or amounts of financial exposure. We study the sample cost of certifying the ratio of expected erroneous exposure to expected exposure. For fixed mean exposure and a specified safe alternative, we reduce the finite-sample minimax problem over bounded observation laws to an exactly solvable three-point experiment. The resulting sample requirement has order $\log(1/\delta)/[\mu\,\mathrm{kl}(r,\alpha)]$, where μ is normalized mean exposure, r is risk, $\alpha > r$ is the risk cap, and δ bounds false certification, with α and target power fixed. This benchmark distinguishes necessary data from conservative test design and loose sufficient bounds. In one zero-risk setting, the optimum requires 27 observations, an implementable sampling test requires 29, and its sufficient bound gives 360. We also connect sampling likelihood ratios to direct product betting and identify the same information scale in worst-case log growth. Controlled experiments and multilabel prediction tasks show when direct tests use population structure that sampling discards, and when sampling remains more effective. Together, the analysis and experiments explain how a reduction can preserve worst-case power while leaving room for improvement on particular populations.

1 INTRODUCTION

A predictive model often creates more than one consequence from a single input. An image classifier can assign several labels, while a lending policy can approve loans of very different sizes. Before deployment, a practitioner must decide whether the errors across those labels or the losses across that capital are acceptable. A risk certificate can support this decision by controlling the chance of approving an unsafe policy. The practitioner also needs to know whether the available data can certify a useful safe policy. Otherwise, an unsuccessful certification leaves two very different responses unresolved: collect more data or use the existing data more effectively.

The distinction starts with the unit in which risk is measured. Suppose 95% of inputs receive one correct label each, while 5% receive 100 labels each, including 60 errors. The mean error fraction per input is 3%, but 50.4% of all predicted labels are wrong. A 10% risk guarantee for the first quantity would permit a system that fails the second requirement. When each issued label or unit of capital matters, risk must weight an input by the exposure it creates. Moreover, many low-exposure observations can provide less evidence than the raw sample size suggests. A useful data budget must account for both the risk target and how exposure is distributed across inputs.

Existing risk-control methods provide a starting point. Risk-controlling prediction sets (Bates et al., 2021) and Learn-then-Test (LTT) (Angelopoulos et al., 2025) turn risk requirements into valid tests, including tests of bounded exposure-weighted losses. Validity, however, constrains the approval of unsafe policies without requiring frequent approval of safe ones. To assess the latter, we study power: the probability that a safe policy passes certification. A sufficient sample bound guarantees a desired power but can be loose, while an observed advantage over one test can reflect that test's conservatism. Neither alone identifies the data requirement imposed by the statistical problem.

We address this gap through an exact worst-case benchmark. We consider all bounded exposure and loss distributions with a specified mean exposure and safe risk. Following the mean-preserving randomization principle of Schlag (2007), we transform each observation into one of three outcomes:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

no retained unit, a correct unit, or an erroneous unit. Every population in the class gives the same transformed experiment, and that experiment is itself an admissible population. These two properties make its optimal certification power equal to the best worst-case power in the original problem, at every sample size. The result turns a problem over an unrestricted class of distributions into a finite calculation.

Our contributions are:

- **An exact, computable benchmark for certification cost.** We characterize the optimal finite-sample power for exposure-weighted risk certification and separate the smallest attainable budget from the budget of an implementable test and a sufficient bound for that test (Section 4). In a representative zero-risk setting, these budgets are 27, 29, and 360 observations, respectively. We also distinguish the oracle benchmark, which uses known mean exposure and a specified safe alternative, from procedures that remain valid without that knowledge.
- **A comparison of sampling and direct evidence.** We connect the likelihood ratio induced by exposure sampling to direct product betting on the original observations (Section 5). The two constructions share the same exposure-specific worst-case information scale, while direct tests can exploit more favorable population structure. We further show that, with the same betting coefficients, a quadratic approximation can discard evidence relative to the exact product.
- **An empirical study of power and policy selection.** Controlled populations isolate the role of exposure structure, and the Scene and Yeast tasks show how the resulting power differences affect the selection of trained multilabel policies (Section 6). Stronger direct tests reverse some apparent advantages over conservative comparators, whereas sampling remains useful at tighter risk caps. A lending study further separates the effects of candidate choice, joint constraints, and calendar transfer on deployment-oriented certification.

2 Related Work

Our analysis connects risk control for selective prediction with statistical tests of bounded observations. Selective classification decides which predictions to issue and measures risk on the selected population. Its basic ratio already permits weights in $[0, 1]$ and bounded losses (El-Yaniv & Wiener, 2010). Selective deep prediction and reject rules study the tradeoff between prediction quality and coverage (Geifman & El-Yaniv, 2017; 2019; Franc et al., 2023; Hendrickx et al., 2024). We use the same ratio structure, with weights that represent deployed units or capital, and characterize the data needed to certify it.

Risk-controlling prediction sets (Bates et al., 2021) and LTT (Angelopoulos et al., 2025) provide finite-sample guarantees, including selection among candidate policies. Conformal risk control (Angelopoulos et al., 2024) gives an expectation guarantee for monotone bounded loss families. Its multilabel false-negative example and LTT's per-image false-discovery example concern different losses. Prediction-set coverage (Romano et al., 2020) and probability calibration (Guo et al., 2017) address other aspects of predictive uncertainty. We fix an exposure-weighted population risk and control false certification with high probability. Related selective methods control deployment loss, selected conditional risk, or expert-call budgets (Bai & Jin, 2026; Xu et al., 2026; Pona et al., 2026). We instead ask how many observations a safe policy needs to pass and give an exact finite-sample benchmark for a fixed exposure-risk ratio.

Mean-preserving randomization supplies the statistical basis for this benchmark. Schlag (2007) develops the reduction of bounded vectors to minimax problems through correlated binary transformations. Rejection sampling and cost-proportionate resampling also retain observations according to bounded weights (Devroye, 1986; Zadrozny et al., 2003). We specialize the reduction to the pair consisting of exposure and exposure-weighted loss. Its triangular domain gives an explicit three-point experiment and a finite-sample likelihood-ratio solution. This characterization lets us compare an implemented test's data requirement with the statistical optimum and with sufficient proof bounds.

Product betting and coefficients chosen from past observations are established constructions for bounded means (Waudby-Smith & Ramdas, 2024). Arnold & Clerico (2026) optimize worst-case and relative worst-case expected log growth for single-round bounded-mean e-values. Waudby-Smith

et al. (2026) establish instance-wise asymptotic log optimality and matching expected rejection times in the high-confidence limit, under regret and moment conditions. These growth and time criteria differ from fixed-sample certification power. We connect the exposure-specific power benchmark to direct product wealth and its worst-case growth. Appendix Table 17 compares the result types.

3 Risk Per Unit Exposure

We first define the risk that a policy creates and the evidence available to certify it. This setup separates control of unsafe approvals from the probability of approving a safe policy, which is the quantity our sample budget analysis will address.

A frozen policy π maps an input X to a prediction or action. Let $E_\pi(X) \in [0, b_\pi]$ denote the resulting exposure, with a positive known cap b_π fixed before certification. Write $e_\pi = E_\pi / b_\pi \in [0,1]$ and let $L_\pi(X,Y) \in [0,1]$ be the average loss per exposed unit. Set $L_\pi = 0$ when $e_\pi = 0$. For $\mu_\pi = \mathbb{E}[e_\pi] > 0$, define

$$R_{\exp}(\pi) = \frac{\mathbb{E}[e_\pi L_\pi]}{\mu_\pi}, \qquad R_{\mathrm{macro}}(\pi) = \mathbb{E}[L_\pi]. \tag{1}$$

The cap must bound population exposure, rather than just the observed sample. A loose cap weakens certification, while clipping exposure changes the risk target. A common conversion of exposure units and the cap leaves both risks unchanged. Positive population exposure makes the ratio defined.

For multilabel prediction, let $C_\pi(X)$ be the predicted set in a vocabulary of K labels and Y the true set. With $m = |C_\pi(X)|$ and $f = |C_\pi(X) \setminus Y|$, take $e_\pi = m/K$ and $L_\pi = f/(m \vee 1)$. Then

$$R_{\exp} = \frac{\mathbb{E}[f]}{\mathbb{E}[m]}, \qquad R_{\mathrm{macro}} = \mathbb{E}\left[\frac{f}{m \vee 1}\right]. \tag{2}$$

We call the first quantity pooled false-prediction risk. LTT's multilabel example uses the second, per-example quantity. Lending provides another instance: exposure is approved principal and unit loss indicates default (Appendix F).

Having fixed the target, we separate the data used to construct a policy from the fresh data used to certify it. This separation is what allows the later selection guarantee to treat the candidate family as fixed.

Assumption 1 (Frozen development and fresh certification)**.** *Conditional on development information $\mathcal{F}_0$, the finite measurable candidate family Π, policies, exposure caps, loss functions, risk caps, fixed sample budget, testing parameters, and any candidate order are fixed. Certification observations are i.i.d. from a population P, independently of $\mathcal{F}_0$. Auxiliary sampling coins are independent of the data.*

Development includes training and any score or policy calibration. A valid returned set $\widehat{\Pi}_{\mathrm{cert}}$ obeys

$$\Pr\{\exists \pi \in \widehat{\Pi}_{\mathrm{cert}} : R_{\exp}(\pi) > \alpha_\pi \mid \mathcal{F}_0\} \le \delta. \tag{3}$$

Utility-based selection from that entire set preserves the guarantee:

$$\widehat{\pi} \in \arg\max_{\pi \in \widehat{\Pi}_{\mathrm{cert}}} \widehat{V}_{\mathrm{cal}}(\pi), \qquad V(\pi) = \mathbb{E}[u_\pi(X,Y)]. \tag{4}$$

An empty set authorizes no policy. The guarantee bounds the probability of returning any unsafe policy, rather than its risk conditional on a return.

This validity requirement raises the central quantitative question: how many fresh observations are needed to certify a safe fixed policy with high probability? The following population classes make that question precise.

Fix $\alpha, \beta \in (0,1)$, $0 < \delta < 1 - \beta$, and $\mu \in (0,1]$. Let $\mathcal{P}_{\mu,r}$ contain all populations of (e_π, L_π) on $[0,1]^2$, with $L_\pi = 0$ when $e_\pi = 0$, mean exposure $\mathbb{E}[e_\pi] = \mu$ and exposure risk $R_{\exp}(\pi) = r$. The null class is $\mathcal{P}_{\mu,\ge\alpha} = \bigcup_{q \in [\alpha,1]} \mathcal{P}_{\mu,q}$, with the same mean exposure. A certifier ϕ_n maps n fresh i.i.d. observations and independent internal randomness to a decision in $\{0,1\}$. It is valid at level δ if $\sup_{P \in \mathcal{P}_{\mu,\ge\alpha}} \Pr_P(\phi_n = 1) \le \delta$. The fixed-policy minimax sample complexity is

$$n^\star(\mu, r) = \min\left\{ n \in \mathbb{N}_0 : \exists \text{ a certifier } \phi_n \text{ valid at level } \delta \text{ with } \inf_{P \in \mathcal{P}_{\mu,r}} \Pr_P(\phi_n = 1) \ge 1 - \beta \right\}, \tag{5}$$

the fewest observations for which some level-δ certifier attains power $1-\beta$ for every population with parameters (μ, r), where $r < \alpha$. We take the minimum of an empty set to be $+\infty$. This definition fixes the mean exposure in both the null and alternative. Let $n^\star_{\mathrm{all}}(\mu, r)$ instead require validity over $\bigcup_{\mu' \in (0,1]} \mathcal{P}_{\mu', \geq \alpha}$, with the same alternative $\mathcal{P}_{\mu,r}$. Then $n^\star \leq n^\star_{\mathrm{all}}$. Section 4 establishes their common sample order and explains why their finite-sample optima can differ.

4 The Exact Worst-Case Certification Problem

We begin with the smallest data budget that any valid certifier can achieve on the hardest population. The key step is to reduce the observed exposure and loss to a three-outcome experiment, solve that finite problem, and then compare its optimum with an implementable sampling test. Fix a policy, write $\mu = \mu_\pi$, $r = R_{\exp}(\pi)$, and drop the policy subscript, with $\mu \in (0, 1]$. Let $\mathrm{kl}(u, v) = u \log(u/v) + (1-u)\log((1-u)/(1-v))$ with the usual endpoint conventions, so $\mathrm{kl}(0, v) = -\log(1-v)$ and $\mathrm{kl}(u, 0) = +\infty$ for $u > 0$.

The reduction represents each observation by at most one exposed unit. This construction removes variation in exposure magnitude while retaining the risk of interest. Draw independent uniforms U_i, V_i on $[0, 1)$, independent of the data, and retain example i when $U_i < e_i$. Record an erroneous unit when additionally $V_i < L_i$.

Lemma 1 (Exact reduction to Bernoulli certification). *Let $T_i = \mathbf{1}\{U_i < e_i\}$ and $B_i = T_i \mathbf{1}\{V_i < L_i\}$. Then $\Pr(T_i = 1) = \mu$ and $\Pr(B_i = 1) = \mathbb{E}[e_i L_i] = \mu r$, so conditional on retention the error indicator is $\mathrm{Bernoulli}(r)$. With $N_T = \sum_i T_i$ and $K_T = \sum_i B_i$,*

$$N_T \sim \mathrm{Bin}(n, \mu), \qquad K_T \mid N_T = m \sim \mathrm{Bin}(m, r). \tag{6}$$

The law in Equation 6 holds for every joint distribution of (e, L) on $[0, 1]^2$ and depends only on (μ, r). To include the minimum-count rule, define

$$p_T = \begin{cases} 1, & N_T < 5, \\ F_{N_T, \alpha}(K_T), & N_T \geq 5, \end{cases} \tag{7}$$

where $F_{m,\alpha}$ is the binomial cumulative distribution function. Then $\Pr(p_T \leq t) \leq t$ for every $t \in [0, 1]$ under $R_{\exp} \geq \alpha$. This count guard belongs to the implemented test; it does not restrict the minimax class of certifiers. The procedure retains at most one unit per example and allows arbitrary dependence within each example. For multilabel prediction, a uniform label slot supplies the retention coin. The observed correctness of that slot replaces the auxiliary variable V_i.

The sampling rule above is already implementable; the remaining question is whether any other valid certifier can do better in the worst case. To answer it, we specialize the mean-preserving binary-reduction principle of Schlag (2007) to the triangular observation (e, eL). The resulting count experiment yields an explicit exposure-risk benchmark. For fixed $\alpha \in (0, 1)$, $\delta \in (0, 1)$, $\mu \in (0, 1]$, and $r \in [0, \alpha)$, define the optimal worst-case power at budget n by

$$\mathcal{V}_n(\mu, r) = \sup_{\phi : \sup_{P \in \mathcal{P}_{\mu, \geq \alpha}} \mathbb{E}_{P^{\otimes n}} \phi \leq \delta} \inf_{P \in \mathcal{P}_{\mu, r}} \mathbb{E}_{P^{\otimes n}} \phi. \tag{8}$$

Here a measurable $\phi \in [0, 1]$ is a conditional probability of certification, implemented by an independent coin. The optimization fixes both μ and the design alternative r. Let $Q_{\mu,q}$ put probabilities $(1-\mu, \mu(1-q), \mu q)$ on $(e, L) = (0, 0), (1, 0), (1, 1)$, respectively.

Theorem 2 (Exact finite-sample minimax power). *Under Assumption 1, fix $\alpha, \delta \in (0, 1)$, $0 < \mu \leq 1$, $0 \leq r < \alpha$, and $n \in \mathbb{N}_0$. For the classes in Equation 8,*

$$\mathcal{V}_n(\mu, r) = \max_{\psi : \sup_{q \in [\alpha, 1]} \mathbb{E}_{Q_{\mu,q}^{\otimes n}} \psi \leq \delta} \mathbb{E}_{Q_{\mu,r}^{\otimes n}} \psi. \tag{9}$$

The maximum is attained by a likelihood-ratio test on the retained count N and error count K:

$$\Lambda_r(N, K) = \left(\frac{r}{\alpha}\right)^K \left(\frac{1-r}{1-\alpha}\right)^{N-K}, \qquad \psi^\star = \mathbf{1}\{\Lambda_r > c\} + \theta \mathbf{1}\{\Lambda_r = c\}, \tag{10}$$

where $c \geq 0$ and $\theta \in [0, 1]$ are calibrated under $Q_{\mu,\alpha}^{\otimes n}$ so that $\mathbb{E}\psi^\star = \delta$. Use $0^0 = 1$, including when $r = 0$ or $n = 0$. Its certification probability is nonincreasing in q, so calibration at $q = \alpha$ controls the entire composite null. Applying $\psi^\star$ after the randomized reduction of Lemma 1 attains Equation 8.

The equality follows from two facts: every admissible population reduces to the same count experiment, and the three-point law belongs to the original class. The optimum uses known μ and a fixed design risk r. The conditional binomial test needs neither parameter, but can have lower power. Appendix A gives the proof and boundary cases. The exact budget is $n^\star(\mu, r) = \min\{n : \mathcal{V}_n(\mu, r) \geq 1 - \beta\}$.

4.1 The optimal budget, a test budget, and a sufficient bound

The exact characterization above describes optimal power at every fixed sample size. To compare it with an implementable procedure, we now translate the power curves into three corresponding sample budgets. Let $n_T(\mu, r)$ be the smallest budget at which the guarded sampling test has power at least $1 - \beta$. Let $n_{\mathrm{suff}}(\mu, r)$ be the explicit budget in Theorem 8. These objects satisfy

$$n^\star(\mu, r) \leq n^\star_{\mathrm{all}}(\mu, r) \leq n_T(\mu, r) \leq n_{\mathrm{suff}}(\mu, r). \tag{11}$$

The first quantity permits fixed-mean oracle calibration. The second requires validity at every positive mean. The third fixes the implemented test, and the fourth uses a sufficient analysis of that test. Theorem 9 bounds every admissible certifier from below. Together, these bounds establish the following rate.

Corollary 3 (Exposure mass times entropy)**.** *Fix $\alpha, \beta \in (0, 1)$ and let $0 < \delta \leq (1 - \beta)/2$. For the minimax sample size $n^\star$ defined in Equation 5 over the bounded class,*

$$n^\star(\mu, r) = \Theta_{\alpha,\beta}\left(\frac{\log(1/\delta)}{\mu\,\mathrm{kl}(r, \alpha)}\right) \qquad \textit{for all } 0 < \mu \leq 1, \quad 0 \leq r < \alpha. \tag{12}$$

The constants can depend only on β, uniformly over $0 < \alpha \leq 1/2$ (Appendix C.3). Since $\mathrm{kl}(r, \alpha) \sim (\alpha - r)^2/[2\alpha(1 - \alpha)]$ as $r \uparrow \alpha$, the familiar inverse-squared-margin law $\Theta(\log(1/\delta)/[\mu(\alpha - r)^2])$ is the local form of Equation 12.

The same order holds for $n^\star_{\mathrm{all}}$ and n_T by equation 11. The bounds count original examples: the sampler retains $n\mu$ units on average, while the three-point family has per-example relative entropy $\mu\,\mathrm{kl}(r, q)$ against risk q. Rare unit-exposure errors make every other observation uninformative. At zero risk, the denominator remains $-\mu \log(1 - \alpha)$, so an absence of observed errors still requires enough exposure to exclude such errors. Population risk r describes power and cannot be set to zero merely because the sample contains no errors.

The rate leaves finite-budget differences unresolved. We compare the optimal budget, implemented test, and sufficient bound experimentally, and can also separate the two validity classes analytically.

The distinction between validity classes already appears with one observation. Take $\mu = 1/2$, $r = 0$, $\alpha = 1/10$, and $\delta = 1/20$. The fixed-mean optimum certifies with probability $1/18$, whereas the all-mean optimum is $19/360$. The latter must protect against null populations with a different proportion of unexposed observations. This example also explains why an estimated μ cannot simply replace known μ in the oracle calibration without an additional validity argument.

The benchmark is also computable. At budget n, we sort at most $(n + 1)(n + 2)/2$ count states by likelihood ratio and sum their null probabilities, randomizing on the final tied group. This uses $O(n^2 \log n)$ arithmetic operations, with additional integer-precision cost under rational arithmetic. A budget sweep then finds the first crossing of the target power.

5 Direct Tests and Population Structure

The previous section identifies the hardest population in the bounded class. We now ask how a test can exploit additional structure in easier populations. Instead of reducing each observation to a sampled unit, direct tests use its exposure and loss together. The key variable is $Z = e(\alpha - L)$, for which the null $R_{\exp} \geq \alpha$ is equivalent to $\mathbb{E}Z \leq 0$. We write $S = \sum_i Z_i$ and $Q = \sum_i Z_i^2$. The following results relate tests of this bounded mean to the sampling benchmark and explain why their power can differ across populations.

5.1 SAMPLING LIKELIHOOD RATIOS AND PRODUCT WEALTH

We begin by relating the sampling construction to a test that keeps the original exposure values. This connection makes clear what information is discarded by sampling and what a direct product test can retain. Bounded-mean betting (Waudby-Smith & Ramdas, 2024) forms a nonnegative product $M_n = \prod_i (1 + \lambda_i Z_i)$. If each stake is chosen before observation i and lies in $[0, 1/(1-\alpha)]$, then $\mathbb{E}[M_n] \le 1$ under the null. Markov's inequality makes $\min\{1, M_n^{-1}\}$ a valid fixed-time p-value, with value one at zero wealth. A fixed mixture of these products is also valid. For a uniform J-stake grid with $0 \le \lambda_j \le [2(1-\alpha)]^{-1}$, compare

$$\mathcal{E}_{\mathrm{prod}} = \frac{1}{J} \sum_j \prod_i (1 + \lambda_j Z_i), \qquad \mathcal{E}_{\mathrm{quad}} = \frac{1}{J} \sum_j \exp(\lambda_j S - \lambda_j^2 Q). \tag{13}$$

We convert each wealth to a p-value and require at least five positive-exposure observations. Unit sampling instead requires five retained units. These guards only remove rejections.

Proposition 4 (Sampling and direct product wealth)**.** *Under Assumption 1, $\mathcal{E}_{\mathrm{prod}} \ge \mathcal{E}_{\mathrm{quad}}$ on every data set, and $p_{\mathrm{prod}} \le p_{\mathrm{quad}}$ under identical guards. For $t \in [0, \alpha)$, put $\lambda_t = (\alpha - t)/(\alpha(1-\alpha))$ and let $\Lambda_t(N_T, K_T)$ be the likelihood ratio in equation 10 with design alternative t. Then*

$$\mathbb{E}[\Lambda_t(N_T, K_T) \mid (e_i, L_i)_{i=1}^n] = \prod_i [1 + \lambda_t e_i (\alpha - L_i)]. \tag{14}$$

The inequality follows from $\exp(x - x^2) \le 1 + x$ for $x \ge -1/2$. For the identity, average one likelihood-ratio factor over its three possible sampled states, then multiply the conditional expectations. Appendix B supplies both proofs and the validity argument for predictable stakes. The identity removes auxiliary randomness from a likelihood ratio. The product p-value and conditional binomial test use different calibrations, so it does not order their rejection powers.

5.2 A COMMON WORST-CASE INFORMATION SCALE

The preceding identity compares the two tests path by path. We next compare their population-level information and identify the case in which no direct test can improve on the reduction. Direct observations can contain information beyond the reduced counts. Define the best constant-stake log growth on a population P as

$$G(P) = \sup_{0 \le \lambda < 1/(1-\alpha)} \mathbb{E}_P \log(1 + \lambda e(\alpha - L)). \tag{15}$$

Theorem 5 (Worst-case logarithmic growth)**.** *For the bounded class $\mathcal{P}_{\mu,r}$ with $0 < \mu \le 1$ and $0 \le r < \alpha < 1$,*

$$\inf_{P \in \mathcal{P}_{\mu,r}} G(P) = \mu \, \mathrm{kl}(r, \alpha). \tag{16}$$

The three-point law $Q_{\mu,r}$ attains the infimum.

Concavity bounds each log factor below by its interpolation over the three vertices. The bound is tight on $Q_{\mu,r}$, whose optimal stake is $(\alpha - r)/(\alpha(1-\alpha))$ for $r > 0$. The case $r = 0$ follows by an endpoint limit. Thus the same information scale controls both worst-case testing cost and worst-case optimal growth. Particular populations can have larger growth. For example, known constant exposure makes $R_{\exp} = \mathbb{E}L$ and removes the need to exclude rare unit-exposure errors. Growth alone does not specify a finite-sample power curve.

Log growth describes asymptotic evidence accumulation, but it does not by itself determine performance at a fixed budget. Second moments give a complementary finite-sample statement. Write $d = \mathbb{E}Z = \mu(\alpha - r) > 0$ and $v = \mathbb{E}Z^2$. For $\alpha \le 1/2$, Proposition 18 gives a sufficient budget of order $[v/d^2 + (1-\alpha)/d][\log(J/\delta) + \log(2/\beta)]$ for the quadratic mixture, and hence for the same-grid product mixture, when the grid contains a stake in $[\lambda_*/2, \lambda_*]$, where $\lambda_* = \min\{1/[2(1-\alpha)], d/(2v)\}$. This bound applies to all $L \in [0, 1]$. It explains why equal mean exposure and risk need not give equal power for direct tests.

Our experiments compare fixed grids with a predictable risk plug-in, which estimates risk from past observations and clips the next coefficient to keep each factor positive. Appendix B.3 gives its definition and consistency proof.

5.3 INDEPENDENT CANDIDATE SEARCH

The results so far concern one policy fixed before certification. Deployment also requires choosing among many such policies without invalidating the certificate. We use LTT's fixed sequence and independent Pareto filtering (Angelopoulos et al., 2025; Laufer-Goldshtein et al., 2023). Development data freeze the candidates and their order. Each test uses level δ, and the sequence stops at its first non-rejection. Its probability of returning anything is exactly the first candidate's pass probability. Later candidates affect the certified set and utility. We compare orders based on estimated risk, utility, test evidence, and $\widehat{\mu}\,\mathrm{kl}(\widehat{r}, \alpha)$, setting the last score to zero for unsafe estimates or zero exposure. Bonferroni testing and intersection–union tests handle alternative searches and joint constraints (Berger, 1982). All procedures preserve equation 3, with their definitions in Appendix G.4.

6 EXPERIMENTS

The experiments examine three consequences of the analysis: how closely an implementable test approaches the exact budget, how exposure structure changes the value of direct observations, and whether those differences affect useful policy selection. We first describe the tasks and comparison design, then relate each result to the corresponding theoretical claim.

6.1 EXPERIMENTAL DESIGN

We use controlled populations to separate statistical effects and two public multilabel tasks to assess their practical consequences. Scene contains 2,407 images with six possible scene labels (Boutell et al., 2004), and Yeast contains 2,417 examples with 14 functional labels (Elisseeff & Weston, 2001). The task is to issue as many correct labels as possible while certifying the pooled false-prediction risk in equation 2. We train nonlinear support-vector predictors and form candidate policies by thresholding and gating their scores. All tests receive the same trained policies, development order, and certification observations. Five data splits separate training, development, certification, and evaluation. Within each split, 100 shared random permutations compare tests at the same budgets. Full model, grid, and split specifications appear in Appendices D and E.

With the tasks and data splits fixed, the remaining comparisons isolate how each test uses the available evidence. Unit sampling is the conditional binomial test from Section 4. Direct tests include the Hoeffding–Bentkus (HB) bounded-mean test used in LTT (Bentkus, 2004; Angelopoulos et al., 2025), quadratic and product mixtures with the same 16 coefficients, a product mixture over the wider valid range, and predictable adaptive betting. The HB test uses the loss $\widetilde{L} = \alpha - Z \in [0, 1]$ and conditions on positive exposure. We evaluate several sample budgets and risk caps at false-certification level $\delta = .05$. For a fixed safe policy, power is its probability of passing. For policy selection, return rate measures how often at least one policy passes, and utility counts correct labels per input. We average utility over all runs, counting an empty return as zero (U_0), and separately over returned runs (U_{ret}). Held-out empirical risk provides an additional check on the returned policies.

6.2 HOW CLOSE IS THE SAMPLING TEST TO THE OPTIMAL BUDGET?

We first compute the exact minimax budget and compare it with the sampling test and its sufficient bound. Table 1 shows that the implemented test can be much closer to the optimum than the bound suggests. In the zero-risk case, it needs 29 observations for 80% power, just two more than the randomized optimum, whereas the sufficient bound gives 360. The two nonzero-risk cases also show moderate test-to-optimum gaps alongside much larger sufficient bounds. These comparisons support the distinction between a statistical data requirement and a conservative guarantee. The remaining gap includes parameter knowledge, conditional calibration, boundary randomization, and the minimum-count rule. Appendix E gives the complete power curves and calculation details.

6.3 WHEN DO DIRECT OBSERVATIONS CONTAIN MORE USEFUL EVIDENCE?

The worst-case benchmark does not determine every population's difficulty. To examine the difference, we compare sparse unit exposure, distributed as $Q_{\mu,r}$, with constant exposure $e = \mu$ and the same Bernoulli loss of risk r. Both populations have the same sampling law, but constant exposure has a

Table 1: Sample budgets at level $\delta = .05$ and target power .8. The implemented sampling test (n_T) is close to the randomized fixed-mean optimum ($n^\star$) in these cases, while the sufficient bound (n_{suff}) is substantially larger. The optimum uses the specified mean and safe alternative.

μ	r	α	$n^\star$	n_T	n_{suff}
1	0	.1	27	29	360
.5	.05	.3	27	34	292
.25	.1	.5	31	42	296

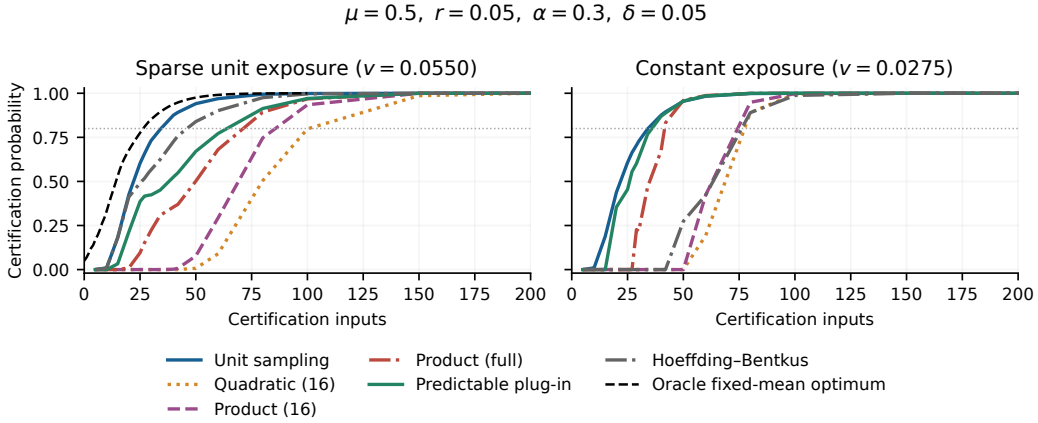

Figure 1: Certification power at equal mean exposure and risk. Direct tests benefit from constant exposure, while unit sampling has the same distribution in both panels. Each curve uses 2,000 independent runs per setting. The dashed black curve is the exact randomized optimum for the three-point family with known design parameters, rather than an upper bound for the constant-exposure family. Product (16) and Quadratic (16) share their coefficient grid. Product (full) uses the wider valid range.

smaller second moment of Z. We estimate power across sample budgets using 2,000 independent runs per setting. Figure 1 shows the resulting contrast: direct tests benefit from constant exposure, while sampling power stays unchanged apart from simulation noise. The result illustrates the mechanism in Section 5. A reduction can preserve the hardest population's power and still discard useful structure elsewhere.

6.4 Do these differences improve multilabel policy selection?

The controlled populations show that direct tests can exploit structure that sampling discards. We now ask whether that advantage survives policy search on real multilabel data. The public-data results show that the choice of direct test materially changes the comparison with sampling. At 500 certification inputs, adaptive betting raises returns on Scene at the 20% cap from 463/500 to 500/500 and U_0 from .5523 to .6024 (Table 2). On Yeast at the 30% cap, it raises returns from 358/500 to 463/500 and U_0 from .9793 to 1.3158. These results show that a fixed data budget can support more useful output through the test alone. Sampling remains competitive at tighter caps. On Scene at 10%, it returns 104/500 times while the compared direct tests return none. The five paired utility differences favor adaptation on Scene, but one favors sampling on Yeast. Thus an average gain does not imply a gain for every split or risk target.

The matched comparisons help explain these outcomes. With identical coefficients, the product test improves on its quadratic approximation: at the Scene 30% cap, returns rise from 477/500 to 490/500. A wider coefficient range yields further gains, and updating coefficients is not always necessary to obtain them. On Scene at 20%, the fixed full-range mixture achieves $U_0 = .6048$, slightly above the adaptive rule's .6024 despite fewer returns. On Yeast at 30%, adaptation improves on the full-range mixture. Appendix E.5 compares tests on identical fixed candidates, and Appendix E reports all caps,

Table 2: Multilabel policy selection with 500 certification inputs. Each block summarizes 100 shared draws on each of five data splits. U_0 includes empty returns as zero, while U_{ret} averages returned runs. Ex./ret. counts held-out empirical risk exceedances among returns. Conditional metrics are undefined when no policy is returned.

Data / α	Test	Returns	U_0	U_{ret}	Ex./ret.
Scene / 0.2	HB	0/500	0.000	N/A	N/A
Scene / 0.2	Quadratic 16	0/500	0.000	N/A	N/A
Scene / 0.2	Product 16	0/500	0.000	N/A	N/A
Scene / 0.2	Product full	487/500	0.605	0.621	0/487
Scene / 0.2	Adaptive	500/500	0.602	0.602	0/500
Scene / 0.2	Unit sampling	463/500	0.552	0.596	0/463
Yeast / 0.3	HB	0/500	0.000	N/A	N/A
Yeast / 0.3	Quadratic 16	3/500	0.007	1.144	0/3
Yeast / 0.3	Product 16	15/500	0.034	1.132	0/15
Yeast / 0.3	Product full	410/500	1.131	1.380	0/410
Yeast / 0.3	Adaptive	463/500	1.316	1.421	0/463
Yeast / 0.3	Unit sampling	358/500	0.979	1.368	1/358

budgets, and paired utilities. Together, the comparisons show why gains over a conservative direct test alone do not establish an advantage of exposure sampling.

6.5 Certification under lending constraints

The multilabel study concerns one pooled risk target. The lending application extends the same question to policies that must satisfy several deployment requirements. We evaluate loan approval and review policies on 589,488 mature loans from the Lending Club archive (George, 2019), controlling default risk by both loan count and principal while limiting review workload. At the reference contract, the same-quarter comparisons return policies in 14/15 or 15/15 runs, depending on the candidate order. Holding the order fixed removes this return difference across the compared tests, as predicted by the first-candidate property in Section 5. Calendar-separated evaluation also reveals empirical exceedances, so the fixed-population certificate does not resolve distribution shift. Appendix K provides the full lending results. This application illustrates why test power, candidate selection, and population change must be assessed separately.

7 Discussion and Conclusion

The analysis gives a way to interpret an unsuccessful risk certification. The exact three-point experiment measures the data needed in the worst case, while direct tests can use structure beyond that benchmark. The 27-versus-29 observation example shows that a practical sampling test can be close to the optimum even when its sufficient bound is much larger. The multilabel results show the other possibility: a conservative direct test can leave usable evidence untapped. A comparison of certificates therefore needs both a statistical benchmark and appropriate test choices.

The scope of each conclusion is explicit. The exact minimax benchmark uses known mean exposure and a specified safe alternative, whereas practical certifiers must remain valid with unknown population parameters. Our product identities concern wealth and growth, so finite-sample power also depends on the coefficient rule and calibration. The multilabel experiments reuse finite pools across five splits. The lending analysis covers mature accepted loans, with unresolved historical label availability and no outcomes for rejected applicants (Ehrhardt et al., 2021). Calendar transfer and cash-flow losses require assumptions beyond the fixed-population target. Within this scope, the paper separates unavoidable data cost from losses due to reduction, test design, and candidate search, giving a clearer basis for deciding how to use a limited certification budget.

AI USE STATEMENT

We used generative AI to polish the writing, check theoretical proofs, assist with the design and implementation of the experimental code, refine proofs, analyze the synthetic experiments, and support literature searches. We checked this work through code tests, independent numerical calculations, source verification, and comparisons with retained experimental records. The authors take responsibility for the final text, mathematical claims, experimental results, and research artifacts.

ETHICS STATEMENT

An aggregate certificate can mask subgroup disparities. Credit access, appeals, human review, and fairness requirements (Hardt et al., 2016) remain separate from the aggregate risk target. The lending experiments evaluate retrospective data; any deployment would require a separate assessment of these issues.

REPRODUCIBILITY STATEMENT

Appendices A, B, C, G, G.8, and G.9 give the assumptions and complete proofs. Appendices D, E, and L document the data splits, model settings, randomization, and reproduction commands. Anonymous supplementary code includes the implementations, experiment runners, frozen configuration generator, and tests. The project retains prediction caches, candidate p-values, orders, rejection masks, and raw repetitions for the reported figures and tables.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Sebastian Arnold and Eugenio Clerico. Optimal e-values for testing the mean of a bounded random variable against a composite alternative. *arXiv preprint arXiv:2601.11347v2*, 2026.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704v1*, 2026.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Vidmantas Bentkus. On Hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300, 1982.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

Adrien Ehrhardt, Christophe Biernacki, Vincent Vandewalle, Philippe Heinrich, and Sébastien Beben. Reject inference methods in credit scoring. *Journal of Applied Statistics*, 48(13–15):2734–2754, 2021.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019.

Nathan George. All Lending Club loan data. Kaggle, 2019. Version 3.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *Machine Learning*, 113(5):3073–3110, 2024.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Efficiently controlling multiple risks with Pareto testing. In *International Conference on Learning Representations*, 2023.

Edoardo Pona, Milad Kazemi, Mehran Hosseini, Yali Du, David Watson, Osvaldo Simeone, and Nicola Paoletti. Calibrate-then-delegate: Safety monitoring with risk and budget guarantees via model cascades. *arXiv preprint arXiv:2604.14251v2*, 2026.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020.

Karl H. Schlag. How to attain minimax risk with applications to distribution-free nonparametric estimation and testing. EUI Working Papers ECO 2007/04, European University Institute, 2007.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Ian Waudby-Smith, Ricardo J. Sandoval, and Michael I. Jordan. Instance-log-optimality of portfolio-based E-processes and their sequential hypothesis tests. *arXiv preprint arXiv:2504.02818v2*, 2026.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844v2*, 2026.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pp. 435–442. IEEE, 2003.

A EXACT FINITE-SAMPLE MINIMAX CERTIFICATION

The main text states the exact minimax characterization and uses it as a benchmark for practical tests. This section supplies the full reduction and finite-sample solution, then treats the boundary cases, the role of known mean exposure, and the finite-state calculation.

We condition on the fixed development information in Assumption 1. Throughout this section, $\alpha, \delta \in (0,1)$, $0 < \mu \le 1$, $0 \le r < \alpha$, and $n \in \mathbb{N}_0$. The sample consists of the complete pairs (e, L). The class $\mathcal{P}_{\mu,q}$ contains all laws on $[0,1]^2$ with $L = 0$ when $e = 0$, $\mathbb{E}e = \mu$, and $\mathbb{E}[eL] = \mu q$. Randomization is independent of the sample. Integrating out this randomization identifies any test with a measurable function into $[0,1]$. Conversely, a function ϕ is implemented by certifying when an independent uniform variable is less than ϕ.

A.1 THE REDUCTION PRESERVES THE EXACT MINIMAX VALUE

The first step is to show that replacing each bounded observation by a three-outcome draw loses no worst-case power. Once this equivalence is in place, the remaining optimization is finite dimensional.

Proof of Theorem 2. Define a Markov kernel $\mathcal{K}$ from an observation (e, L) to $\{(0,0), (1,0), (1,1)\}$ by the probabilities

$$\mathcal{K}((e,L), \cdot) = (1 - e, e(1-L), eL). \tag{17}$$

All three entries are nonnegative and sum to one. For every $P \in \mathcal{P}_{\mu,q}$, integration gives $P\mathcal{K} = Q_{\mu,q}$. Independent applications to the n observations therefore give $P^{\otimes n}\mathcal{K}^{\otimes n} = Q_{\mu,q}^{\otimes n}$. For a three-point test ψ, set $\phi = \mathcal{K}^{\otimes n}\psi$. This is a measurable $[0,1]$-valued function and satisfies, for every $P \in \mathcal{P}_{\mu,q}$,

$$\mathbb{E}_{P^{\otimes n}}\phi = \mathbb{E}_{Q_{\mu,q}^{\otimes n}}\psi. \tag{18}$$

Thus every feasible three-point test transfers with exactly the same size and power, proving that the left side of Equation 9 is at least its right side.

Conversely, $Q_{\mu,q} \in \mathcal{P}_{\mu,q}$ for every $q \in [0,1]$. Restrict any feasible full-observation test ϕ to the three-point sample space and call this restriction ψ. It obeys the three-point null constraints, while

$$\inf_{P \in \mathcal{P}_{\mu,r}} \mathbb{E}_{P^{\otimes n}}\phi \le \mathbb{E}_{Q_{\mu,r}^{\otimes n}}\psi. \tag{19}$$

Taking suprema proves the opposite inequality. No interchange of a supremum and an infimum is used.

It remains to solve the finite three-point experiment and prove attainment. Write N for the number of nonzero-exposure observations and K for the number of erroneous exposed observations. Its count mass is

$$w_q(m,k) = \binom{n}{m}\binom{m}{k}(1-\mu)^{n-m}\mu^m(1-q)^{m-k}q^k, \quad 0 \le k \le m \le n. \tag{20}$$

For $\mu = 1$, use only the support $m = n$; for $\mu < 1$, all count pairs have positive mass under $q = \alpha$. The reference law $Q_{\mu,\alpha}^{\otimes n}$ dominates every $Q_{\mu,q}^{\otimes n}$ on the common possible sample space. On that space, the likelihood ratio of the r law to the α law is exactly $\Lambda_r(N,K)$ in Equation 10. The multinomial multiplicities and exposure factors cancel. This also holds on the ordered sample space, before reducing it to counts.

There are finitely many distinct likelihood-ratio values. Choose a value $c \ge 0$ with

$$a := \Pr_{\mu,\alpha}(\Lambda_r > c) \le \delta \le \Pr_{\mu,\alpha}(\Lambda_r \ge c) = a + b, \qquad b := \Pr_{\mu,\alpha}(\Lambda_r = c) > 0, \tag{21}$$

and put $\theta = (\delta - a)/b$. Such a value exists by traversing all distinct values in decreasing order. The rule in Equation 10 has size exactly δ at α. A single common θ is used on the entire tied level, including states with different values of N.

For completeness, the finite Neyman–Pearson argument is immediate from a pointwise inequality, so no external optimal-testing result is needed. For any $[0,1]$-valued test η with $\mathbb{E}_{\mu,\alpha}\eta \le \delta$, $(\psi^\star - \eta)(\Lambda_r - c) \ge 0$ at every point. Consequently,

$$\begin{aligned}\mathbb{E}_{\mu,r}(\psi^\star - \eta) &= \mathbb{E}_{\mu,\alpha}[(\psi^\star - \eta)(\Lambda_r - c)] + c\,\mathbb{E}_{\mu,\alpha}(\psi^\star - \eta) \\ &\ge 0.\end{aligned} \tag{22}$$

This proves optimality against the simple null at α even among ordered-sample tests, not only count-based tests.

We next establish feasibility for the entire composite null. If $r > 0$, then for every m and $k < m$,

$$\frac{\Lambda_r(m, k+1)}{\Lambda_r(m, k)} = \frac{r(1-\alpha)}{\alpha(1-r)} < 1. \tag{23}$$

If $r = 0$, then $\Lambda_0(m, 0) = (1-\alpha)^{-m}$ and $\Lambda_0(m, k) = 0$ for $k \geq 1$, so it is still nonincreasing in k. Since $x \mapsto \mathbf{1}\{x > c\} + \theta\mathbf{1}\{x = c\}$ is nondecreasing, $k \mapsto \psi^\star(m, k)$ is nonincreasing. Couple $K_q = \sum_{j=1}^m \mathbf{1}\{U_j < q\}$ using the same independent uniforms. Then K_q is nondecreasing in q, and hence $\mathbb{E}[\psi^\star(m, K_q)]$ is nonincreasing in q. The distribution $N \sim \mathrm{Bin}(n, \mu)$ is independent of q; averaging over N proves

$$\sup_{q \in [\alpha, 1]} \mathbb{E}_{\mu,q}\psi^\star = \mathbb{E}_{\mu,\alpha}\psi^\star = \delta. \tag{24}$$

Thus the simple-null optimum is feasible for the composite null. Together with the two transfer inequalities, this proves the theorem, including attainment in the original problem. □

A.2 Boundary cases and exact sample budgets

The reduction gives the general solution, but several endpoints simplify and help interpret it. We record those cases before comparing different null classes.

At $n = 0$, the sample space has one point, $\Lambda_r = 1$, and choosing $c = 1$, $\theta = \delta$ gives $\mathcal{V}_0 = \delta$. This is the power of an independent random certification decision; the unrestricted minimax class allows it. The implemented minimum-count guard is an additional restriction of that particular test.

At $\mu = 1$, $N = n$ almost surely, and the optimal test is the randomized one-sided binomial test. No conditioning on an impossible count is needed. For $r = 0$, the alternative has $K = 0$ almost surely. At the null boundary,

$$\Pr_{\mu,\alpha}(K = 0) = (1 - \mu\alpha)^n. \tag{25}$$

If δ exceeds this mass, the construction can take $c = 0$ and

$$\theta = \frac{\delta - (1-\mu\alpha)^n}{1 - (1-\mu\alpha)^n}. \tag{26}$$

The denominator is positive whenever this case occurs. The power is then one. At equality, take $c = 0$, $\theta = 0$; below it, the threshold is positive and selects zero-error states in decreasing order of N. When $n = 0$, the preceding single-point rule applies instead.

If $\mu = 1$ and $r = 0$, this specializes to

$$\mathcal{V}_n(1, 0) = \min\{1, \delta(1-\alpha)^{-n}\}. \tag{27}$$

Indeed, the only state of positive alternative probability has null mass $(1-\alpha)^n$, so the size constraint permits certification there with probability at most $\delta/(1-\alpha)^n$, capped at one; the likelihood ratio rule attains this value. For $\beta \in (0, 1)$ with $\delta < 1 - \beta$,

$$n^\star(1, 0) = \left\lceil \frac{\log((1-\beta)/\delta)}{-\log(1-\alpha)} \right\rceil. \tag{28}$$

The guarded, nonrandomized binomial rule instead has exact budget

$$n_T(1, 0) = \max\left\{5, \left\lceil \frac{\log(1/\delta)}{-\log(1-\alpha)} \right\rceil\right\}. \tag{29}$$

This formula holds for any target power in $(0, 1]$, since its power at $r = 0$, $\mu = 1$ jumps from zero to one. At $(\alpha, \delta, \beta) = (.1, .05, .2)$, the respective budgets are 27 and 29. The exact powers at $n = 26, 27$ are approximately .773864 and .859849. The earlier information lower bound rounds up to 19 and the sufficient condition gives 360.

The optimal value is nondecreasing in n: a procedure with $n + 1$ observations can ignore the last observation and implement any n-observation procedure. Attainment therefore gives the exact budget $\min\{n : \mathcal{V}_n \geq 1 - \beta\}$, including zero if $\delta \geq 1 - \beta$. This argument does not assert monotonicity of a particular nonrandomized binomial test, whose discrete rejection threshold can cause power dips. If the unsafe null is written as $q > \alpha$, its optimum is unchanged: for every finite-sample test, the three-point rejection probability is a polynomial in q, and the constraints for $q > \alpha$ imply the one at α by continuity. The kernel construction then gives the same upper and lower comparisons.

A.3 KNOWN MEAN EXPOSURE VERSUS A UNIVERSAL NULL

The preceding optimum calibrates the null at a known mean exposure. We now remove that knowledge to show exactly where the stronger, universal null can reduce finite-sample power.

Let $\mathcal{V}_n^{\mathrm{all}}(\mu, r)$ replace the null in Equation 8 by $\bigcup_{\nu\in(0,1]} \mathcal{P}_{\nu,\geq\alpha}$, while retaining the alternative class $\mathcal{P}_{\mu,r}$. Its value is at most $\mathcal{V}_n(\mu, r)$ because its feasible set is smaller. The same kernel and embedded-family proof give the exact reduction

$$\mathcal{V}_n^{\mathrm{all}}(\mu, r) = \max_{\psi:\, \sup_{\nu\in(0,1],\, q\in[\alpha,1]} \mathbb{E}_{Q_{\nu,q}^{\otimes n}}\psi\leq\delta} \mathbb{E}_{Q_{\mu,r}^{\otimes n}}\psi. \tag{30}$$

The finite-dimensional feasible set is closed and bounded, being the intersection of the box $[0, 1]^{3^n}$ with closed halfspaces, so this maximum is attained. The fixed-μ likelihood-ratio calibration need not satisfy these stronger constraints.

An exact separation occurs at $n = 1$, $\mu = 1/2$, $r = 0$, $\alpha = 1/10$, and $\delta = 1/20$. Write the test probabilities on the three atoms as (x_0, x_1, x_2). The fixed-μ optimum is $(0, 1/9, 0)$, with power $1/18$. Its size at $\nu = 1$, $q = \alpha$ is $1/10 > \delta$. For the universal null, taking $\nu \downarrow 0$ implies $x_0 \leq \delta$, while $\nu = 1$, $q = \alpha$ implies $(1 - \alpha)x_1 + \alpha x_2 \leq \delta$, hence $x_1 \leq \delta/(1 - \alpha)$. The alternative power is therefore at most

$$\tfrac{1}{2}(x_0 + x_1) \leq \tfrac{1}{2}\left(\delta + \frac{\delta}{1-\alpha}\right) = \frac{19}{360} < \frac{1}{18}. \tag{31}$$

The rule $(\delta, \delta/(1 - \alpha), 0)$ attains the bound, since for every $\nu \in (0, 1]$ and $q \geq \alpha$ its size is $(1 - \nu)\delta + \nu(1 - q)\delta/(1 - \alpha) \leq \delta$. The reduction transfers it to all bounded populations. Thus equality of sample-complexity orders does not imply equality of the finite-sample power curves for the two null classes.

A.4 FINITE-STATE CALCULATION AND RESOURCE ACCOUNTING

Having established the statistical characterization, we turn to its exact evaluation. The construction below implements the likelihood-ratio rule and also makes its computational cost explicit.

For $\mu < 1$, enumerate the $(n + 1)(n + 2)/2$ pairs (m, k); for $\mu = 1$, only $n + 1$ pairs are possible. Compute the two masses w_α, w_r from Equation 20, sort the ratios w_r/w_α in decreasing order, and aggregate equal ratios into complete levels. Fill the null probability budget until a level would reach or exceed δ. On that level use $\theta = (\delta - a)/b$ from Equation 21, and assign zero to all lower levels. The resulting exact power is

$$\mathcal{V}_n(\mu, r) = \sum_{m,k} w_r(m, k)[\mathbf{1}\{\Lambda_r(m, k) > c\} + \theta\mathbf{1}\{\Lambda_r(m, k) = c\}]. \tag{32}$$

A useful independent certificate is the finite dual identity

$$\mathcal{V}_n(\mu, r) = c\delta + \sum_{m,k}[w_r(m, k) - cw_\alpha(m, k)]_+. \tag{33}$$

It follows by substituting the threshold rule: below c the positive part is zero, above c it is $w_r - cw_\alpha$, and the tie contribution to the power is c times its allocated null mass.

Precomputing powers and binomial coefficients gives $O((n + 1)^2)$ mass operations. Sorting and summation use $O((n + 1)^2 \log(n + 2))$ arithmetic/comparison operations and $O((n + 1)^2)$ storage, with constants independent of the statistical parameters. These are arithmetic, not unit-cost bit, bounds. If μ, r, α, δ are rational numbers of at most B bits, all masses can use common denominators of at most $O(nB)$ bits; binomial coefficients, ratios, partial sums, and the randomized threshold require $O((n + 1)(B + \log(n + 2)))$ bits. Exact rational operations on integers of this size give a finite exact algorithm. For arbitrary real inputs, the stated operation count presumes exact real arithmetic and comparisons; finite precision requires an additional numerical tolerance and tie policy. After calibration, implementing the kernel and accumulating (N, K) takes $O(n)$ operations and $O(1)$ additional counters, followed by one boundary randomization.

The accompanying independent script uses rational arithmetic for both probability sums and tie comparisons. It also maximizes the small-n linear program on the original ordered three-point sample

space by exhaustively enumerating its vertices. For a box intersected with one size constraint, every vertex has at most one coordinate strictly between zero and one: two such coordinates would admit a nonzero feasible perturbation preserving the size. This provides a separate finite check of the likelihood-ratio calculation. These checks support implementation verification; the proof above establishes the parameter-uniform claim.

B Direct Betting: Validity, Information, and Adaptation

Section 5 introduced direct betting as a way to retain the exposure information discarded by sampling. Here we justify that method in full: we establish validity, connect product wealth to the sampling likelihood ratio, and analyze the predictable risk plug-in used in the experiments. We condition on development information and use the fixed-policy setting of Assumption 1. Write $Z_i = e_i(\alpha - L_i)$ and $M_n = \prod_{i=1}^{n}(1 + \lambda_i Z_i)$, with $M_0 = 1$. The risk class has $L = 0$ at zero exposure as in Section 3.

B.1 Product validity and the conditional identity

We first verify that the product construction is valid under predictable stakes. The same calculation then reveals its precise connection to the likelihood ratio induced by exposure-unit sampling.

Proof of Proposition 4. Let $\mathcal{F}_{i-1}$ include development information and all previous certification observations. Predictability means that λ_i is $\mathcal{F}_{i-1}$-measurable. Since $Z_i \geq -(1-\alpha)$, every factor in M_n is nonnegative, and independence gives

$$\mathbb{E}[1 + \lambda_i Z_i \mid \mathcal{F}_{i-1}] = 1 + \lambda_i \mu(\alpha - r) \leq 1.$$

Multiplication by the nonnegative M_{i-1} proves the supermartingale property, including paths where wealth is zero. Fixed convex mixing preserves it. With $p_n = \min\{1, M_n^{-1}\}$ (and $p_n = 1$ at $M_n = 0$), Markov's inequality gives $\Pr(p_n \leq u) \leq u$ for $u \in (0,1)$. For $x \geq -1/2$, the function $f(x) = \log(1+x) - x + x^2$ has derivative $f'(x) = x(1+2x)/(1+x)$, is decreasing on $[-1/2, 0]$ and increasing on $[0, \infty)$, and has minimum $f(0) = 0$. Thus $\exp(x - x^2) \leq 1 + x$. Applying this to every $x = \lambda_j Z_i$, multiplying across observations, and summing with the same nonnegative weights proves the pathwise inequality. Replacing either p-value by one on a common guard-failure event preserves its validity and the stated ordering.

For the conditional identity in equation 14, the conditional expectation of the likelihood-ratio factor contributed by row i is

$$\begin{aligned}(1 - e_i) + e_i(1 - L_i)\frac{1-t}{1-\alpha} + e_i L_i \frac{t}{\alpha} \\ = 1 + \frac{\alpha - t}{\alpha(1-\alpha)} e_i(\alpha - L_i).\end{aligned}$$

Auxiliary draws are independent across rows conditional on all original observations, so the conditional expectation of their product is the product of these expectations. The formula also holds at $t = 0$ using the explicit convention in the statement. □

B.2 The exposure-specific worst-case growth rate

Validity alone does not describe when direct betting is informative. We now compute its least favorable log-growth rate and compare it with the information scale of the sampling reduction.

Proof of Theorem 5. Fix $0 \leq \lambda < 1/(1-\alpha)$, so all logarithms are finite and bounded. Concavity, first along the exposure coordinate and then along the loss coordinate, gives the pointwise inequalities

$$\begin{aligned}\log(1 + \lambda e(\alpha - L)) &\geq e \log(1 + \lambda(\alpha - L)) \\ &\geq e\{(1 - L)\log(1 + \lambda\alpha) + L \log(1 - \lambda(1-\alpha))\}.\end{aligned}$$

Taking expectations yields $\mathbb{E}_P \log(1 + \lambda Z) \geq \mu g_r(\lambda)$, where

$$g_r(\lambda) = (1 - r)\log(1 + \lambda\alpha) + r \log(1 - \lambda(1-\alpha)).$$

For $r > 0$, differentiation shows that g_r is maximized at the interior point $\lambda_* = (\alpha - r)/(\alpha(1-\alpha))$. At this point $1 + \lambda_*\alpha = (1-r)/(1-\alpha)$ and $1 - \lambda_*(1-\alpha) = r/\alpha$, hence $g_r(\lambda_*) = \mathrm{kl}(r, \alpha)$. For $r = 0$, g_0 is increasing and its supremum as $\lambda \uparrow 1/(1-\alpha)$ is $-\log(1-\alpha) = \mathrm{kl}(0, \alpha)$. Thus every P has $G(P) \geq \mu\,\mathrm{kl}(r, \alpha)$. On the stated three-point law, both concavity inequalities are equalities for every allowed λ, so its supremum is exactly this value. No interchange of an infimum and a supremum is needed. □

B.3 The predictable risk plug-in baseline

The worst-case calculation uses the best constant stake for a known alternative. For the experiments, we instead need a data-adaptive rule that remains valid without knowing the risk in advance. The numerical comparison fixes $\tau = 2$ and $\varepsilon = 10^{-6}$ before certification and uses

$$\widehat{r}_{i-1} = \frac{\tau\alpha/2 + \sum_{k<i} e_k L_k}{\tau + \sum_{k<i} e_k}, \qquad \lambda_i = \min\left\{\frac{1-\varepsilon}{1-\alpha}, \max\left\{0, \frac{\alpha - \widehat{r}_{i-1}}{\alpha(1-\alpha)}\right\}\right\}. \tag{34}$$

These stakes satisfy the predictable validity conditions proved above, so the terminal inverse wealth is valid. The prior contributes a fixed exposure mass τ at risk $\alpha/2$; it is not an estimate from certification outcomes. This baseline is not asserted to maximize growth on each law.

Proposition 6 (Consistency of the predictable baseline)**.** *For every fixed law with $\mu > 0$ and $r < \alpha$, the wealth from equation 34 grows exponentially almost surely. Consequently, its fixed-time guarded test has power tending to one at any fixed level $\delta > 0$.*

Proof. The strong law gives $\widehat{r}_{i-1} \to r$ almost surely. Therefore $\lambda_i \to \bar{\lambda} = \min\{(1-\varepsilon)/(1-\alpha), (\alpha - r)/(\alpha(1-\alpha))\} > 0$. On the allowed compact stake interval, $1 + \lambda Z \geq \varepsilon$, and the derivative of $\log(1 + \lambda Z)$ with respect to λ has absolute value at most $1/\varepsilon$, because $|Z| \leq 1$. The Cesàro average of $|\lambda_i - \bar{\lambda}|$ tends to zero. Hence

$$\frac{1}{n}\log M_n - \frac{1}{n}\sum_{i=1}^{n} \log(1 + \bar{\lambda} Z_i) \longrightarrow 0 \quad \text{almost surely.}$$

The second term tends to $\mathbb{E}\log(1 + \bar{\lambda}Z)$ by the strong law for bounded variables. This limit is at least $\mu g_r(\bar{\lambda}) > 0$: g_r is strictly increasing between zero and its positive maximizing stake (or its endpoint supremum when $r = 0$), and $g_r(0) = 0$. Also $\Pr(e > 0) \geq \mu > 0$, so eventually there are five positive-exposure rows and positive total exposure almost surely. The guarded p-value therefore tends to zero almost surely, which implies the power claim. □

C Exposure-Weighted Certification: Proofs and Scope

The exact minimax result gives a benchmark, while the main text uses a simpler sampling test with explicit complexity bounds. This section fills in that second part of the analysis: we prove the sampling law, establish the midpoint comparison, derive the upper and lower sample bounds, and prove Corollary 3. All inferential statements condition on the development information $\mathcal{F}_0$ of Assumption 1; a fixed policy is a measurable function of a fresh observation, not of its certification labels. Write $\mu = \mathbb{E}[e] \in (0, 1]$, $r = R_{\exp} = \mathbb{E}[eL]/\mu$, and $\mathrm{kl}(u, v) = u\log(u/v) + (1-u)\log((1-u)/(1-v))$, with $\mathrm{kl}(0, v) = -\log(1-v)$ and $\mathrm{kl}(u, 0) = +\infty$ for $u > 0$.

C.1 Exact reduction to Bernoulli certification

We begin with the probabilistic step used throughout the paper: converting fractional exposure and loss into a retained-unit count. The proof also shows why arbitrary dependence within an example does not affect the resulting binomial law.

Proof of Lemma 1. Fix example i with $(e_i, L_i) \in [0, 1]^2$, and draw U_i, V_i uniform on $[0, 1)$, mutually independent and independent of the data. Set $T_i = \mathbf{1}\{U_i < e_i\}$ and $Z_i = \mathbf{1}\{V_i < L_i\}$, so $B_i = T_i Z_i$. Conditioning on (e_i, L_i),

$$\Pr(T_i = 1 \mid e_i, L_i) = e_i, \qquad \Pr(T_i = 1, Z_i = 1 \mid e_i, L_i) = e_i L_i, \tag{35}$$

because $U_i \perp V_i$. Taking expectations gives $\Pr(T_i = 1) = \mathbb{E}[e_i] = \mu$ and $\Pr(B_i = 1) = \mathbb{E}[e_i L_i] = \mu r$, hence $\Pr(B_i = 1 \mid T_i = 1) = \mu r/\mu = r$. The quadruples (e_i, L_i, U_i, V_i) are i.i.d. across i, so the pairs (T_i, B_i) are i.i.d. with the three-cell law

$$\Pr(T_i = 1, B_i = 1) = \mu r, \quad \Pr(T_i = 1, B_i = 0) = \mu(1 - r), \quad \Pr(T_i = 0) = 1 - \mu. \tag{36}$$

The marginal count $N_T = \sum_i T_i$ follows $\mathrm{Bin}(n, \mu)$. Conditional on $N_T = m$, the retained error count $K_T = \sum_i B_i$ counts successes in m independent draws with success probability $\mu r/\mu = r$. It therefore follows $\mathrm{Bin}(m, r)$, as stated in Equation 6. The conditional statement applies to counts with positive probability, including the point mass at $N_T = n$ when $\mu = 1$.

To verify the p-value, fix $t < 1$ and $m \geq 5$. If an integer k satisfies $F_{m,\alpha}(k) \leq t$, let k_t be the largest such integer. A coupling with common uniforms makes a binomial count with parameter $r \geq \alpha$ at least as large as the count with parameter α. Thus

$$\Pr_r(p_T \leq t \mid N_T = m) = F_{m,r}(k_t) \leq F_{m,\alpha}(k_t) \leq t. \tag{37}$$

If no k_t exists, the event is empty. For $m < 5$ the guarded p-value is one, so the rejection probability is zero. Averaging over the count proves $\Pr_r(p_T \leq t) \leq t$ for every joint law on $[0, 1]^2$; the case $t = 1$ is immediate. For multilabel prediction, the retention coin selects one of the K label slots uniformly, so $e_i = m_i/K$ when the model predicts m_i labels. We observe the retained slot's correctness directly and use at most one label per example, without V_i. □

C.2 The midpoint entropy comparison

The sampling law reduces certification to binomial tails. To compare the upper and lower budgets on a common scale, we next relate the two midpoint tail exponents to $\mathrm{kl}(r, \alpha)$.

Lemma 7 (Midpoint entropy comparison)**.** *Let $0 \leq r < \alpha < 1$ and $s = (r + \alpha)/2$. Then*

$$\mathrm{kl}(r, \alpha) \leq \frac{(\alpha - r)^2}{\alpha(1 - \alpha)}, \qquad \min\{\mathrm{kl}(s, \alpha), \mathrm{kl}(s, r)\} \geq \frac{(\alpha - r)^2}{2}, \tag{38}$$

and consequently,

$$\min\{\mathrm{kl}(s, \alpha), \mathrm{kl}(s, r)\} \geq c_\alpha \, \mathrm{kl}(r, \alpha), \qquad c_\alpha = \begin{cases} 1/8, & 0 < \alpha \leq 1/2, \\ \alpha(1 - \alpha)/2, & 1/2 < \alpha < 1. \end{cases} \tag{39}$$

The constant is independent of α when $\alpha \leq 1/2$.

Both parts use the following integral representation for $u, v \in (0, 1)$. We treat the endpoint $r = 0$ explicitly below:

$$\mathrm{kl}(u, v) = \int_v^u \frac{u - t}{t(1 - t)} \, dt, \tag{40}$$

obtained from $\frac{d}{dt} \mathrm{kl}(u, t) = (t - u)/[t(1 - t)]$ and $\mathrm{kl}(u, u) = 0$. The integrand keeps the sign of $u - t$, so the integral is nonnegative in either direction.

Proof of Lemma 7. For the upper bound, take $0 \leq r < \alpha$, so $u = r < v = \alpha$ and equation 40 integrates over $t \in [r, \alpha]$ where $1 - t \geq 1 - \alpha$. Hence

$$\mathrm{kl}(r, \alpha) \leq \frac{1}{1 - \alpha} \int_r^\alpha \frac{t - r}{t} \, dt = \frac{1}{1 - \alpha} \big[(\alpha - r) - r \log(\alpha/r)\big]. \tag{41}$$

Substituting $x = r/\alpha \in [0, 1)$, the bracket equals $\alpha \, h(x)$ with $h(x) = (1 - x) + x \log x$ (and $h(0) = 1$). A direct computation gives

$$(1 - x)^2 - h(x) = x\big[x - 1 - \log x\big] \geq 0, \tag{42}$$

since $\log x \leq x - 1$. Therefore $h(x) \leq (1 - x)^2$ and

$$\mathrm{kl}(r, \alpha) \leq \frac{\alpha(1 - x)^2}{1 - \alpha} = \frac{(\alpha - r)^2}{\alpha(1 - \alpha)}. \tag{43}$$

At $r = 0$ this reads $-\log(1-\alpha) \le \alpha/(1-\alpha)$.

For the lower bounds, since $t(1-t) \le 1/4$, equation 40 gives the Pinsker inequality $\mathrm{kl}(u,v) \ge 4\int_{v\wedge u}^{v\vee u} |u-t|\,dt = 2(u-v)^2$ for all u, v. With $s = (r+\alpha)/2$ and $\alpha - s = s - r = (\alpha - r)/2$,

$$\mathrm{kl}(s,\alpha) \ge 2(s-\alpha)^2 = \tfrac{1}{2}(\alpha-r)^2, \qquad \mathrm{kl}(s,r) \ge 2(s-r)^2 = \tfrac{1}{2}(\alpha-r)^2, \tag{44}$$

which is the second inequality of Equation 38. Combining equation 43 and equation 44,

$$\min\{\mathrm{kl}(s,\alpha), \mathrm{kl}(s,r)\} \ge \frac{(\alpha-r)^2}{2} = \frac{\alpha(1-\alpha)}{2} \cdot \frac{(\alpha-r)^2}{\alpha(1-\alpha)} \ge \frac{\alpha(1-\alpha)}{2}\,\mathrm{kl}(r,\alpha), \tag{45}$$

which gives the stated c_α for $\alpha > 1/2$.

For the sharper constant when $\alpha \le 1/2$, on $[0,\alpha]$ the map $t \mapsto t(1-t)$ increases, so $t(1-t) \le \alpha(1-\alpha)$ and $1/[t(1-t)] \ge 1/[\alpha(1-\alpha)]$ throughout $[r,\alpha]$. Applying this to equation 40 at both midpoints,

$$\min\{\mathrm{kl}(s,\alpha), \mathrm{kl}(s,r)\} \ge \frac{1}{\alpha(1-\alpha)} \cdot \frac{(\alpha-r)^2}{8} = \frac{1}{8} \cdot \frac{(\alpha-r)^2}{\alpha(1-\alpha)} \ge \frac{1}{8}\,\mathrm{kl}(r,\alpha), \tag{46}$$

which gives $c_\alpha = 1/8$ for $\alpha \le 1/2$. At $r = 0$, $\mathrm{kl}(s,0) = +\infty$, and the finite integral for $\mathrm{kl}(s,\alpha)$ gives the same lower bound. We use this piecewise convention for c_α throughout. □

C.3 Sufficient sample size and its explicit constant

With the reduction and entropy comparison in hand, we can turn retained-unit concentration into a sufficient budget in the original sample size. The following result keeps the constants explicit so that the numerical gap can be assessed later.

Theorem 8 (Sufficient data for unit sampling). *Let $\gamma, \beta \in (0,1)$ and $0 \le r < \alpha < 1$. Under Assumption 1, put $s = (r+\alpha)/2$ and*

$$m_{\mathrm{cert}} = \max\left\{5, \left\lceil \frac{\log(1/\gamma)}{\mathrm{kl}(s,\alpha)} \right\rceil, \left\lceil \frac{\log(2/\beta)}{\mathrm{kl}(s,r)} \right\rceil \right\}, \tag{47}$$

where m_{cert} is a sufficient retained-unit threshold for certification and the third term is absent at $r = 0$. If an integer $n \ge 0$ satisfies $n\mu \ge \max\{2m_{\mathrm{cert}}, 8\log(2/\beta)\}$, then $\Pr(p_T \le \gamma) \ge 1-\beta$. With $\gamma = \delta \le (1-\beta)/2$, Lemma 7 yields

$$n \;\ge\; \frac{C_{\alpha,\beta}\log(1/\delta)}{\mu\,\mathrm{kl}(r,\alpha)} \quad\Longrightarrow\quad \textit{power} \;\ge 1-\beta, \tag{48}$$

for an explicit constant $C_{\alpha,\beta}$ depending only on the cap and the target power.

Proof of Theorem 8. For $K \sim \mathrm{Bin}(m,q)$, the exponential moment is $\mathbb{E}[e^{tK}] = (1-q+qe^t)^m$. Markov's inequality, optimized at $e^t = x(1-q)/[q(1-x)]$, gives the binomial Chernoff bounds (Hoeffding, 1963):

$$\Pr(K/m \le x) \le e^{-m\,\mathrm{kl}(x,q)}\ (x < q), \qquad \Pr(K/m \ge x) \le e^{-m\,\mathrm{kl}(x,q)}\ (x > q). \tag{49}$$

The lower-tail formula at $x = 0$ is exact. The cases with $q = 0$ or $q = 1$ follow from the corresponding constant counts. Condition on $\{N_T = m\}$. For $m \ge m_{\mathrm{cert}}$ the binomial Chernoff bound $F_{m,\alpha}(k) \le \exp\{-m\,\mathrm{kl}(k/m,\alpha)\}$ for $k/m < \alpha$, together with the fact that $\mathrm{kl}(\cdot,\alpha)$ decreases below α, gives $p_T \le \gamma$ on the event $K_T/m \le s$. The upper-tail Chernoff bound at the true retained parameter r gives $\Pr(K_T/m > s \mid N_T = m) \le \exp\{-m\,\mathrm{kl}(s,r)\} \le \beta/2$ when $r > 0$; when $r = 0$, $K_T = 0$ almost surely and this term is absent, matching the dropped third entry of Equation 47. Finally $n\mu \ge 2m_{\mathrm{cert}}$ and the lower-tail Chernoff bound for $\mathrm{Bin}(n,\mu)$ give $\Pr(N_T < m_{\mathrm{cert}}) \le \exp\{-n\mu/8\} \le \beta/2$ whenever $n\mu \ge 8\log(2/\beta)$. For completeness, if $\lambda = n\mu$, the count bound follows from

$$\Pr(N_T \le \lambda/2) \le e^{\lambda/4}(1-\mu+\mu e^{-1/2})^n \le e^{\lambda(1/4+e^{-1/2}-1)} \le e^{-\lambda/8}, \tag{50}$$

using $1 + z \le e^z$ and $e^{-1/2} \le 1 - 1/2 + 1/8$. A union bound over the two failure modes yields power at least $1-\beta$.

For the closed form, set $\gamma = \delta \le (1-\beta)/2$, write $L = \log(1/\delta) \ge \log\frac{2}{1-\beta} \ge \log 2$, and $\kappa = \mathrm{kl}(r, \alpha)$, which satisfies $0 < \kappa \le \mathrm{kl}(0, \alpha) = -\log(1-\alpha)$ because $\mathrm{kl}(\cdot, \alpha)$ decreases on $[0, \alpha)$. By Lemma 7, $\mathrm{kl}(s, \alpha)$ and $\mathrm{kl}(s, r)$ are each at least $c_\alpha \kappa$, so

$$m_{\mathrm{cert}} \le \max\Big\{5,\ \frac{L}{c_\alpha \kappa} + 1,\ \frac{\log(2/\beta)}{c_\alpha \kappa} + 1\Big\}. \tag{51}$$

We bound each term by a constant multiple of $L/(c_\alpha \kappa)$. Since $L \ge \log\frac{2}{1-\beta}$ we have $\log(2/\beta) \le C_1 L$ with $C_1 = \log(2/\beta)/\log\frac{2}{1-\beta}$, and since $\kappa \le -\log(1-\alpha)$ and $L \ge \log 2$, we also have $1 \le C_3\, L/(c_\alpha \kappa)$ with $C_3 = c_\alpha(-\log(1-\alpha))/\log 2$. Hence $m_{\mathrm{cert}} \le C_4\, L/(c_\alpha \kappa)$ with $C_4 = \max\{5C_3,\ 1 + C_3,\ C_1 + C_3\}$. The two requirements are $n\mu \ge 2m_{\mathrm{cert}}$ and $n\mu \ge 8\log(2/\beta) \le 8C_1(-\log(1-\alpha))\, L/\kappa$, so the sufficient condition holds whenever

$$n \ \ge\ \frac{C_{\alpha,\beta}\log(1/\delta)}{\mu\,\mathrm{kl}(r,\alpha)}, \qquad C_{\alpha,\beta} = \max\Big\{\tfrac{2C_4}{c_\alpha},\ 8C_1(-\log(1-\alpha))\Big\}, \tag{52}$$

which depends only on the cap α and target power $1-\beta$. This is Equation 48.

If $0 < \alpha \le 1/2$, the convention $c_\alpha = 1/8$ and the inequality $-\log(1-\alpha) \le \log 2$ give $C_3 \le 1/8$ and $C_4 \le \max\{9/8, C_1 + 1/8\}$. Therefore the complete sufficient constant has the uniform upper bound

$$C_{\alpha,\beta} \le C_\beta := \max\{18, 16C_1 + 2\}. \tag{53}$$

This bound includes the minimum count and the random retention cost, not only the midpoint entropy comparison. □

C.4 Necessary sample size and the error budget

The sufficient condition shows what unit sampling can achieve. We now use a matching hard family to show that no uniformly valid certifier can improve the dependence on exposure mass, risk separation, or the error budget.

Theorem 9 (Necessary data over the bounded class)**.** *Fix $0 \le r < \alpha < 1$, $\beta \in (0,1)$, and $0 < \delta < 1-\beta$. Consider the family whose exposure is 1 with probability μ and 0 otherwise, with unit loss $\mathrm{Bernoulli}(q)$ conditional on positive exposure and $L = 0$ on zero exposure for every q. Suppose a certifier uses n fresh observations and independent internal randomness. If it satisfies $\Pr(\textit{certify}) \le \delta$ whenever $q > \alpha$ and attains power $1-\beta$ at parameter r, then*

$$n \ \ge\ \frac{\mathrm{kl}(1-\beta,\delta)}{\mu\,\mathrm{kl}(r,\alpha)} \ \ge\ \frac{c_\beta \log(1/\delta)}{\mu\,\mathrm{kl}(r,\alpha)}, \tag{54}$$

where the second inequality holds for $\delta \le (1-\beta)/2$ with $c_\beta = \mathrm{kl}(1-\beta, (1-\beta)/2)/\log(2/(1-\beta)) > 0$.

Proof of Theorem 9. Use the family from Section 4, with exposure 1 with probability μ and 0 otherwise. Conditional on positive exposure, the unit loss is $\mathrm{Bernoulli}(q)$, while $L = 0$ on zero exposure for every q, so $R_{\mathrm{exp}} = q$. For $q' > q$, the populations P_q and $P_{q'}$ differ only in this loss parameter and have per-observation divergence $\mathrm{KL}(P_q \| P_{q'}) = \mu\,\mathrm{kl}(q, q')$. Fix a certifier with certification event C using n fresh observations and independent internal randomness, with $\Pr_{P_r}(C) \ge 1-\beta$ and $\Pr_{P_q}(C) \le \delta$ for every $q > \alpha$. The internal randomness has the same law under both populations, so it carries no information and the joint input divergence is $n\mu\,\mathrm{kl}(r,q)$. Write $\overline{P}_q = P_q^{\otimes n} \otimes \nu$ for the full input law, where ν is the common law of the independent random seed. Let $a_C = \Pr_{\overline{P}_r}(C)$ and $b_C = \Pr_{\overline{P}_q}(C)$. Splitting the likelihood-ratio integral over C and its complement and applying Jensen's inequality to each normalized integral gives $\mathrm{KL}(\overline{P}_r \| \overline{P}_q) \ge a_C \log(a_C/b_C) + (1-a_C)\log((1-a_C)/(1-b_C))$. This is the log-sum, or data-processing, inequality for the binary statistic $\mathbf{1}_C$, with limiting conventions at zero probabilities. Consequently,

$$n\mu\,\mathrm{kl}(r,q) \ \ge\ \mathrm{kl}\big(\Pr_{P_r}(C), \Pr_{P_q}(C)\big) \ \ge\ \mathrm{kl}(1-\beta,\delta), \tag{55}$$

where the last inequality uses the monotonicity of binary relative entropy on the region where its first argument exceeds its second. Here $\Pr_{P_r}(C) \ge 1-\beta > \delta \ge \Pr_{P_q}(C)$. Letting $q \downarrow \alpha$ and using continuity of $\mathrm{kl}(r,\cdot)$ at $\alpha < 1$ gives $n\mu\,\mathrm{kl}(r,\alpha) \ge \mathrm{kl}(1-\beta,\delta)$, the first inequality of Equation 54.

For the second inequality define, on $\delta \in (0, (1-\beta)/2]$,

$$\phi(\delta) = \frac{\mathrm{kl}(1-\beta, \delta)}{\log(1/\delta)}, \qquad c_\beta = \inf_{0<\delta\le(1-\beta)/2} \phi(\delta). \tag{56}$$

Expanding $\mathrm{kl}(1-\beta,\delta) = (1-\beta)\log(1/\delta) + (1-\beta)\log(1-\beta) + \beta\log\beta - \beta\log(1-\delta)$ and dividing by $\log(1/\delta)$ shows $\phi(\delta) \to 1-\beta$ as $\delta \downarrow 0$, so ϕ extends continuously to $[0, (1-\beta)/2]$ with $\phi(0) = 1-\beta > 0$. A continuous function on a compact interval that is strictly positive everywhere attains a positive minimum, so $c_\beta > 0$; the closed range excludes the singular point $\delta = 1-\beta$. This proves $\mathrm{kl}(1-\beta,\delta) \ge c_\beta \log(1/\delta)$ and hence Equation 54. To locate the infimum, put $a = 1-\beta$ and define binary entropy $H(x) = -x\log x - (1-x)\log(1-x)$. For $0 < z < a$,

$$\frac{d}{dz}\frac{\mathrm{kl}(a,z)}{\log(1/z)} = \frac{(1-a)H(z)/(1-z) - H(a)}{z\log^2(1/z)} < 0. \tag{57}$$

Indeed, the derivative of $H(z)/(1-z)$ is $-\log z/(1-z)^2 > 0$, so $H(z)/(1-z) < H(a)/(1-a)$ when $z < a$. Thus ϕ decreases, and its infimum over $(0, (1-\beta)/2]$ equals its value at the right endpoint. Table 4 reports this explicit c_β. □

Proof of Corollary 3. Theorem 8 gives a certifier meeting power $1-\beta$ with $n = O_{\alpha,\beta}(\log(1/\delta)/[\mu\,\mathrm{kl}(r,\alpha)])$, and Theorem 9 shows every valid certifier over the bounded class needs $n = \Omega_{\alpha,\beta}(\log(1/\delta)/[\mu\,\mathrm{kl}(r,\alpha)])$. To account for integer budgets, put $B = \log(1/\delta)/[\mu\,\mathrm{kl}(r,\alpha)]$ and $K_\alpha = -\log(1-\alpha)$. Since $B \ge \log 2/K_\alpha$,

$$c_\beta B \le n^\star \le \lceil C_{\alpha,\beta} B \rceil \le (C_{\alpha,\beta} + K_\alpha/\log 2)B. \tag{58}$$

This proves equation 12. For $\alpha \le 1/2$, the upper constant is at most $C_\beta + 1$ by equation 53, so both sides are uniform in the cap. A second-order Taylor expansion of $\mathrm{kl}(\cdot,\alpha)$ at α gives $\mathrm{kl}(r,\alpha) = \frac{(\alpha-r)^2}{2\alpha(1-\alpha)}(1+o(1))$ as $r \uparrow \alpha$. At $r = 0$, the denominator instead equals the finite value $-\mu\log(1-\alpha)$. □

C.5 Boundary cases

The two bounds share the same interior scaling, but their limiting cases are also informative. The following table collects those endpoints and states how each follows from the finite-sample results above.

Table 3 records the limiting behavior of the characterization; each row is consistent with the finite-sample statements above and with the endpoint conventions.

Table 3: Boundary behavior of the exposure-weighted complexity $n^\star \asymp \log(1/\delta)/[\mu\,\mathrm{kl}(r,\alpha)]$. Every entry is a limit of the finite-sample bounds, not a separate assumption.

Regime	Governing quantity	Behavior
$r = 0$	$\mathrm{kl}(0,\alpha) = -\log(1-\alpha)$	finite; $n^\star \asymp \log(1/\delta)/[\mu(-\log(1-\alpha))]$
$r \uparrow \alpha$	$\mathrm{kl}(r,\alpha) \to 0$	$n^\star \asymp \log(1/\delta)/[\mu(\alpha-r)^2] \to \infty$
$\mu = 1$	every example exposed	reduces to mean certification; no $1/\mu$ penalty
$\mu \downarrow 0$	exposure mass vanishes	$n^\star \propto 1/\mu \to \infty$ at fixed r, α
$\delta = (1-\beta)/2$	$c_\beta = \phi(\delta)$ (Table 4)	tightest budget in the stated range
$\alpha > 1/2$	$c_\alpha = \alpha(1-\alpha)/2$	constant degrades from $1/8$; rate unchanged
multilabel dependence	sampling law equation 36	arbitrary within-example dependence allowed

C.6 Numerical verification record

We close the theoretical appendix with numerical checks of the analytic inequalities and constants. These calculations verify the implementation; they do not replace the proofs above.

The script `scripts/verify_exposure_complexity.py` checks the analytic inequalities on dense grids and reports their numerical margins. It uses exact binomial tails rather than Monte Carlo estimates and evaluates $\alpha \in \{.02, .05, .09, .1, .2, .3, .5, .7, .9, .98\}$, $\beta \in \{.1, .2, .3, .5\}$, and 20,001-point risk and level grids. The grids provide numerical checks alongside the proofs above.

Table 4: Verification summary. Upper block: the sufficient budget of Theorem 8 attains the target power on exact binomial tails, staying within a factor 12.8–20.0 of the necessary budget of Theorem 9 ($\beta = .2$, $\delta = .05$). Lower block: the error-budget constant $c_\beta = \phi((1-\beta)/2)$ and its limit $1-\beta$. The midpoint constant c_α is met with slack $\geq 1.5\times$; Pinsker holds with ratio ≥ 1; the range-only test has log-log slope -2.00 against -1.00 for unit sampling. We set n_{nec} to the ceiling of the lower bound when computing the displayed ratios.

μ	r	α	m_{cert}	n_{suff}	exact power	$n_{\mathrm{suff}}/n_{\mathrm{nec}}$
1.000	.00	.10	180	360	1.0000	18.9
0.500	.05	.10	796	3,184	1.0000	13.9
0.050	.09	.10	21,247	849,880	1.0000	12.8
0.005	.00	.10	180	72,000	1.0000	19.9
0.200	.00	.20	82	820	1.0000	19.1
0.200	.15	.20	1,485	14,850	1.0000	13.0
0.100	.25	.30	1,980	39,600	1.0000	12.8
0.020	.00	.09	202	20,200	1.0000	20.0
0.020	.08	.09	19,297	1,929,700	1.0000	12.8

Error budget $\phi(\delta) = \mathrm{kl}(1-\beta, \delta)/\log(1/\delta)$

β	$\delta_{\max}$	c_β	limit $1-\beta$
.10	.450	0.5678	0.900
.20	.400	0.3654	0.800
.30	.350	0.2412	0.700
.50	.250	0.1038	0.500

D Multilabel Protocol and Full Diagnostic Scope

The main text reports the multilabel comparison at the level needed to interpret the findings. This section supplies the shared protocol behind those results and then gives the linear-model and aggregation diagnostics that complement the primary nonlinear comparison.

D.1 Public data, models, and split independence

We begin by fixing the data partitions and prediction models, because every subsequent test comparison relies on the same frozen candidates. The split design keeps evaluation labels outside model fitting, calibration, and policy search. We use the Mulan project's public Scene and Yeast training/test files and XML label descriptions. We verify their contents against the upstream Git blob identifiers and retain SHA-256 hashes of every file. We read label columns from XML metadata rather than their positions. Scene contains 1,211 training and 1,196 test rows with 294 real-valued features and six labels. Yeast contains 1,500 training and 917 test rows with 103 features and 14 labels. The complete label vector belongs to one example throughout.

Five fixed, label-independent permutations divide each data set into model-fitting, development, certification-pool, and evaluation partitions. The exact seeds and row assignments are recorded in the supplementary configurations. We fit the standardization parameters only on the model-fitting rows. The linear classifier is one-versus-rest logistic regression, and the nonlinear classifier is one-versus-rest support vector classification with a radial basis kernel and internal probability estimation. Their fixed implementation settings are also recorded in the configurations. We reserve evaluation labels for the final assessment and exclude them from preprocessing, model fitting, probability estimation, and candidate selection.

We added the nonlinear model after the linear model yielded few certificates. Both models use the same input partitions, threshold families, sample budgets, caps, and random streams. We report them together and perform no hyperparameter search on the evaluation set.

D.2 Candidates and risk targets

With the prediction models fixed, we next define the policies that will be searched and the risk each policy must satisfy. This common candidate family ensures that later differences come from

certification rather than model training. The candidate family is the Cartesian product of 19 label thresholds $t \in \{.05, .10, \ldots, .95\}$ and three input gates $g \in \{0, .8, .9\}$. For frozen scores $s_k(x)$, the policy is

$$C_{t,g}(x) = \begin{cases} \{k : s_k(x) \geq t\}, & \max_k s_k(x) \geq g, \\ \varnothing, & \text{otherwise.} \end{cases}$$

The raw records retain all 57 candidate identifiers, including candidates with equivalent predictions. Development Pareto filtering resolves exact coordinate ties by original index. Every order receives the same frontier minimizing $(\widehat{R}_{\text{exp}}, -\widehat{V})$; empirical zero exposure is represented by risk one for ordering and an evidence score of zero. Utility is the number of correct predicted labels per input. The fixed order uses stable candidate identifiers to break score ties.

For each of 100 repetitions, a random permutation of the reserved certification pool supplies a prefix of length 100, 250, or 500. We pair the three budgets through these prefixes and share each prefix across all tests and orders. We draw a fresh uniform vocabulary slot for each example and repetition and share it across candidate policies. Its retention probability is m/K, and its retained-error probability is f/K. Conditional on the retained count, the binomial test uses risk $\mathbb{E}[f]/\mathbb{E}[m]$. The 100 repetitions share a fixed finite pool, so they represent dependent resamples rather than independent datasets. We preserve this distinction when reporting uncertainty.

D.3 Testing and summary conventions

The candidate construction determines what may be selected; this subsection specifies how each candidate is tested and how repeated runs are summarized. These conventions are shared by all reported multilabel comparisons. The primary direct HB comparator conditions on positive exposure. The all-example HB transform, range Hoeffding, quadratic betting, wide-grid quadratic betting, exposure-unit sampling, and macro-HB are also retained. Macro-HB tests $\mathbb{E}[f/(m \vee 1)]$, not pooled risk; using the same candidate frontier is valid because the frontier remains independent of certification data. Its target differs from the primary comparison.

At every fixed $\alpha \in \{.1, .2, .3\}$ and $\delta = .05$, we evaluate all seven tests with five frontier orders: risk, utility, exposure score, development HB, and development wide betting. Bonferroni tests the original 57 candidates. The fixed-sequence test stops at its first failure, and the returned policy maximizes calibration correct-label utility over the certified prefix, breaking ties by original index. Each arm has its own fixed-procedure guarantee; the multi-cap diagnostic grid is not a simultaneously certified menu.

Return counts sum 100 repetitions across five splits. We average conditional utility and risk over returns within each split, then over splits with returns. We leave missing conditional values undefined and separately report utility under the convention that no return yields zero output. Empirical exceedance counts record returned repetitions whose held-out ratio exceeds the cap. These counts can include the same policy repeatedly and therefore do not estimate independent failures or the population event in Equation 3.

D.4 Known-population aggregation check

The preceding experiments target risk per predicted label. To show why this choice matters, we now compare it with a macro risk that averages once per input, using a population for which the two targets differ sharply.

The two-type example has probability .95 of $(e, L) = (.01, 0)$ and probability .05 of $(e, L) = (1, .6)$. Therefore $\mu = .0595$, $R_{\text{macro}} = .03$, and $R_{\text{exp}} = .03/.0595 = .50420168\ldots$. We convert the fractional loss to a binary observation using independent exposure and loss coins. At $n = 100$, macro-HB returns 3,060/5,000 times; at $n = 500$ and $2,000$ it returns 5,000/5,000 times. All six micro-risk implementations return zero times at all three budgets. The two-sided 95% exact binomial interval for zero out of 5,000 is $[0, .0007375]$; for 5,000 out of 5,000 it is $[.9992625, 1]$. These intervals quantify simulation uncertainty at each setting, not simultaneous coverage across the complete table.

Table 5: Linear-model reference with the same protocol, $n = 500$ and the development evidence-score order. Conditional metrics use the legacy split-averaging convention described above; Table 2 pools returned runs. Tight-cap failures are retained. Ex./ret. divides empirical exceedances by returns; N/A denotes an empty return.

Data	α	Test	Return	Utility	$\widehat{R}_{\exp}$	Ex./ret.
Scene	0.2	HB	0/500	–	–	N/A
Scene	0.2	Wide betting	0/500	–	–	N/A
Scene	0.2	Unit sampling	0/500	–	–	N/A
Scene	0.3	HB	0/500	–	–	N/A
Scene	0.3	Wide betting	0/500	–	–	N/A
Scene	0.3	Unit sampling	138/500	0.492	0.249	17/138
Yeast	0.2	HB	0/500	–	–	N/A
Yeast	0.2	Wide betting	0/500	–	–	N/A
Yeast	0.2	Unit sampling	7/500	0.536	0.209	5/7
Yeast	0.3	HB	0/500	–	–	N/A
Yeast	0.3	Wide betting	0/500	–	–	N/A
Yeast	0.3	Unit sampling	124/500	0.852	0.222	0/124

Table 6: Mean within-split Spearman correlations with single-policy unit-sampling pass frequency, nonlinear models and $n = 500$. Higher is more predictive. Risk-only uses $-\widehat{r}$, mass-only uses $\widehat{\mu}$, and development HB uses its negative p-value. Five splits contribute, except four nonconstant splits for Yeast at $\alpha = .1$. The correlations describe ranking agreement among dependent candidates.

Data	α	Risk only	Mass only	Mass–KL	Dev. HB
Scene	0.1	0.944	-0.861	0.882	0.873
Scene	0.2	0.968	-0.867	0.974	0.970
Scene	0.3	0.919	-0.796	0.943	0.928
Yeast	0.1	0.241	-0.188	-0.049	-0.039
Yeast	0.2	0.638	-0.559	0.716	0.814
Yeast	0.3	0.639	-0.487	0.906	0.937

D.5 Numerical and artifact checks

Finally, we record the checks that connect the reported summaries to the frozen predictions and run artifacts. This audit is separate from the statistical comparisons above but supports their reproducibility. We retain data hashes, package versions, fitted estimators, split indices, complete probability matrices, per-example labels, calibration subset membership, sampled slots, candidate p-values, orders, selected identifiers, and rejection masks. Direct tests use the actual fractional-loss moment $\sum_i e_i^2(\alpha - L_i)^2$, rather than the binary simplification. The local numerical library emitted floating-point status warnings during linear-model matrix products; all fitted coefficients and outputs were finite. We recomputed held-out predictions with explicit elementwise sums instead of matrix multiplication. Predictions agreed to machine precision across the linear fits, and the run records retain the original warnings.

We generate the main summary with `scripts/build_general_exposure_artifacts.py`. The two configuration files are `configs/general_multilabel.json` and `configs/general_multilabel_rbf.json`; `scripts/run_multilabel_exposure.py` is the shared entry point. The controlled aggregation run uses `scripts/simulate_aggregation_targets.py`. Every completed result directory contains its actual configuration, environment, raw records, logs, and completion marker. The original linear run is preserved separately from the replay that captures all numerical warnings.

E Additional Power and Utility Comparisons

The main experiments focus on representative settings. Here we widen the comparison across budgets, risk caps, and data splits, then use growth and fixed-policy diagnostics to explain why test power and selected-policy utility need not move together.

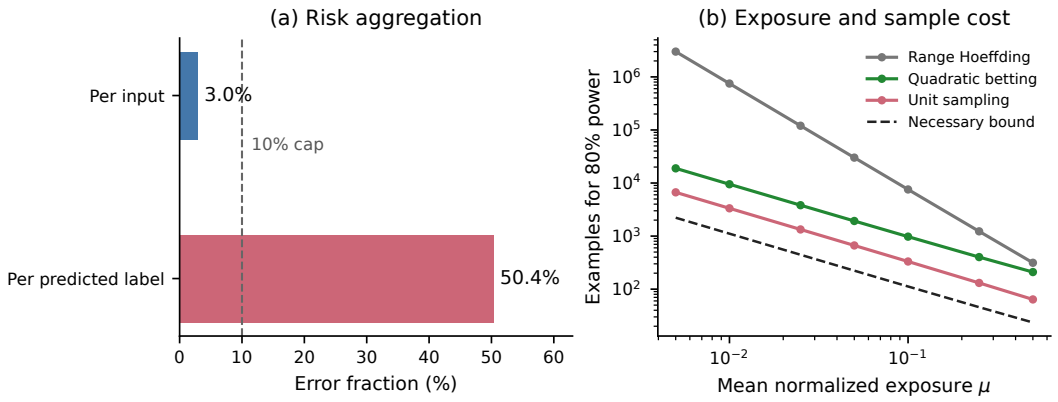

Figure 2: Risk units and evidence requirements. (a) The same population has 3% error per input and 50.4% error per predicted label. (b) Exact sample sizes for 80% power on a zero-loss population with exposure a on half the inputs and zero otherwise, so $\mu = a/2$, at $\alpha = .1, \delta = .05$. The direct range test has inverse-squared exposure cost; unit sampling and the quadratic betting mixture have inverse-exposure cost. The dashed contamination lower bound permits rare unit-exposure errors (Appendix F); it is not a per-population lower bound under a known constant-exposure restriction.

E.1 Frozen-cache comparison of direct tests

We compare tests on frozen radial-basis-kernel predictions for Scene and Yeast. Each data set has five train/development/calibration-pool/evaluation splits, 57 threshold–gate candidates, and 100 cached calibration permutations per split. The tests share evaluation sets, development frontiers and orders, and uniform label-slot draws. This design keeps model fitting and candidate construction from confounding the comparison. Certification uses permutation prefixes of $50, 100, 150, 250, 350, 500$ observations, all within the cached horizon of 500. Larger budgets extend the same reserved pool. Repeated draws and overlapping splits therefore support descriptive paired comparisons rather than independent replications.

E.1.1 Tests and guards

The frozen candidates allow a clean comparison of the certification rules. We therefore specify the stake grids and guard conditions before defining the evaluation metrics. Set $c = 1 - \alpha$. The same-grid controls use either 16 fixed stakes $\mathrm{geomspace}(1/(64c), 1/(2c), 16)$ or the horizon-dependent grid $\lambda_j = (2c)^{-1}2^{-j}$ for $j = 0, \ldots, \lceil \log_2 n \rceil + 1$. For each grid, the uniform product mixture is paired with the quadratic mixture. A further full-range product control uses 16 geometric stakes between $1/(64c)$ and $(1 - 10^{-6})/c$. The predictable plug-in baseline is exactly equation 34, with prior mass 2, prior risk $\alpha/2$, and endpoint buffer 10^{-6}. Its stakes have the same maximum as the full-range control. Hence the same-grid comparison isolates the quadratic approximation, whereas comparisons involving the adaptive baseline must also account for stake range and weighting. Every direct test requires five positive-exposure rows. Unit sampling uses the shared label slots and requires five retained units. HB is the Hoeffding–Bentkus test conditioned on positive exposure; all-row HB and Hoeffding remain in the result archive.

E.1.2 Local and selected-policy evaluations

The tests can differ both on a single fixed policy and after searching a candidate sequence. We report both views so that gains in test power are not automatically attributed to better policy selection. For each split and cap, three fixed policies are chosen on development information alone: the first frontier candidate in the risk, evidence, and HB orders. We report each policy's development and evaluation estimates of mean exposure, exposure risk, and the second moment of $Z = e(\alpha - L)$, together with every local pass frequency. The full 57-policy records are also retained. Selection uses the fixed sequences and the Bonferroni comparison, with calibration utility breaking choices using the same rule for every test. A fixed sequence returns precisely when its first candidate passes; later candidates affect which policy is returned and its utility.

For run b, define $U_{0,b}$ as the selected policy's held-out utility, or zero for an empty return. We report $\widehat{U}_0 = B^{-1}\sum_b U_{0,b}$, conditional utility averaged over returned runs, return frequency, certified candidate count, and empirical risk exceedances divided by returns. An empty-return row has undefined conditional utility and an exceedance entry of N/A (0/0). Empirical exceedances use the finite evaluation set; they are not exact counts of population risk violations. Each metric is computed from individual run records, followed by split-level paired differences. In particular, U_0 is not reconstructed by multiplying a split-averaged conditional utility by an aggregate return rate.

E.1.3 KNOWN-POPULATION EXPERIMENT

The frozen-cache study mixes policy search with population heterogeneity. To isolate the mechanism predicted by the theory, we also compare sparse and constant exposure under matched values of (μ, r). The three (μ, r, α) settings are $(1, 0, .1)$, $(.5, .05, .3)$, and $(.25, .1, .5)$. We use 2,000 independent Monte Carlo runs at each setting. In the sparse-unit family, $e \sim \text{Bernoulli}(\mu)$ and a latent loss $B \sim \text{Bernoulli}(r)$ independently, with observed $L = B$ when $e > 0$ and $L = 0$ when $e = 0$. In the constant-exposure family, $e \equiv \mu$ and the same Bernoulli loss law is used. Both families have identical μ, r, but their second moments are respectively $\mu\{(1-r)\alpha^2 + r(1-\alpha)^2\}$ and $\mu^2\{(1-r)\alpha^2 + r(1-\alpha)^2\}$. Common random streams pair methods and structures, and nested prefixes pair budgets. The oracle fixed-mean power curve is obtained by full count-state enumeration and shown only on the sparse-unit family. The Monte Carlo standard error of a reported frequency $\widehat{p}$ is $\sqrt{\widehat{p}(1-\widehat{p})/2000}$, at most .0112.

E.1.4 REPRODUCTION

The preceding protocol is fully determined by the archived predictions, random draws, and configurations. For completeness, the following commands reproduce the corresponding analyses from the repository root with the listed scientific Python dependencies available. Completed result directories are immutable, so a new execution should use fresh output paths.

```
rev_out=results_revision_20260926
PYTHONPATH=src python scripts/replay_strong_betting.py \
  --include-full-grid \
  --output "$rev_out/strong_betting_full"
PYTHONPATH=src python scripts/simulate_strong_betting.py \
  --include-full-grid \
  --output "$rev_out/strong_betting_synthetic_full_canonical"
PYTHONPATH=src python scripts/verify_exact_minimax.py --verify
PYTHONPATH=src python scripts/compute_controlled_growth.py
PYTHONPATH=src python scripts/exact_exposure_power.py
PYTHONPATH=src python scripts/build_strong_betting_artifacts.py
PYTHONPATH=src python scripts/build_strong_betting_artifacts.py \
  --language zh
PYTHONPATH=src python scripts/build_paired_utility_appendix.py
PYTHONPATH=src python -m unittest discover -s tests -v
```

The archived protocol, source hashes, prediction and draw hashes, individual selections, candidate p-values, plotting inputs, split summaries, and paired differences allow numerical checks without model retraining.

E.2 CONSTANT-STAKE GROWTH IN THE CONTROLLED POPULATIONS

The controlled populations permit a direct calculation of the growth criterion in equation 15. In the sparse population, $e \sim \text{Bernoulli}(\mu)$ and the loss conditional on $e = 1$ is $\text{Bernoulli}(r)$, with $L = 0$ when $e = 0$. In the constant population, $e \equiv \mu$ and $L \sim \text{Bernoulli}(r)$. Retain the same universally valid stake interval $0 \le \lambda < 1/(1-\alpha)$ as in the definition of $G(P)$. Writing

$$g_r(t) = (1-r)\log(1+t\alpha) + r\log(1-t(1-\alpha)),$$

the two expected log factors are $H_{\mathrm{s}}(\lambda) = \mu g_r(\lambda)$ and $H_{\mathrm{c}}(\lambda) = g_r(\mu\lambda)$. For $0 \le t < 1/(1-\alpha)$,

$$g_r'(t) = \frac{\alpha - r - t\alpha(1-\alpha)}{(1+t\alpha)(1-t(1-\alpha))}, \qquad g_r''(t) < 0.$$

Table 7: Expected log-wealth growth per original observation (natural logarithms). A dagger marks an endpoint supremum approached from below. Stakes are rounded to six decimals and growth rates to nine decimals. These are deterministic population calculations for the three simulated settings, rather than fitted estimates.

(μ, r, α)	Population	λ^*	$G(P)$	$\bar{\lambda}$	$H_P(\bar{\lambda})$
$(1, 0, .1)$	Sparse	$1.111111^\dagger$	0.105360516	1.111110	0.105360416
$(1, 0, .1)$	Constant	$1.111111^\dagger$	0.105360516	1.111110	0.105360416
$(.5, .05, .3)$	Sparse	1.190476	0.100262297	1.190476	0.100262297
$(.5, .05, .3)$	Constant	$1.428571^\dagger$	0.149790855	1.190476	0.129138074
$(.25, .1, .5)$	Sparse	1.600000	0.092016052	1.600000	0.092016052
$(.25, .1, .5)$	Constant	$2.000000^\dagger$	0.172060989	1.600000	0.141775046

The derivative changes sign at $(\alpha - r)/[\alpha(1-\alpha)]$ when $r > 0$; for $r = 0$ the function is increasing. Thus the optimal stakes, allowing a right-endpoint limit, are

$$\lambda_{\mathrm{s}}^* = \frac{\alpha - r}{\alpha(1-\alpha)}, \qquad \lambda_{\mathrm{c}}^* = \min\left\{\frac{\alpha - r}{\mu\alpha(1-\alpha)}, \frac{1}{1-\alpha}\right\}. \tag{59}$$

Consequently $G(P_{\mathrm{s}}) = \mu\,\mathrm{kl}(r, \alpha)$ and $G(P_{\mathrm{c}}) = g_r(\mu\lambda_{\mathrm{c}}^*)$. The constant-population optimum is an endpoint supremum precisely when $r \le \alpha(1-\mu)$. These values use the original universal stake constraint; exploiting a separately known smaller exposure bound would change that constraint.

Table 7 compares these optima with the limiting stake of the predictable risk plug-in in equation 34, $\bar{\lambda} = \min\{(1-\varepsilon)/(1-\alpha), (\alpha - r)/[\alpha(1-\alpha)]\}$, where $\varepsilon = 10^{-6}$. The consistency proof identifies its almost-sure log-wealth rate with $H_P(\bar{\lambda})$. In the two nontrivial constant-exposure cases, the plug-in leaves a positive growth gap because its limiting stake is smaller than the optimum. In the first case the two populations coincide, and only the endpoint clipping produces the small gap. Bounded scalar optimization, with the endpoint limits checked separately, agrees with the analytic growth values to within 2.8×10^{-17}. This calculation illustrates the existing growth criterion on the experimental populations. Log growth is an asymptotic wealth criterion; finite-sample power also depends on fluctuations, mixture penalties, calibration, and count guards, so the table does not imply a finite-budget power ordering.

E.3 Exact budgets and simulated power

The growth calculation describes asymptotic evidence accumulation. We next return to finite samples, comparing exact minimax power with the implemented test and then examining the two controlled exposure structures. Figure 3 separates the exact minimax envelope from the implemented conditional test across budgets. The latter can show local power decreases because its nonrandomized rejection counts change discretely. We scan budgets in increasing order when finding its first crossing of the target power. The minimax envelope cannot decrease: a test at a larger budget can ignore its final observation.

Figures 4 and 5 supplement Figure 1. Unit exposure in the first case makes the two population constructions identical. Smaller mean exposure in the third case increases the contrast between their second moments.

E.4 All risk caps and calibration budgets

The main text highlights representative caps and budgets. Here we broaden the view to the full experimental grid, keeping the candidate order fixed so that the curves reflect changes in the tests and available evidence. Table 8 reports the full comparison at $n = 500$. Figures 6 and 7 show the budget curves and the five underlying splits. All methods use the original development exposure-score order. The light curves make variation across splits visible; their overlap does not imply independent sampling uncertainty. The individual records also retain each development order, the Bonferroni search, and both wider grids.

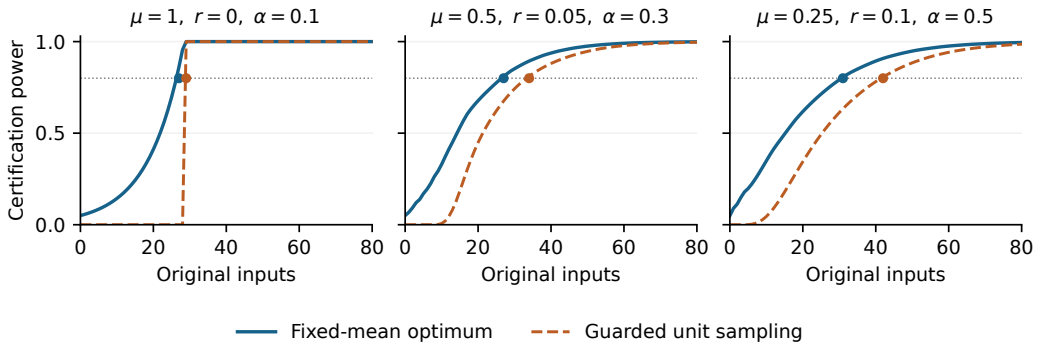

Figure 3: Exact randomized fixed-mean power and the implemented guarded conditional-binomial test in three settings. The three-point count law provides the minimax benchmark for the entire bounded class at fixed mean exposure and design alternative. Markers identify the first budget with power at least .8.

Both direct betting methods improve on Unit in all five Scene splits and in four Yeast splits; both worsen Yeast split 7. An average utility improvement therefore does not imply a uniform split-level improvement.

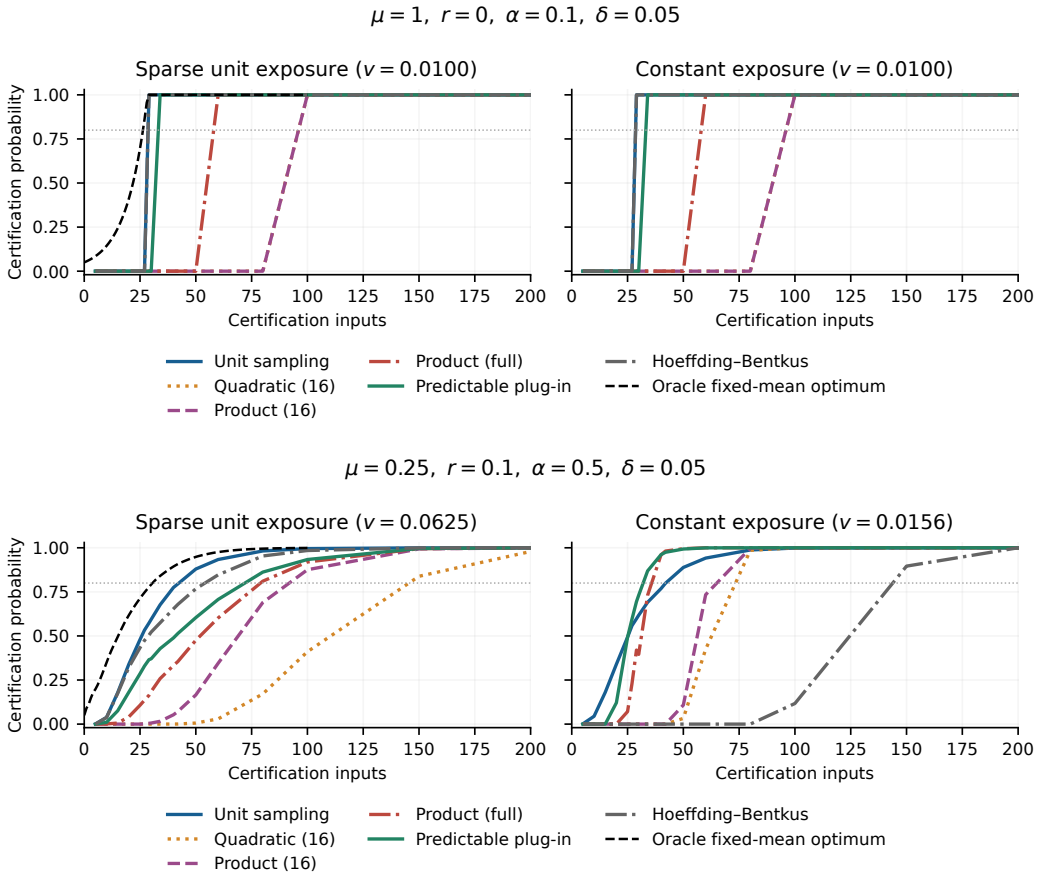

Figure 4: Additional known-population power curves. Each frequency uses 2,000 independent runs. The randomized fixed-mean envelope applies to the sparse-unit panels; direct tests can exploit the additional structure in constant-exposure panels.

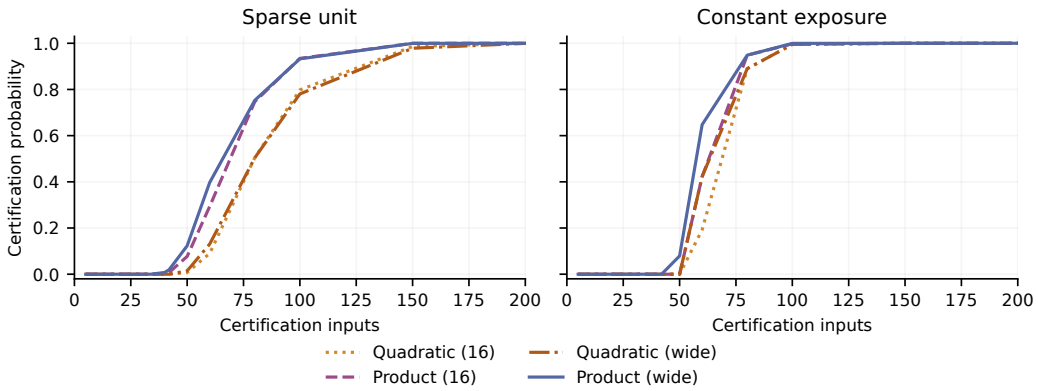

Figure 5: Same-grid product and quadratic mixtures with the horizon-dependent stake grid. This grid extends toward zero as the budget grows. It differs from the full-range product control, which extends toward the largest valid product stake.

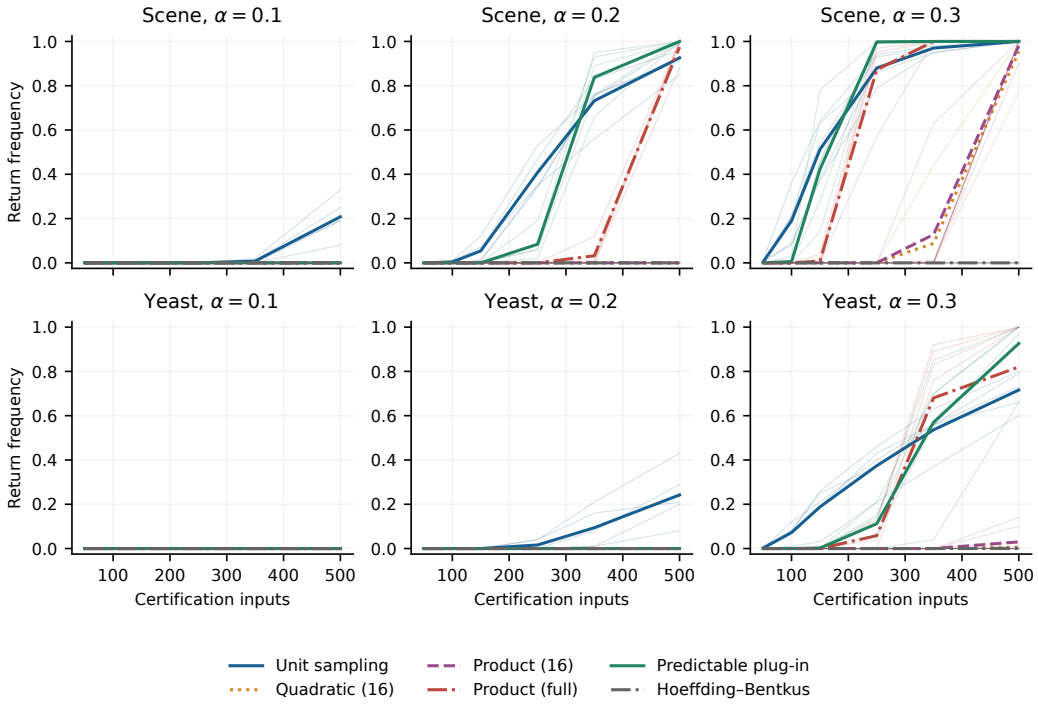

Figure 6: Return frequencies across calibration budgets at three risk caps. Light curves show each of five frozen splits; dark curves show their mean. Each split uses 100 shared permutations of its reserved finite pool.

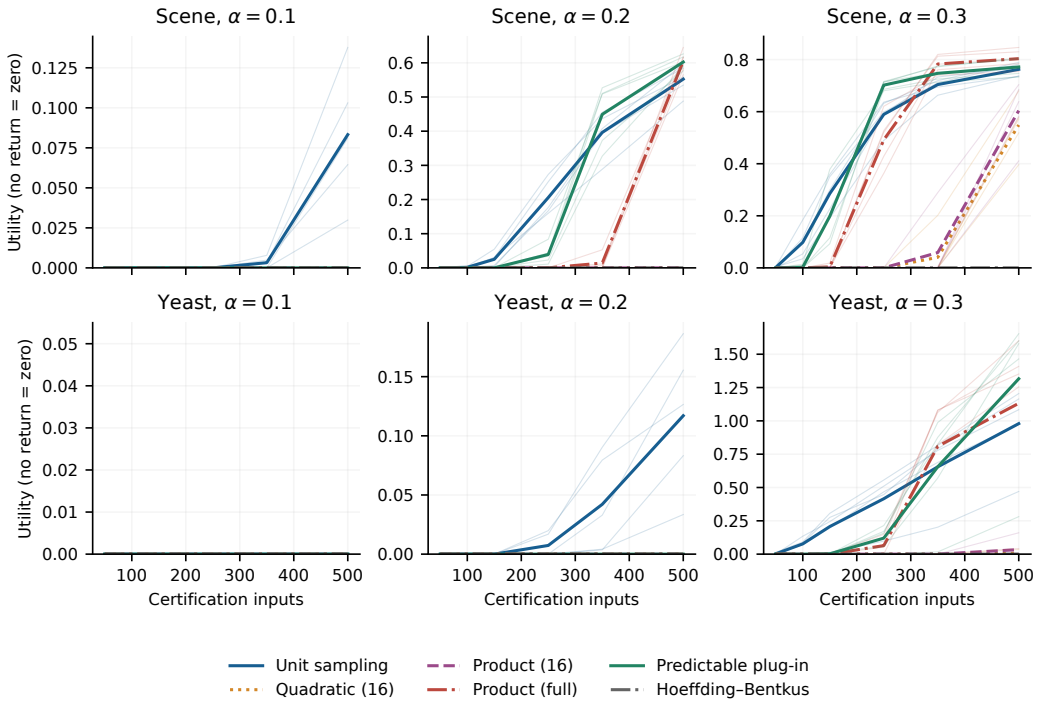

Figure 7: Utility per input across all runs, with zero utility for an empty return. The higher return frequency of a test need not imply higher utility, because tests can certify different sets of candidates.

Table 8: Frozen-cache replay at $n = 500$, using the original development evidence order. U_0 averages utility over all runs with zero for an empty return; U_{ret} averages only returned runs. Ex./ret. counts held-out empirical risk exceedances among returns, not population type-I errors. Each row pools 100 reused-pool draws from each of five splits; split-paired results are reported separately.

Data / α	Test	Returns	U_0	U_{ret}	Ex./ret.
Scene / 0.1	Adaptive	0/500	0.000	–	N/A (0/0)
Scene / 0.1	Quad16	0/500	0.000	–	N/A (0/0)
Scene / 0.1	HB	0/500	0.000	–	N/A (0/0)
Scene / 0.1	Prod16	0/500	0.000	–	N/A (0/0)
Scene / 0.1	ProdFull	0/500	0.000	–	N/A (0/0)
Scene / 0.1	Unit	104/500	0.083	0.400	0/104
Scene / 0.2	Adaptive	500/500	0.602	0.602	0/500
Scene / 0.2	Quad16	0/500	0.000	–	N/A (0/0)
Scene / 0.2	HB	0/500	0.000	–	N/A (0/0)
Scene / 0.2	Prod16	0/500	0.000	–	N/A (0/0)
Scene / 0.2	ProdFull	487/500	0.605	0.621	0/487
Scene / 0.2	Unit	463/500	0.552	0.596	0/463
Scene / 0.3	Adaptive	500/500	0.772	0.772	0/500
Scene / 0.3	Quad16	477/500	0.549	0.575	0/477
Scene / 0.3	HB	0/500	0.000	–	N/A (0/0)
Scene / 0.3	Prod16	490/500	0.604	0.616	0/490
Scene / 0.3	ProdFull	500/500	0.804	0.804	0/500
Scene / 0.3	Unit	500/500	0.763	0.763	2/500
Yeast / 0.1	Adaptive	0/500	0.000	–	N/A (0/0)
Yeast / 0.1	Quad16	0/500	0.000	–	N/A (0/0)
Yeast / 0.1	HB	0/500	0.000	–	N/A (0/0)
Yeast / 0.1	Prod16	0/500	0.000	–	N/A (0/0)
Yeast / 0.1	ProdFull	0/500	0.000	–	N/A (0/0)
Yeast / 0.1	Unit	0/500	0.000	–	N/A (0/0)
Yeast / 0.2	Adaptive	0/500	0.000	–	N/A (0/0)
Yeast / 0.2	Quad16	0/500	0.000	–	N/A (0/0)
Yeast / 0.2	HB	0/500	0.000	–	N/A (0/0)
Yeast / 0.2	Prod16	0/500	0.000	–	N/A (0/0)
Yeast / 0.2	ProdFull	0/500	0.000	–	N/A (0/0)
Yeast / 0.2	Unit	121/500	0.117	0.483	2/121
Yeast / 0.3	Adaptive	463/500	1.316	1.421	0/463
Yeast / 0.3	Quad16	3/500	0.007	1.144	0/3
Yeast / 0.3	HB	0/500	0.000	–	N/A (0/0)
Yeast / 0.3	Prod16	15/500	0.034	1.132	0/15
Yeast / 0.3	ProdFull	410/500	1.131	1.380	0/410
Yeast / 0.3	Unit	358/500	0.979	1.368	1/358

Table 9: Within-split paired unconditional utility at $n = 500$, with the original development evidence order. Unit sampling is the reference; Adaptive and Full denote predictable risk plug-in and full-range fixed product betting. Each utility averages all 100 runs in that split, with zero for an empty return. Differences are new method minus Unit. Shared permutations pair the comparisons; overlapping data splits support descriptive differences, not independent-sample significance claims.

Data / α	Split	Unit U_0	Adaptive U_0	Difference	Full U_0	Difference
Scene / 0.2	7	0.5585	0.6028	+0.0442	0.6164	+0.0578
Scene / 0.2	19	0.5960	0.6261	+0.0301	0.6450	+0.0490
Scene / 0.2	42	0.5347	0.5892	+0.0545	0.5650	+0.0303
Scene / 0.2	73	0.5844	0.6173	+0.0329	0.6157	+0.0313
Scene / 0.2	101	0.4880	0.5766	+0.0886	0.5820	+0.0940
Yeast / 0.3	7	0.4702	0.2813	-0.1889	0.0418	-0.4283
Yeast / 0.3	19	0.9732	1.5792	+0.6059	1.2547	+0.2815
Yeast / 0.3	42	1.1643	1.6539	+0.4896	1.4086	+0.2443
Yeast / 0.3	73	1.0840	1.5997	+0.5157	1.6008	+0.5169
Yeast / 0.3	101	1.2049	1.4647	+0.2598	1.3515	+0.1466

Table 10: Fixed-policy diagnostic at $n = 500$. Each policy is selected using development data alone: the first frontier candidate under the risk, evidence, or HB order. Moments are equal-weight means of held-out estimates across five splits; pass frequencies are also equal-weight split means (100 reused-pool draws per split). The policy may differ across splits; the risk estimates are not population ground truth.

Data / α	Dev. rule	$\hat{\mu}$	$\hat{r}$	$\hat{v}$	Unit	Quad16	Prod16	ProdFull	Adaptive	HB
Scene / 0.1	evidence	0.054	0.015	0.00020	0.208	0.000	0.000	0.000	0.000	0.000
Scene / 0.1	hb	0.054	0.015	0.00020	0.208	0.000	0.000	0.000	0.000	0.000
Scene / 0.1	risk	0.054	0.015	0.00020	0.208	0.000	0.000	0.000	0.000	0.000
Scene / 0.2	evidence	0.061	0.021	0.00054	0.926	0.000	0.000	0.974	1.000	0.000
Scene / 0.2	hb	0.061	0.021	0.00054	0.926	0.000	0.000	0.974	1.000	0.000
Scene / 0.2	risk	0.054	0.015	0.00044	0.948	0.000	0.000	0.940	1.000	0.000
Scene / 0.3	evidence	0.070	0.026	0.00118	1.000	0.954	0.980	1.000	1.000	0.000
Scene / 0.3	hb	0.088	0.048	0.00161	0.994	1.000	1.000	1.000	1.000	0.000
Scene / 0.3	risk	0.054	0.015	0.00086	1.000	0.454	0.570	1.000	1.000	0.000
Yeast / 0.1	evidence	0.018	0.030	0.00009	0.000	0.000	0.000	0.000	0.000	0.000
Yeast / 0.1	hb	0.028	0.065	0.00019	0.000	0.000	0.000	0.000	0.000	0.000
Yeast / 0.1	risk	0.005	0.000	0.00001	0.000	0.000	0.000	0.000	0.000	0.000
Yeast / 0.2	evidence	0.031	0.068	0.00039	0.242	0.000	0.000	0.000	0.000	0.000
Yeast / 0.2	hb	0.033	0.073	0.00044	0.266	0.000	0.000	0.000	0.000	0.000
Yeast / 0.2	risk	0.005	0.000	0.00002	0.000	0.000	0.000	0.000	0.000	0.000
Yeast / 0.3	evidence	0.071	0.153	0.00168	0.716	0.006	0.030	0.820	0.926	0.000
Yeast / 0.3	hb	0.058	0.138	0.00145	0.740	0.002	0.034	0.974	0.990	0.000
Yeast / 0.3	risk	0.005	0.000	0.00005	0.004	0.000	0.000	0.000	0.000	0.000

E.5 FIXED-POLICY DIAGNOSTICS

Aggregate selection results can hide which candidate drives a difference. We therefore freeze one policy per development rule and compare the tests on that same policy before returning to the full search results. Table 10 averages each fixed-policy diagnostic over five splits. The following tables retain the split-specific policies and moments so that these averages do not hide which policy was tested. Each development rule chooses a policy before certification. Held-out moment estimates describe the population seen by that policy; they are not used to select stakes or choose candidates.

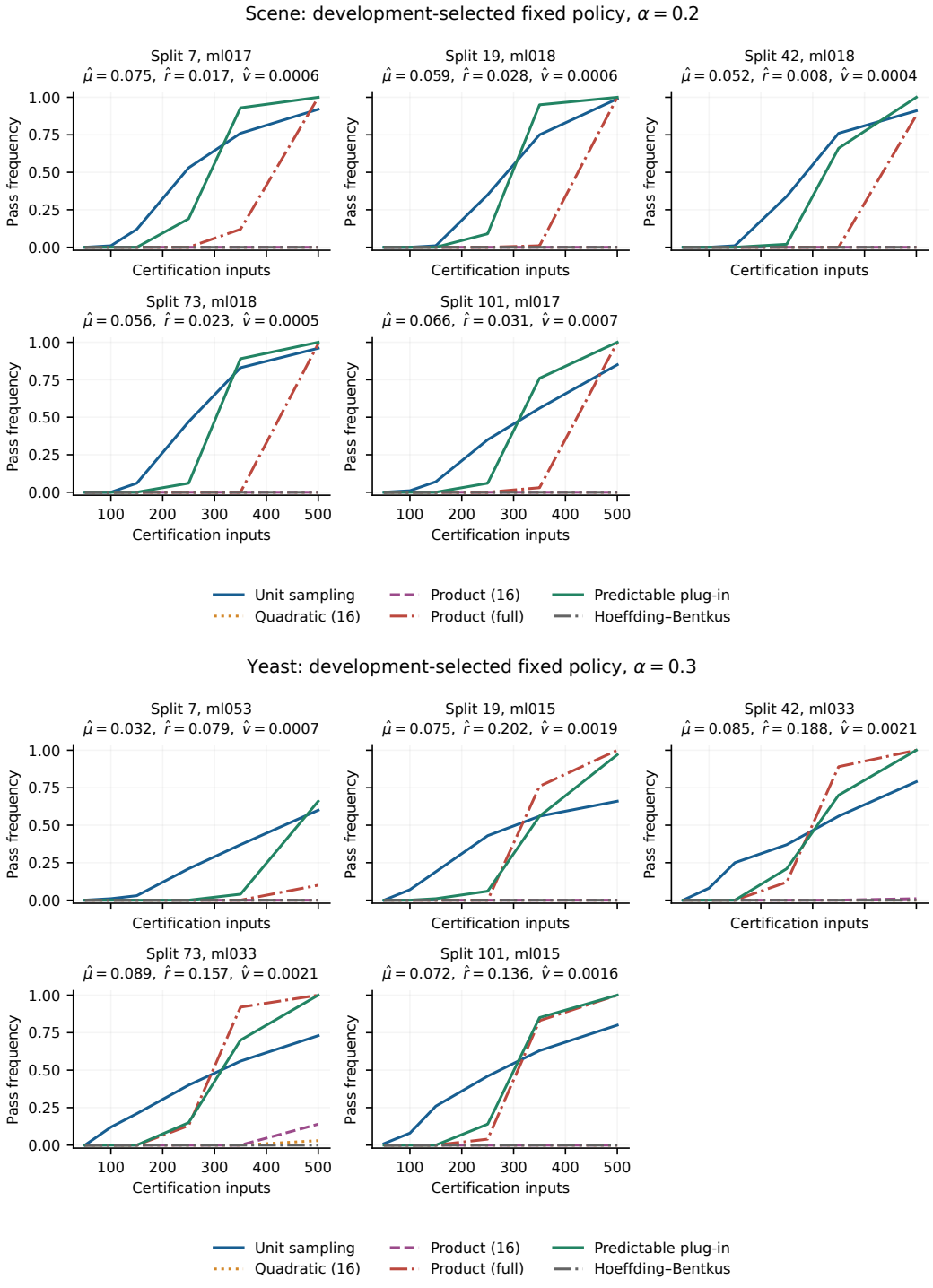

Figure 8: Pass frequencies for the policy chosen before certification by the exposure-score rule. Top: Scene at $\alpha = .2$. Bottom: Yeast at $\alpha = .3$. Each panel shows one split; curves compare tests, and panel titles give held-out estimates of μ, r, and v.

Table 11: Split-specific fixed-policy diagnostics for Scene at $\alpha = 0.1$, with $n = 500$. The development rule chooses each policy before certification. Moments use held-out evaluation data; entries in test columns are pass frequencies over 100 cached draws.

Dev. rule	Split / policy	$\hat{\mu}$	$\hat{r}$	$\hat{v}$	Unit	Quad16	Prod16	ProdFull	Adaptive	HB
evidence	7 / ml018	0.054	0.015	0.00020	0.190	0.000	0.000	0.000	0.000	0.000
evidence	19 / ml018	0.059	0.028	0.00032	0.250	0.000	0.000	0.000	0.000	0.000
evidence	42 / ml018	0.052	0.008	0.00014	0.190	0.000	0.000	0.000	0.000	0.000
evidence	73 / ml018	0.056	0.023	0.00026	0.330	0.000	0.000	0.000	0.000	0.000
evidence	101 / ml018	0.048	0.000	0.00008	0.080	0.000	0.000	0.000	0.000	0.000
hb	7 / ml018	0.054	0.015	0.00020	0.190	0.000	0.000	0.000	0.000	0.000
hb	19 / ml018	0.059	0.028	0.00032	0.250	0.000	0.000	0.000	0.000	0.000
hb	42 / ml018	0.052	0.008	0.00014	0.190	0.000	0.000	0.000	0.000	0.000
hb	73 / ml018	0.056	0.023	0.00026	0.330	0.000	0.000	0.000	0.000	0.000
hb	101 / ml018	0.048	0.000	0.00008	0.080	0.000	0.000	0.000	0.000	0.000
risk	7 / ml018	0.054	0.015	0.00020	0.190	0.000	0.000	0.000	0.000	0.000
risk	19 / ml018	0.059	0.028	0.00032	0.250	0.000	0.000	0.000	0.000	0.000
risk	42 / ml018	0.052	0.008	0.00014	0.190	0.000	0.000	0.000	0.000	0.000
risk	73 / ml018	0.056	0.023	0.00026	0.330	0.000	0.000	0.000	0.000	0.000
risk	101 / ml018	0.048	0.000	0.00008	0.080	0.000	0.000	0.000	0.000	0.000

Table 12: Split-specific fixed-policy diagnostics for Scene at $\alpha = 0.2$, with $n = 500$. The development rule chooses each policy before certification. Moments use held-out evaluation data; entries in test columns are pass frequencies over 100 cached draws.

Dev. rule	Split / policy	$\hat{\mu}$	$\hat{r}$	$\hat{v}$	Unit	Quad16	Prod16	ProdFull	Adaptive	HB
evidence	7 / ml017	0.075	0.017	0.00063	0.920	0.000	0.000	1.000	1.000	0.000
evidence	19 / ml018	0.059	0.028	0.00056	0.990	0.000	0.000	1.000	1.000	0.000
evidence	42 / ml018	0.052	0.008	0.00039	0.910	0.000	0.000	0.880	1.000	0.000
evidence	73 / ml018	0.056	0.023	0.00050	0.960	0.000	0.000	0.990	1.000	0.000
evidence	101 / ml017	0.066	0.031	0.00065	0.850	0.000	0.000	1.000	1.000	0.000
hb	7 / ml017	0.075	0.017	0.00063	0.920	0.000	0.000	1.000	1.000	0.000
hb	19 / ml018	0.059	0.028	0.00056	0.990	0.000	0.000	1.000	1.000	0.000
hb	42 / ml018	0.052	0.008	0.00039	0.910	0.000	0.000	0.880	1.000	0.000
hb	73 / ml018	0.056	0.023	0.00050	0.960	0.000	0.000	0.990	1.000	0.000
hb	101 / ml017	0.066	0.031	0.00065	0.850	0.000	0.000	1.000	1.000	0.000
risk	7 / ml018	0.054	0.015	0.00045	0.950	0.000	0.000	0.970	1.000	0.000
risk	19 / ml018	0.059	0.028	0.00056	0.990	0.000	0.000	1.000	1.000	0.000
risk	42 / ml018	0.052	0.008	0.00039	0.910	0.000	0.000	0.880	1.000	0.000
risk	73 / ml018	0.056	0.023	0.00050	0.960	0.000	0.000	0.990	1.000	0.000
risk	101 / ml018	0.048	0.000	0.00032	0.930	0.000	0.000	0.860	1.000	0.000

Table 13: Split-specific fixed-policy diagnostics for Scene at $\alpha = 0.3$, with $n = 500$. The development rule chooses each policy before certification. Moments use held-out evaluation data; entries in test columns are pass frequencies over 100 cached draws.

Dev. rule	Split / policy	$\hat{\mu}$	$\hat{r}$	$\hat{v}$	Unit	Quad16	Prod16	ProdFull	Adaptive	HB
evidence	7 / ml017	0.075	0.017	0.00122	1.000	1.000	1.000	1.000	1.000	0.000
evidence	19 / ml018	0.059	0.028	0.00099	1.000	0.770	0.900	1.000	1.000	0.000
evidence	42 / ml017	0.071	0.012	0.00113	1.000	1.000	1.000	1.000	1.000	0.000
evidence	73 / ml016	0.080	0.042	0.00142	1.000	1.000	1.000	1.000	1.000	0.000
evidence	101 / ml017	0.066	0.031	0.00114	1.000	1.000	1.000	1.000	1.000	0.000
hb	7 / ml016	0.086	0.034	0.00152	1.000	1.000	1.000	1.000	1.000	0.000
hb	19 / ml015	0.096	0.061	0.00187	1.000	1.000	1.000	1.000	1.000	0.000
hb	42 / ml013	0.110	0.072	0.00212	0.970	1.000	1.000	1.000	1.000	0.000
hb	73 / ml016	0.080	0.042	0.00142	1.000	1.000	1.000	1.000	1.000	0.000
hb	101 / ml017	0.066	0.031	0.00114	1.000	1.000	1.000	1.000	1.000	0.000
risk	7 / ml018	0.054	0.015	0.00087	1.000	0.380	0.560	1.000	1.000	0.000
risk	19 / ml018	0.059	0.028	0.00099	1.000	0.770	0.900	1.000	1.000	0.000
risk	42 / ml018	0.052	0.008	0.00080	1.000	0.260	0.420	1.000	1.000	0.000
risk	73 / ml018	0.056	0.023	0.00092	1.000	0.830	0.900	1.000	1.000	0.000
risk	101 / ml018	0.048	0.000	0.00072	1.000	0.030	0.070	1.000	1.000	0.000

Table 14: Split-specific fixed-policy diagnostics for Yeast at $\alpha = 0.1$, with $n = 500$. The development rule chooses each policy before certification. Moments use held-out evaluation data; entries in test columns are pass frequencies over 100 cached draws.

Dev. rule	Split / policy	$\hat{\mu}$	$\hat{r}$	$\hat{v}$	Unit	Quad16	Prod16	ProdFull	Adaptive	HB
evidence	7 / ml054	0.027	0.060	0.00015	0.000	0.000	0.000	0.000	0.000	0.000
evidence	19 / ml054	0.018	0.040	0.00007	0.000	0.000	0.000	0.000	0.000	0.000
evidence	42 / ml018	0.007	0.000	0.00001	0.000	0.000	0.000	0.000	0.000	0.000
evidence	73 / ml018	0.006	0.000	0.00001	0.000	0.000	0.000	0.000	0.000	0.000
evidence	101 / ml053	0.031	0.053	0.00022	0.000	0.000	0.000	0.000	0.000	0.000
hb	7 / ml054	0.027	0.060	0.00015	0.000	0.000	0.000	0.000	0.000	0.000
hb	19 / ml053	0.027	0.035	0.00011	0.000	0.000	0.000	0.000	0.000	0.000
hb	42 / ml051	0.033	0.085	0.00030	0.000	0.000	0.000	0.000	0.000	0.000
hb	73 / ml054	0.020	0.093	0.00016	0.000	0.000	0.000	0.000	0.000	0.000
hb	101 / ml053	0.031	0.053	0.00022	0.000	0.000	0.000	0.000	0.000	0.000
risk	7 / ml018	0.005	0.000	0.00000	0.000	0.000	0.000	0.000	0.000	0.000
risk	19 / ml018	0.004	0.000	0.00000	0.000	0.000	0.000	0.000	0.000	0.000
risk	42 / ml018	0.007	0.000	0.00001	0.000	0.000	0.000	0.000	0.000	0.000
risk	73 / ml018	0.006	0.000	0.00001	0.000	0.000	0.000	0.000	0.000	0.000
risk	101 / ml018	0.004	0.000	0.00000	0.000	0.000	0.000	0.000	0.000	0.000

Table 15: Split-specific fixed-policy diagnostics for Yeast at $\alpha = 0.2$, with $n = 500$. The development rule chooses each policy before certification. Moments use held-out evaluation data; entries in test columns are pass frequencies over 100 cached draws.

Dev. rule	Split / policy	$\hat{\mu}$	$\hat{r}$	$\hat{v}$	Unit	Quad16	Prod16	ProdFull	Adaptive	HB
evidence	7 / ml054	0.027	0.060	0.00028	0.080	0.000	0.000	0.000	0.000	0.000
evidence	19 / ml053	0.027	0.035	0.00029	0.200	0.000	0.000	0.000	0.000	0.000
evidence	42 / ml051	0.033	0.085	0.00041	0.430	0.000	0.000	0.000	0.000	0.000
evidence	73 / ml051	0.037	0.106	0.00059	0.210	0.000	0.000	0.000	0.000	0.000
evidence	101 / ml053	0.031	0.053	0.00038	0.290	0.000	0.000	0.000	0.000	0.000
hb	7 / ml053	0.032	0.079	0.00039	0.210	0.000	0.000	0.000	0.000	0.000
hb	19 / ml053	0.027	0.035	0.00029	0.200	0.000	0.000	0.000	0.000	0.000
hb	42 / ml051	0.033	0.085	0.00041	0.430	0.000	0.000	0.000	0.000	0.000
hb	73 / ml050	0.041	0.114	0.00071	0.200	0.000	0.000	0.000	0.000	0.000
hb	101 / ml053	0.031	0.053	0.00038	0.290	0.000	0.000	0.000	0.000	0.000
risk	7 / ml018	0.005	0.000	0.00002	0.000	0.000	0.000	0.000	0.000	0.000
risk	19 / ml018	0.004	0.000	0.00001	0.000	0.000	0.000	0.000	0.000	0.000
risk	42 / ml018	0.007	0.000	0.00004	0.000	0.000	0.000	0.000	0.000	0.000
risk	73 / ml018	0.006	0.000	0.00002	0.000	0.000	0.000	0.000	0.000	0.000
risk	101 / ml018	0.004	0.000	0.00002	0.000	0.000	0.000	0.000	0.000	0.000

Table 16: Split-specific fixed-policy diagnostics for Yeast at $\alpha = 0.3$, with $n = 500$. The development rule chooses each policy before certification. Moments use held-out evaluation data; entries in test columns are pass frequencies over 100 cached draws.

Dev. rule	Split / policy	$\hat{\mu}$	$\hat{r}$	$\hat{v}$	Unit	Quad16	Prod16	ProdFull	Adaptive	HB
evidence	7 / ml053	0.032	0.079	0.00068	0.600	0.000	0.000	0.100	0.660	0.000
evidence	19 / ml015	0.075	0.202	0.00188	0.660	0.000	0.000	1.000	0.970	0.000
evidence	42 / ml033	0.085	0.188	0.00213	0.790	0.000	0.010	1.000	1.000	0.000
evidence	73 / ml033	0.089	0.157	0.00211	0.730	0.030	0.140	1.000	1.000	0.000
evidence	101 / ml015	0.072	0.136	0.00163	0.800	0.000	0.000	1.000	1.000	0.000
hb	7 / ml049	0.042	0.095	0.00094	0.770	0.000	0.000	0.940	0.980	0.000
hb	19 / ml015	0.075	0.202	0.00188	0.660	0.000	0.000	1.000	0.970	0.000
hb	42 / ml050	0.035	0.093	0.00077	0.770	0.000	0.000	0.930	1.000	0.000
hb	73 / ml050	0.041	0.114	0.00101	0.740	0.000	0.000	1.000	1.000	0.000
hb	101 / ml031	0.098	0.186	0.00264	0.760	0.010	0.170	1.000	1.000	0.000
risk	7 / ml018	0.005	0.000	0.00004	0.000	0.000	0.000	0.000	0.000	0.000
risk	19 / ml018	0.004	0.000	0.00003	0.000	0.000	0.000	0.000	0.000	0.000
risk	42 / ml018	0.007	0.000	0.00008	0.020	0.000	0.000	0.000	0.000	0.000
risk	73 / ml018	0.006	0.000	0.00006	0.000	0.000	0.000	0.000	0.000	0.000
risk	101 / ml018	0.004	0.000	0.00004	0.000	0.000	0.000	0.000	0.000	0.000

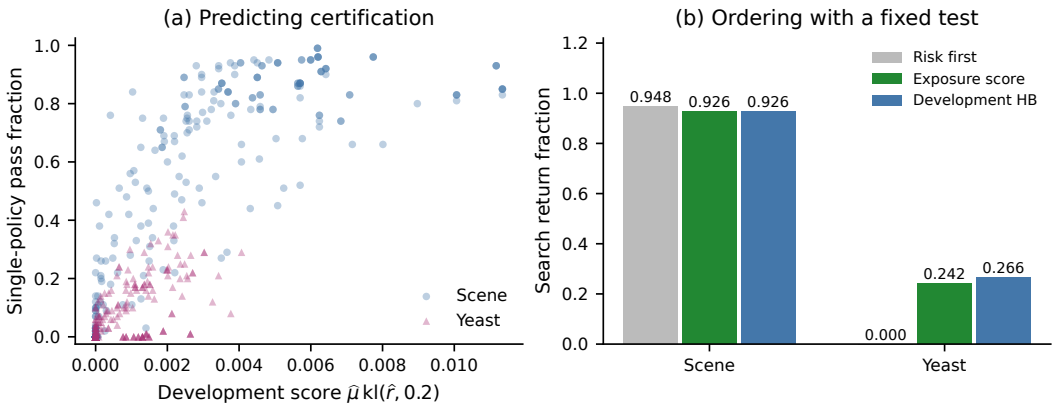

Figure 9: Development-only difficulty prediction and search, nonlinear models with $n = 500, \alpha = .2$. (a) Each point is a fixed candidate within a model split; its y-coordinate averages 100 repeated calibration subsets. Candidates and repetitions are dependent. (b) The certification test is held fixed to exposure-unit sampling. The exposure score improves Yeast's return fraction over risk-first ordering but not over development HB, and slightly lowers Scene's return fraction relative to risk-first.

Table 17: The guarantee being optimized matters. Exact finite-sample power, expected log growth, and expected sequential rejection time are distinct objects. The exposure reduction specializes an established binary-reduction principle.

Work	Guarantee object	Model and time regime	Role of distributional structure
Schlag (2007)	Minimax risk; worst power at fixed means	Independent bounded-vector samples; finite sample	Mean-preserving binary transformation must preserve the model class
Waudby-Smith & Ramdas (2024)	Valid confidence sequences and intervals	Bounded observations; predictable stakes; time-uniform or fixed time	Direct product wealth adapts to observed variation
Arnold & Clerico (2026)	Worst expected log growth, absolute or relative	Bounded scalar mean; single-round e-values	Extremal distributions depend on the objective
Waudby-Smith et al. (2026)	Instance log growth; expected rejection time	i.i.d. portfolios; regret and moment assumptions; asymptotic optimality	Oracle growth depends on the alternative distribution
Xu et al. (2026), calibration-only	High-probability selected conditional risk	i.i.d. calibration; finite grids; finite sample	Separate numerator and selection-probability bounds
Exposure specialization	Exact worst-case power at fixed (μ, r)	i.i.d. (e, L); randomized tests; each finite n	Three-point least-favorable family; direct growth may be higher elsewhere

E.6 Auxiliary development-order diagnostics

The fixed-policy results separate the tests; the remaining question is how well development data anticipate certification difficulty. We address that question by comparing development rankings with independent pass frequencies. Figure 9 compares development information with independent certification outcomes. For unit sampling at $n = 500, \alpha = .2$, the mean within-split Spearman correlations between the exposure score and single-policy pass frequency are .974 (Scene) and .716 (Yeast). This rank correlation measures agreement between two orderings. Risk-only scores give .968 and .638; development HB gives .970 and .814. The score therefore provides a useful complexity diagnostic, while sample-level test information can be more predictive. At the tighter Yeast cap .1, the mean correlation is $-.049$ over four nonconstant splits, with three negative values and one positive value (Table 6).

The effect of candidate order depends on the dataset. With the sampling test fixed, Scene's return rate changes from .948 under risk-first ordering to .926 under the exposure score. Yeast changes from zero to .242, but development HB reaches .266. At $\alpha = .3$, the corresponding Yeast rates are .004, .716, and .740. Higher return rate need not mean higher output: at that cap, a betting-derived development order returns less often (.568) but produces more correct labels when missing returns count as zero (1.110/input versus .979 for the exposure-score order). The complete cross-order results show that the score's benefit depends on the dataset, risk cap, and comparison order.

F THE LENDING CONTRACT AS AN EXPOSURE APPLICATION

The general exposure framework becomes a joint decision problem in lending: a policy must control default risk, principal-weighted exposure, and review workload at the same time. This section defines that contract and shows how the single-policy tests combine into a valid post-selection certificate.

In this application, $Y \in \{0, 1\}$ denotes default and a policy approves, rejects, or defers an application. Let A_π, D_π be its approval and deferral events and $w(X) \in [0, 1]$ the normalized requested principal. The general exposure is $e_\pi = w\mathbf{1}_{A_\pi}$, with unit loss $L_\pi = Y$ on approval and zero otherwise. The three constraints are

$$R_A(\pi) = \Pr(Y = 1 \mid A_\pi), \quad R_D(\pi) = \Pr(D_\pi), \quad R_W(\pi) = \frac{\mathbb{E}[wY\mathbf{1}_{A_\pi}]}{\mathbb{E}[w\mathbf{1}_{A_\pi}]}. \tag{60}$$

The contract requires $R_A \le \alpha_A$, $R_W \le \alpha_W$, and $R_D \le B$. The count-only contract omits R_W. Principal associated with defaulting loans measures exposure to default, while cash-flow loss also depends on recovery and timing. We define the target using a known principal cap; principal clipping would define a different target. Approval probability and population approved exposure must both be positive.

For approval count N_A, default count K_A, and deferral count K_D, conditional sampling gives

$$p_A = F_{N_A,\alpha_A}(K_A), \qquad p_D = F_{n,B}(K_D). \tag{61}$$

Approval-dependent tests return a p-value of one when $N_A < 5$. On the approved sample, set $Z_i = w_i(\alpha_W - Y_i)$ and

$$S = \sum_{i:A_i} Z_i, \qquad Q = \sum_{i:A_i} Z_i^2. \tag{62}$$

The lending runs use the 16-stake grid from $1/[64(1 - \alpha_W)]$ to $1/[2(1 - \alpha_W)]$ in Equation 13. Thinning retains an approved observation with probability w_i and tests its binary default outcome. The Hoeffding and HB comparators use the same approved sample. The comparisons use the same fixed candidate parameters and evaluate each fixed order with the different certification tests.

Proposition 10 (Complete-contract certification)**.** *Under Assumption 1, combine a valid exposure p-value with Equation 61 using $p_\pi = \max(p_A, p_D, p_W)$. Bonferroni-IUT, Holm-IUT, or an independently ordered fixed sequence gives*

$$\Pr\{\exists\pi \in \widehat{\Pi}_{\text{cert}} : R_A(\pi) > \alpha_A \text{ or } R_W(\pi) > \alpha_W \text{ or } R_D(\pi) > B \mid \mathcal{F}_0\} \le \delta. \tag{63}$$

Every measurable selection within the set inherits this guarantee.

Appendix G proves the constituent statements. The risk caps and adjusted p-values specify the certificate. Appendix G.5 explains why inversion at an observed Holm rank does not give simultaneous numerical confidence bounds.

F.1 CREDIT-SPECIFIC COMPLEXITY CONSTRUCTIONS

The joint contract establishes validity, but it does not yet explain how much data the exposure component needs. We first summarize the empirical reference point and then use simple population constructions to isolate that sample-cost mechanism.

The following binary-loss results support the lending experiments and give a complementary contamination interpretation of exposure cost. The main text uses the more general bounded-loss characterization. The central difficulty exists even for one policy. Consider a fixed approval region containing a type with $w = a \in (0, 1]$ and a type with $w = 1$. If $\rho < 1$, also assume a reject type exists. In a safe population P_0, the policy approves the low-exposure type with probability $\rho \in (0, 1]$, all its labels are zero, and it rejects the remaining mass. Its approved exposure mass is $\mu = \rho a$.

Theorem 11 (Necessary evidence against high-exposure contamination)**.** *Fix the development object and policy above. A certifier uses only n fresh observations and independent internal randomness and satisfies $\Pr_P(C) \le \delta$ for every population in a class whenever the policy violates its exposure cap, conditional on this same development object. Suppose the class contains P_0 and $P_1 = (1-\varepsilon)P_0 + \varepsilon H$, where H is a unit-exposure, defaulting approved type and*

$$q \in (\alpha_W, 1), \qquad \varepsilon = \frac{q\rho a}{1 - q + q\rho a}. \tag{64}$$

Table 18: Same-quarter Lending reference, identical frozen candidates, $\alpha_A = \alpha_W = .09$, $B = .10$, and $\delta = .05$. Utility and exposure risk are held-out means among returning runs. There are five quarters and three model seeds per quarter; model seeds are not independent calendar replications. Ex. counts empirical violations of any component.

Exposure test	Selection	Return	Utility	$\widehat{R}_W$	Ex.
Hoeffding	Bonferroni-IUT	0/15	–	–	–
Hoeffding	Pareto Testing	6/15	0.0918	0.0585	0
Hoeffding–Bentkus	Bonferroni-IUT	6/15	0.0834	0.0547	0
Hoeffding–Bentkus	Pareto Testing	14/15	0.1175	0.0718	0
Quadratic betting	Bonferroni-IUT	15/15	0.1118	0.0693	0
Quadratic betting	Pareto Testing	15/15	0.1279	0.0768	0
Thinning	Bonferroni-IUT	15/15	0.1130	0.0705	0
Thinning	Pareto Testing	15/15	0.1313	0.0783	0

If $\beta \in (0,1)$, $\Pr_{P_0}(C) \geq 1-\beta$, and $0 < \delta < 1-\beta$, then

$$n \geq \frac{\log((1-\beta)/\delta)}{-\log(1-\varepsilon)}. \tag{65}$$

Proof sketch. The risks under the contaminated population are $R_W(P_1) = q$, $R_A(P_1) = qa/(1-q+qa)$, and $R_D(P_1) = 0$. With probability $(1-\varepsilon)^n$, a P_1 sample has no contamination and has exactly the P_0^n distribution, including the certifier's independent randomness. Thus $\delta \geq \Pr_{P_1}(C) \geq (1-\varepsilon)^n(1-\beta)$; rearrangement gives Equation 65. □

This construction violates the exposure constraint while keeping review demand at zero and count risk below its cap for sufficiently small a. For example, $q = .20, a = .01$ give $R_A(P_1) \simeq .00249$ despite $R_W(P_1) = .20$. The bound concerns fresh evidence for a uniformly valid certifier after the policy has been fixed on development data. A known support restriction that excludes high-exposure types changes the population class and this bound.

Theorem 12 (Attainable exposure scaling on the safe family)**.** *On P_0, let $\gamma, \beta \in (0,1)$ be the local test level and failure probability for power. Thinning passes exactly when $N_T \geq m_T$, where*

$$N_T \sim \mathrm{Bin}(n, \rho a), \quad m_T = \max\left\{5, \left\lceil \frac{\log\gamma}{\log(1-\alpha_W)} \right\rceil \right\}. \tag{66}$$

Thus $n\rho a \geq \max\{2m_T, 8\log(1/\beta)\}$ suffices for exposure-test power at least $1-\beta$.

For $0 < \alpha_W \leq 1/4$, let $\lambda_* = [2(1-\alpha_W)]^{-1}$ and $g(a) = \lambda_*\alpha_W a - \lambda_*^2\alpha_W^2 a^2$. The $J = 16$ mixture in Equation 13 passes whenever

$$N_A \geq m_E = \max\left\{5, \left\lceil \frac{\log(J/\gamma)}{g(a)} \right\rceil \right\}, \qquad g(a) \geq \frac{5}{12}\alpha_W a. \tag{67}$$

Therefore $n\rho \geq \max\{2m_E, 8\log(1/\beta)\}$ suffices for the same exposure-test power.

Proof sketch. Every retained label is zero, which gives Equation 66. The largest-λ betting summand gives the lower bound $E \geq J^{-1}\exp\{N_A g(a)\}$. Since $\lambda_*\alpha_W a \leq 1/6$, the bound on $g(a)$ follows. Both power claims use $\Pr\{N < \mathbb{E}N/2\} \leq \exp\{-\mathbb{E}N/8\}$ for a binomial N. Appendix G gives complete proofs and the additional count and capacity conditions for a joint contract. □

For $q = 2\alpha_W \leq 1/2$, fixed β, $\gamma = \delta$, and $0 < \delta \leq (1-\beta)/2$, the theorems match the exposure-component rate

$$n = \Theta\left(\frac{\log(1/\delta)}{\rho a \alpha_W}\right), \tag{68}$$

with constants allowed to depend on fixed β. Joint count/exposure certification additionally uses $\alpha_A \geq \alpha_W$ for the thinning implication and enough total observations to certify zero review demand. For a later policy, the fixed sequence must first reach that policy before the local power bound applies.

The exposure bottleneck extends to heterogeneous populations with nonzero default risk. Write $r = R_W < \alpha_W$, $\mu = \mathbb{E}[w\mathbf{1}_A] > 0$, and let kl denote binary relative entropy.

Theorem 13 (Exposure mass and risk-margin complexity)**.** *Under Assumption 1, fix a policy with $0 < \mu = \mathbb{E}[w\mathbf{1}_A] \le \rho = \Pr(A) \le 1$ and $0 \le r = R_W < \alpha_W < 1$. Use independent uniform thinning coins, independent of development information and certification observations. Then $N_T \sim \mathrm{Bin}(n, \mu)$ and $K_T \mid N_T = m \sim \mathrm{Bin}(m, r)$ for every allowed exposure law and count of positive probability. Let $\gamma, \beta \in (0, 1)$. With $s = (r + \alpha_W)/2$, define*

$$m_\star = \max\left\{5, \left\lceil \frac{\log(1/\gamma)}{\mathrm{kl}(s, \alpha_W)} \right\rceil, \left\lceil \frac{\log(2/\beta)}{\mathrm{kl}(s, r)} \right\rceil \right\}, \tag{69}$$

where the last term is zero for $r = 0$. If $n\mu \ge \max\{2m_\star, 8\log(2/\beta)\}$, its exposure-test power at level $\gamma \in (0, 1)$ is at least $1 - \beta$.

Conversely, consider the zero/unit-exposure family constructed in Appendix G.9: unit-exposure approvals have mass μ and Bernoulli labels, zero-exposure approvals mass $\rho - \mu$, and rejections mass $1 - \rho$. A certifier valid at level $0 < \delta < 1 - \beta$ for every label parameter $q > \alpha_W$ in this family needs

$$n \ge \frac{\mathrm{kl}(1 - \beta, \delta)}{\mu\, \mathrm{kl}(r, \alpha_W)} \tag{70}$$

to achieve power $1 - \beta$ at parameter r. The fixed policy's support must admit every positive-mass type. Both assertions concern a fixed policy and the same frozen development object.

For fixed α_W, β, $\gamma = \delta \le (1 - \beta)/2$, and $r \uparrow \alpha_W$, these bounds match the worst-case order $\Theta(\log(1/\delta)/[\mu(\alpha_W - r)^2])$. The upper bound covers arbitrary exposure distributions, while an explicit zero/unit-exposure subfamily establishes the worst-case lower bound. Appendix G.9 gives the finite-sample proof and the exact power expression.

For a general safe population, let $d = \mu(\alpha_W - R_W) > 0$ and $v = \mathbb{E}[\mathbf{1}_A w^2(\alpha_W - Y)^2]$. These are population moments. With $\alpha_W \le 1/2$ and a predetermined grid covering the positive-drift stake scale, Proposition 18 gives a sufficient sample size of order $[v/d^2 + (1 - \alpha_W)/d][\log(J/\gamma) + \log(2/\beta)]$. Its explicit constants include the approval-count guard. This connects the exposure-only family to nonzero-default populations: both the risk margin and outcome-dependent second moment affect certification power. The grid condition directly affects power: when $0 < d < \lambda_{\min} v$, every log-wealth component has negative drift, so $p_E \to 1$ almost surely despite $R_W < \alpha_W$. For $A \equiv 1$, $w \equiv 1$, $\alpha_W = .1$, and $Y \sim \mathrm{Bernoulli}(.099)$, $d = .001$ and $v = .0892$ satisfy this condition. Thinning's general power guarantee avoids this fixed-grid limitation. Appendix G.8 details both regimes.

In contrast, on P_0 the range-only Hoeffding test has the exact threshold

$$m_H = \max\left\{5, \left\lceil \frac{\log(1/\gamma)}{2\alpha_W^2 a^2} \right\rceil \right\}. \tag{71}$$

Its a^{-2} dependence contrasts with the achievable a^{-1} dependence. This loss of efficiency comes from the range-only test. We therefore also compare HB and betting on real data. A Bonferroni implementation substitutes $\gamma = \delta/M$ in the sufficient conditions; Theorem 11 does not assert a universal $\log M$ lower bound.

G Proofs and Statistical Scope

With the lending contract in place, we can state precisely what its certificate guarantees. This section proves the component tests under fixed development information, combines them into a post-selection guarantee, and then studies the sample requirements of the resulting procedures.

G.1 Objects, conditioning, and conventions

Before proving the component tests, we make explicit what is fixed by development data and what remains random during certification. This conditioning is the common basis for every result in the section. All inferential statements condition on the development information $\mathcal{F}_0$ in Assumption 1. It determines the scorers, routing rules, finite policy family, exposure function, caps, mixture weights,

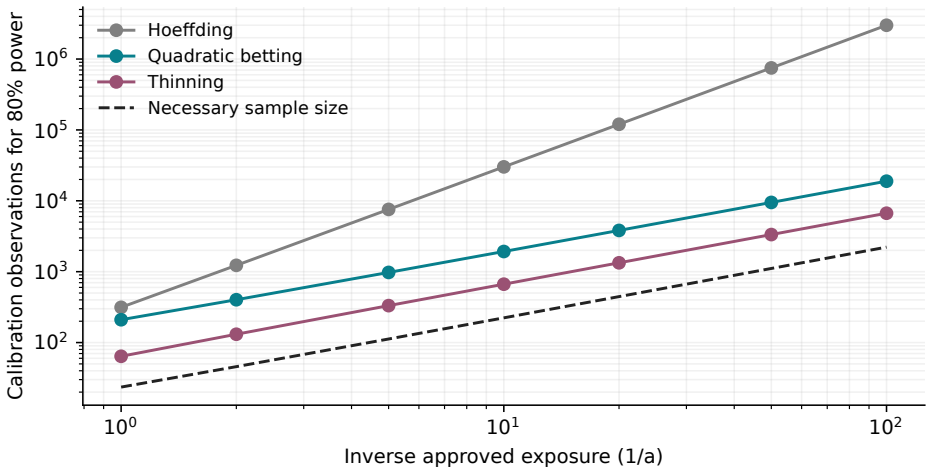

Figure 10: Exact sample sizes for 80% exposure-test power on P_0, with $\rho = .5$, $\alpha_W = .1$, and $\gamma = \delta = .05$. The dashed curve is Theorem 11 with $q = .2$, $\beta = .2$. We evaluate thinning and the full betting mixture with exact binomial tails. The betting curve uses the full mixture rather than the sufficient threshold m_E; the finite-sample constants differ despite matching rates.

minimum-count rules, and any testing order. A fixed policy is therefore a measurable function of a fresh application, rather than a function refitted on its certification labels. We use the independent test set only after selection and allow dependence between different policies' test statistics throughout.

Table 19: Notation used in the proofs. Population quantities refer to the certification distribution.

Symbol	Meaning
A_π, D_π	Approval and deferral indicators of a fixed policy
Y, w	Adverse-outcome indicator and exposure in $[0, 1]$
R_A, R_D, R_W	Conditional count risk, deferral probability, exposure risk
n, N_A, N_T	Total, approved, and thinned-approved sample counts
K_A, K_D, K_T	Approved defaults, deferrals, and retained defaults
α_A, α_W, B	Prespecified operating caps
δ, γ	Family-wise level and local policy-test level
M, J	Number of candidate policies and betting mixture components
ρ, a, μ	Approval probability, safe approved exposure, $\mu = \rho a$
β	Failure probability in a power requirement $1 - \beta$

The implementation uses a minimum of five approvals for approval-dependent tests and five retained approvals for thinning. These conservative rules only remove rejections and thus preserve the validity of the binomial p-value. Exposure tests return one when observed exposure is zero; positive population exposure ensures that the target ratio exists. The proofs count rejection at the cap as a false rejection, so they also control policies whose risk strictly exceeds the cap.

G.2 Conditional sampling and binomial tests

We start with the count components of the contract. Conditioning on the number of approved applications reduces both approval risk and review demand to ordinary binomial statements.

Lemma 14 (A random approval count)**.** *For i.i.d. observations and a fixed policy, conditional on $N_A = m > 0$, the approved observations in source order have the product conditional law $P(\cdot \mid A_\pi = 1)^m$. In particular, $K_A \mid N_A = m \sim \mathrm{Bin}(m, R_A)$.*

Proof. First condition on the complete vector of approval indicators. Independence of the original observations makes the conditional law of every approved observation $P(\cdot \mid A_\pi = 1)$, independently of the others. For any indicator vector with m ones, the ordered list of approved observations has the

same product law. Mixing over indicator vectors conditional on their sum therefore preserves that law. Taking their binary labels proves the binomial statement. □

For a fixed m, the binomial lower-tail p-value $F_{m,\alpha}(K)$ is super-uniform when $K \sim \mathrm{Bin}(m, r)$ and $r \geq \alpha$. To see this, let k_t be the largest integer whose CDF at α is at most t; the rejection event is $K \leq k_t$. Binomial stochastic ordering gives $\Pr_r(K \leq k_t) \leq F_{m,\alpha}(k_t) \leq t$. Empty rejection regions cause no difficulty. Lemma 14 proves this bound conditional on every m, and averaging over N_A proves unconditional validity. Replacing the p-value by one for $m < 5$ can only remove rejections. Deferrals are i.i.d. Bernoulli indicators without this conditioning step.

G.3 Exposure tests on the approved sample

The count components follow from conditional binomial laws. The exposure component instead uses the amount attached to each approved application, so we recenter the weighted loss and compare three valid ways to test its mean. Let $\alpha = \alpha_W$ in this subsection. Positive approved exposure gives

$$\mathbb{E}[Z \mid A_\pi = 1] = \mathbb{E}[w \mid A_\pi = 1](\alpha - R_W), \qquad Z = w(\alpha - Y). \tag{72}$$

The exposure null therefore implies a nonpositive centered mean among approved observations. We condition on their count so that all three bounded-mean tests below use the same approved sample.

The range length of Z is one. Under a nonpositive mean, Hoeffding's inequality (Hoeffding, 1963) gives $\Pr(S/m \geq t \mid N_A = m) \leq \exp(-2mt^2)$ for $t > 0$. Inverting this tail bound yields $p_H = \exp(-2(S_+)^2/m)$. This comparator uses only the known support $[-(1-\alpha), \alpha]$. The support bound must hold for the population, even if the observed exposure maximum is smaller.

G.3.1 The Hoeffding–Bentkus comparator

The first direct comparator maps the exposure problem to a bounded loss and applies a standard lower-tail bound. This provides a familiar baseline for the betting and thinning constructions that follow. Set $L = \alpha - Z = \alpha(1-w) + wY \in [0,1]$ and let $\bar{L}$ be its ordinary mean over the $m = N_A$ approved observations. Then $\mathbb{E}[L \mid A] - \alpha = \mathbb{E}[w \mid A](R_W - \alpha)$. The LTT Hoeffding–Bentkus p-value is

$$p_{\mathrm{HB}} = \min\left\{\exp[-mh_1(\bar{L} \wedge \alpha, \alpha)], \quad eF_{m,\alpha}(\lceil m\bar{L} \rceil)\right\}, \tag{73}$$

where $h_1(u,v) = u\log(u/v) + (1-u)\log((1-u)/(1-v))$ uses continuous endpoint values, including $h_1(0,v) = -\log(1-v)$. This is Proposition 1 of the LTT preprint (Angelopoulos et al., 2025): PDF version 5, page 4, Equation (1); its HTML equation is numbered (2). The Bentkus inequality is due to Bentkus (2004).

The formula requires the ceiling, inclusive binomial CDF, and factor e. Only the binomial branch receives the factor e, and only the KL branch truncates the mean at α. Their minimum remains valid because both branches bound the same ordered tail event; arbitrary tests do not generally share this property. Conditioning on m and using Lemma 14 establishes validity with the random approval count. Our implementation uses stable relative-entropy functions and returns one when the approval guard fails. We use the transformed loss only for testing; Pareto coordinates remain the original empirical risks in Equation 60.

G.3.2 Quadratic betting with sufficient statistics

The Hoeffding–Bentkus test uses only a range bound. Quadratic betting also uses the observed second moment, allowing the evidence to reflect how exposure is distributed within the approved sample.

Lemma 15 (A conditional mixture e-value)**.** *Fix $\alpha \in (0,1)$ and finitely many predetermined $\lambda_j \in [0, 1/(2(1-\alpha))]$ with nonnegative weights summing to one. Under $R_W \geq \alpha$, the weighted mixture of $\exp(\lambda_j S - \lambda_j^2 Q)$ has conditional expectation at most one given every positive approval count, and hence is an e-value after mixing over that count.*

Proof. For $x \geq -1/2$, define $f(x) = \log(1+x) - x + x^2$. Its derivative is $x(1+2x)/(1+x)$, which is nonpositive on $[-1/2, 0]$ and nonnegative on $[0, \infty)$. Since $f(0) = 0$, $f(x) \geq 0$ on the stated domain, or $\exp(x - x^2) \leq 1 + x$.

The lower support bound on Z and the restriction on λ_j give $\lambda_j Z \geq -1/2$. Equation 72 yields

$$\mathbb{E}[\exp(\lambda_j Z - \lambda_j^2 Z^2) \mid A] \leq 1 + \lambda_j \mathbb{E}[Z \mid A] \leq 1. \tag{74}$$

Conditional on $N_A = m$, Lemma 14 makes these factors independent. Their product has mean at most one and equals $\exp(\lambda_j S - \lambda_j^2 Q)$. Averaging with predetermined weights preserves the expectation bound. For $m = 0$ use the constant e-value one. Another expectation over m completes the argument. Markov's inequality now gives $\Pr(\min\{1, E^{-1}\} \leq t) \leq t$ for $0 < t < 1$. □

This conservative construction follows from bounded-mean betting (Waudby-Smith & Ramdas, 2024), with a fixed rather than learned stake grid. For binary Y, the cached moments suffice:

$$S = \alpha \sum_{i:A_i} w_i - \sum_{i:A_i} w_i Y_i, \qquad Q = \alpha^2 \sum_{i:A_i} w_i^2 + (1 - 2\alpha) \sum_{i:A_i} w_i^2 Y_i. \tag{75}$$

The validity argument therefore depends directly on these sums. The code computes the mixture with log-sum-exp for numerical stability. A single predetermined bet also gives a valid test; the argument requires its stake to be independent of the certification outcomes.

G.3.3 Thinning and distribution tilting

Betting keeps the weighted observations intact. Thinning takes the complementary route: it converts exposure into a sampling probability and then applies an exact binomial test to the retained labels.

Lemma 16 (Exposure-proportionate sampling)**.** *Let $T = \mathbf{1}\{A_\pi = 1, U < w\}$ with independent uniform U. If $\mathbb{E}[w\mathbf{1}_A] > 0$, then $\Pr(Y = 1 \mid T = 1) = R_W$. Conditional on the retained count, retained labels are i.i.d. Bernoulli(R_W).*

Proof. Conditioning on the application and label gives $\mathbb{E}[T \mid X, Y] = w\mathbf{1}_A$. Therefore $\Pr(Y = 1, T = 1) = \mathbb{E}[wY\mathbf{1}_A]$ and $\Pr(T = 1) = \mathbb{E}[w\mathbf{1}_A]$; their ratio is R_W. Apply Lemma 14 to the augmented i.i.d. observations (X, Y, U) and selection indicator T. □

The exposure p-value is $F_{N_T, \alpha_W}(K_T)$, which we replace by one when $N_T < 5$. This procedure applies rejection sampling to a label ratio (Devroye, 1986; Zadrozny et al., 2003). We share each application's uniform draw across candidates, inducing dependence that the outer family-wise error procedures allow. The guarantee averages over the coins, and our experiments record one predetermined random stream.

G.4 Complete contracts and post-selection validity

The preceding tests handle one component of one fixed policy at a time. We now combine the components and then account for searching across policies, which yields the complete-contract guarantee used in the experiments. For policy π, write its closed unsafe null as $H_\pi = H_{\pi,A} \cup H_{\pi,W} \cup H_{\pi,D}$. If this union is true, at least one component, say k, has a valid null p-value. Then

$$\Pr(p_\pi \leq t) = \Pr(\max_j p_{\pi,j} \leq t) \leq \Pr(p_{\pi,k} \leq t) \leq t. \tag{76}$$

This intersection–union argument does not need independence or a within-policy Bonferroni factor. It is the multi-risk construction already used by LTT (Berger, 1982; Angelopoulos et al., 2025).

Proof of Proposition 10. For Bonferroni-IUT, union-bound the valid policy-level tests at δ/M over all true policy nulls. For Holm, suppose a true null is first rejected at ordered rank j and there are M_0 true nulls. All true nulls are among the remaining $M - j + 1$ hypotheses, so its critical value is no larger than δ/M_0. A false rejection therefore implies that some true null p-value is at most δ/M_0; another union bound gives δ.

For a fixed sequence, condition on the independent development object and its order. Let j_0 be its first true null. A false rejection requires reaching and rejecting that hypothesis. Its p-value is tested at δ, so this event has probability at most δ, irrespective of the preceding false-null tests. When no true null is in the sequence, the error probability is zero. Pareto filtering and ordering are functions of the same development object, so the identical argument applies. Finally, selecting any member of a wholly safe returned set remains safe, even if calibration utility chooses that member. □

The guarantee bounds the chance of any unsafe return under a fixed contract. A deployment protocol must fix its cap before certification or account for the expanded search when choosing among caps. The reported guarantee does not condition on the event that the procedure returns a policy.

Proposition 17 (What Holm can improve)**.** *On the same finite family and p-values, the Bonferroni-IUT rejected set is a subset of the Holm rejected set, and the two sets are nonempty on exactly the same event. Maximizing the same calibration utility over those sets makes Holm's attained calibration utility weakly larger whenever a policy is returned. This implication concerns the calibration objective.*

Proof. Both procedures reject at least one null exactly when $p_{(1)} \leq \delta/M$. If $p_{(r)} \leq \delta/M$, then for every $j \leq r$, $p_{(j)} \leq \delta/M \leq \delta/(M-j+1)$, so Holm reaches and rejects rank r. This proves inclusion. A maximization over a superset cannot reduce the attained value of the same objective. Independent test utility need not be ordered in the same way. □

G.5 Numerical confidence bounds versus tests

The preceding results certify a set of policies at a fixed level. Reporting a numerical confidence bound after selection is a different task, so we clarify here why a bound obtained from the observed Holm rank is not automatically simultaneous. A fixed-level one-sided Clopper–Pearson upper bound (Clopper & Pearson, 1934) for k events in m trials is

$$U(k,m;\eta) = \begin{cases} 1, & m = 0 \text{ or } k = m, \\ \mathrm{Beta}^{-1}(1-\eta; k+1, m-k), & 0 \leq k < m. \end{cases} \tag{77}$$

For K reported Bernoulli components on each of M frozen policies, using $\eta = \delta/(KM)$ gives simultaneous coverage by the conditional argument and a union bound. Approval and deferral use their respective sample sizes. The historical selector also reports a rejection-error component; its bounds use $K = 3$. Current contract comparisons report caps and adjusted test p-values instead of weighted-risk confidence endpoints.

The Holm critical level depends on the observed rank. An ordinary Clopper–Pearson interval requires a fixed confidence level, so direct inversion of the Holm threshold does not establish simultaneous coverage. The code distinguishes such rank diagnostics from fixed-level confidence bounds. Likewise, the development score bands used to propose routing policies are not certificates for selected portfolios.

G.6 The rare-contamination lower bound

Having established validity, we turn to power. The first construction shows why a seemingly safe low-exposure population can still require many fresh observations when rare high-exposure errors are allowed. Fix the development object d and a policy whose approval region contains both an admissible type with exposure $a \in (0,1]$ and one with exposure one. Both types can have the labels stipulated below. If $\rho < 1$, its reject region contains a type to hold the remaining probability mass. The population class in Theorem 11 includes these support conditions, particularly an approval region that admits high-exposure types.

Under P_0, the first type is approved with probability ρ, has label zero, and contributes exposure a; remaining observations are rejected. Let H put all mass on an approved unit-exposure default, and define P_1 and ε as in Equation 64. Directly,

$$\mathbb{E}_{P_1}[wY\mathbf{1}_A] = \varepsilon, \qquad \mathbb{E}_{P_1}[w\mathbf{1}_A] = (1-\varepsilon)\rho a + \varepsilon, \tag{78}$$

$$\Pr_{P_1}(Y=1, A) = \varepsilon, \qquad \Pr_{P_1}(A) = (1-\varepsilon)\rho + \varepsilon. \tag{79}$$

Substitution yields

$$R_W(P_1) = q, \qquad R_A(P_1) = \frac{qa}{1-q+qa}, \qquad R_D(P_1) = 0. \tag{80}$$

Thus P_1 violates its exposure cap for any $q > \alpha_W$. Its count risk is still at most α_A whenever $a \leq \alpha_A(1-q)/(q(1-\alpha_A))$. This additional inequality is useful for isolating an exposure-only failure but is not needed to make the lower-bound population unsafe.

Full proof of Theorem 11. Generate each P_1 observation by an independent contamination indicator with probability ε, drawing from P_0 otherwise. Let G be the event that all n indicators are zero. Then $\Pr_{P_1}(G) = (1-\varepsilon)^n$ and the sample law conditional on G is exactly P_0^n. The certifier receives the same fixed object d in both populations. Its independent internal randomness has the same law, so $\Pr_{P_1}(C \mid G, d) = \Pr_{P_0}(C \mid d)$. Conditional uniform validity at P_1 and power at P_0 imply

$$\delta \geq \Pr_{P_1}(C \mid d) \geq (1-\varepsilon)^n \Pr_{P_0}(C \mid d) \geq (1-\varepsilon)^n(1-\beta). \tag{81}$$

For $\beta \in (0,1)$ and $0 < \delta < 1-\beta$, logarithms and division by $-\log(1-\varepsilon) > 0$ give the claimed bound. □

The theorem fixes the same development object d under both populations and requires conditional validity for every allowed P. Its lower bound therefore concerns fresh evidence for a fixed policy. A joint analysis of training and certification would require a different population class and validity requirement.

G.7 Attainable power and complete-contract conditions

The lower bound identifies the difficult contamination mechanism. We now show that thinning and betting attain the same exposure scaling on the safe family, while also accounting for the other contract components.

Full proof of Theorem 12. On P_0, every approved label is zero and $N_T \sim \mathrm{Bin}(n, \rho a)$. For $N_T \geq 5$, the thinning p-value is $(1-\alpha_W)^{N_T}$. It is at most γ exactly at the integer threshold in Equation 66. Hence its exact power is $\Pr\{\mathrm{Bin}(n, \rho a) \geq m_T\}$.

For betting, $S = N_A \alpha_W a$ and $Q = N_A \alpha_W^2 a^2$. The prescribed grid contains $\lambda_* = [2(1-\alpha_W)]^{-1}$ with mixture weight $1/J = 1/16$. Keeping this summand alone gives $E \geq J^{-1}\exp(N_A g(a))$. Put $t = \lambda_* \alpha_W a$. For $0 < \alpha_W \leq 1/4$ and $a \leq 1$, $0 < t \leq 1/6$ and

$$g(a) = t(1-t) \geq \frac{5}{6}t = \frac{5\alpha_W a}{12(1-\alpha_W)} \geq \frac{5}{12}\alpha_W a > 0. \tag{82}$$

The sufficient threshold m_E follows by requiring $N_A g(a) \geq \log(J/\gamma)$ and $N_A \geq 5$.

For either binomial count N with mean ν, a multiplicative Chernoff bound gives $\Pr(N < \nu/2) \leq e^{-\nu/8}$. If $\nu \geq 2m$ and $\nu \geq 8\log(1/\beta)$, the event $N < m$ is contained in $N < \nu/2$ and has probability at most β. Substitute $(\nu, m) = (n\rho a, m_T)$ for thinning and $(n\rho, m_E)$ for betting. □

These statements concern the exposure component at a local test level. For completeness, define

$$m_A = \max\left\{5, \left\lceil \frac{\log\gamma}{\log(1-\alpha_A)} \right\rceil\right\}, \qquad n_D = \left\lceil \frac{\log\gamma}{\log(1-B)} \right\rceil. \tag{83}$$

On P_0, the count component passes when $N_A \geq m_A$ and the capacity component passes when $n \geq n_D$, since every deferral count is zero. If $\alpha_A \geq \alpha_W$, then $m_A \leq m_T$ and $N_A \geq N_T$; the thinning exposure event already implies the count event. Thus its exact joint power is

$$\mathbf{1}\{n \geq n_D\}\Pr\{\mathrm{Bin}(n, \rho a) \geq m_T\}. \tag{84}$$

For betting, replace m_E by $\max(m_E, m_A)$ in its sufficient condition and also require $n \geq n_D$. These are events concerning the same approval count; no independence between the component tests is presumed. With Bonferroni-IUT one sets $\gamma = \delta/M$. For a fixed sequence, these are marginal power bounds for a fixed candidate's local test. Let G_j be the event that the sequence reaches candidate j, and let C_j be the event that its complete local contract passes. If $\Pr(C_j^c \mid \mathcal{F}_0) \leq \beta_j$, then

$$\Pr(G_j \cap C_j \mid \mathcal{F}_0) \geq \max\{0, \Pr(G_j \mid \mathcal{F}_0) - \beta_j\}. \tag{85}$$

To obtain this bound, subtract $\Pr(G_j \cap C_j^c \mid \mathcal{F}_0) \leq \beta_j$ from $\Pr(G_j \mid \mathcal{F}_0)$. If each local contract up to position j has failure probability at most β_i, a union bound gives a probability of at least $1 - \sum_{i \leq j}\beta_i$ for certifying candidate j. The tests can be dependent, and a marginal power bound alone does not control power conditional on reaching the candidate.

Take $q = 2\alpha_W \le 1/2$, $\mu = \rho a$, fixed $\beta \in (0,1)$, $\gamma = \delta$, and $0 < \delta \le (1-\beta)/2$. Since

$$-\log(1-\varepsilon) = \log\left(1 + \frac{q\mu}{1-q}\right), \qquad \alpha_W \mu \le -\log(1-\varepsilon) \le 4\alpha_W \mu, \tag{86}$$

the lower bound has order $\log((1-\beta)/\delta)/(\alpha_W\mu)$. For the left inequality, $x = q\mu/(1-q) \le 1$ and $\log(1+x) \ge x/2$; for the right inequality, $\log(1+x) \le x$ and $q/(1-q) \le 4\alpha_W$. The separation $\delta \le (1-\beta)/2$ makes $\log((1-\beta)/\delta)$ comparable to $\log(1/\delta)$ with constants depending on fixed β. The thresholds m_T and m_E give the matching upper order in Equation 68. The separation condition keeps the lower numerator away from zero as a fraction of $\log(1/\delta)$. The complete-contract guarantee also requires the count and capacity conditions stated above.

On P_0, $p_H = \exp(-2N_A\alpha_W^2 a^2)$, so the raw range-only test passes when $N_A \ge \lceil \log(1/\gamma)/(2\alpha_W^2 a^2) \rceil$. Including the implemented approval guard gives m_H in Equation 71. Its exact exposure-test power is $\Pr\{\mathrm{Bin}(n,\rho) \ge m_H\}$. The full betting mixture can pass before m_E: that threshold uses just one mixture summand. Numerical comparisons evaluate the production mixture, not this looser sufficient bound.

G.8 A SECOND-MOMENT POWER BOUND FOR BOUNDED LOSSES

The preceding construction uses binary defaults and a simple exposure pattern. To cover the broader setting used in the paper, we next derive a power condition for arbitrary bounded exposure and loss. The sufficient condition below applies to every bounded exposure and loss, including fractional losses. Let $0 < \alpha \le 1/2$, $c = 1-\alpha$, $\rho = \Pr(e > 0) > 0$, and

$$\widetilde{Z} = e(\alpha - L), \quad d = \mathbb{E}\widetilde{Z} = \mu(\alpha - r) > 0, \quad v = \mathbb{E}\widetilde{Z}^2, \quad \mu = \mathbb{E}[e]. \tag{87}$$

These moments use the full population, with zero contributions from unexposed observations. The power condition concerns their population values. In the lending application, $e = \mathbf{1}_A w$ and $L = Y$ on approval recover the credit-specific quantities.

Proposition 18 (A general sufficient power condition)**.** *Under Assumption 1, suppose the predetermined, uniform J-component betting mixture contains a fixed $\lambda \in (0, 1/(2c)]$ with $h = \lambda d - \lambda^2 v > 0$. For $\gamma, \beta \in (0,1)$, put $\ell = \log(J/\gamma)$ and $t = \log(2/\beta)$. The guarded exposure test has power at least $1-\beta$ at level γ if*

$$n \ge \max\left\{\frac{2\ell}{h}, \quad \left(\frac{18\lambda^2 v}{h^2} + \frac{4\lambda c}{h}\right)t, \quad \frac{\max\{10, 8t\}}{\rho}\right\}. \tag{88}$$

The guard requires five positive-exposure observations. The same conclusion holds for the product mixture on the same grid, which dominates the quadratic mixture pointwise by Proposition 4.

Proof. Let $\xi = \lambda\widetilde{Z} - \lambda^2\widetilde{Z}^2$, so $\mathbb{E}\xi = h$. Since $|\widetilde{Z}| \le c$ and $\lambda c \le 1/2$,

$$|\xi| \le \tfrac{3}{2}\lambda|\widetilde{Z}|, \qquad \mathrm{Var}(\xi) \le \tfrac{9}{4}\lambda^2 v, \qquad |h - \xi| \le 3\lambda c. \tag{89}$$

If $\mathrm{Var}(\xi) = 0$, then $\xi = h$ almost surely and the lower-tail failure event below is impossible. Otherwise, the one-sided Bernstein bound for the independent full-population observations gives

$$\Pr\left(\sum_{i=1}^n \xi_i < \frac{nh}{2}\right) \le \exp\left[-\frac{nh^2}{8\,\mathrm{Var}(\xi) + (4/3)(3\lambda c)h}\right] \le \exp\left[-\frac{nh^2}{18\lambda^2 v + 4\lambda c h}\right]. \tag{90}$$

The second sample-size condition makes this at most $\beta/2$; the first gives $nh/2 \ge \ell$. The third condition and a binomial Chernoff bound imply $\Pr(N_+ < 5) \le e^{-n\rho/8} \le \beta/2$. Here $N_+ = \sum_i \mathbf{1}\{e_i > 0\}$. Outside these two failure events, the chosen mixture summand gives

$$E \ge J^{-1}\exp\left(\sum_i \xi_i\right) = J^{-1}\exp(\lambda S - \lambda^2 Q) \ge 1/\gamma. \tag{91}$$

Since $\ell > 0$, this event also implies positive observed exposure. A union bound proves the result without assuming independence between the positive-exposure count and the wealth.

For completeness, the Bernstein form follows by bounding the centered variable $W = h - \xi$ in absolute value by $b = 3\lambda c$. Expanding its moment generating function and using $k! \geq 2 \cdot 3^{k-2}$ for $k \geq 2$ gives $\log \mathbb{E}e^{sW} \leq s^2 \operatorname{Var}(W)/(2(1 - sb/3))$ for $0 < s < 3/b$. For positive variance, Chernoff optimization at $s = u/(n \operatorname{Var}(W) + bu/3)$ gives $\exp[-u^2/(2n \operatorname{Var}(W) + 2bu/3)]$. Taking $u = nh/2$ gives Equation 90. The zero-variance case was handled separately, so this optimization never uses its inadmissible endpoint $s = 3/b$. □

Define $\lambda_\star = \min\{1/(2c), d/(2v)\}$. If the frozen grid contains a component in $[\lambda_\star/2, \lambda_\star]$, then $h \geq \lambda d/2$ and $1/(\lambda d) \leq 4 \max\{c/d, v/d^2\}$. Consequently a sufficient sample size has order

$$O\left[\left(\frac{v}{d^2} + \frac{c}{d}\right)\left(\log\frac{J}{\gamma} + \log\frac{2}{\beta}\right)\right]. \tag{92}$$

This expression also covers the minimum-count requirement because $d \leq \rho\alpha \leq \rho c$, so $c/d \geq 1/\rho$. The constants can also be made explicit. Since $v \geq d^2 > 0$, the scale $\lambda_\star$ is well-defined. For a covered component,

$$\frac{2\ell}{h} \leq \frac{4\ell}{\lambda d} \leq 16\ell \max\left\{\frac{c}{d}, \frac{v}{d^2}\right\}, \tag{93}$$

$$\left(\frac{18\lambda^2 v}{h^2} + \frac{4\lambda c}{h}\right) t \leq \left(\frac{72v}{d^2} + \frac{8c}{d}\right) t. \tag{94}$$

Also $t > \log 2$ and $d \leq \rho c$ imply $\max\{10, 8t\}/\rho \leq 15ct/d$. Thus the sufficient budget

$$n \geq 16\ell \max\left\{\frac{c}{d}, \frac{v}{d^2}\right\} + \left(\frac{72v}{d^2} + \frac{15c}{d}\right) t \tag{95}$$

implies all three conditions in equation 88 and justifies equation 92 with numerical constants.

The original quadratic grid has adjacent ratio $32^{1/15} < 2$ and spans $[1/(64c), 1/(2c)]$; it supplies this coverage when $\lambda_\star \geq 1/(64c)$. Very small safe margins can violate that condition. A wider grid preserves validity if the experiment fixes it before certification; our experiments use the prespecified grids.

The same drift calculation establishes a failure regime for the quadratic wealth on a fixed grid. If $0 < d < \lambda_{\min} v$, then $\lambda_j d - \lambda_j^2 v < 0$ for every member of the finite grid. The strong law gives $n^{-1}(\lambda_j S - \lambda_j^2 Q) \to \lambda_j d - \lambda_j^2 v < 0$ almost surely. Each wealth component tends to zero, hence $E \to 0$ and the guarded $p_E \to 1$. For example, $e \equiv 1$, $\alpha = .1$, and $L \sim \text{Bernoulli}(.099)$ give $d = .001$ and $v = .0892$, while $\lambda_{\min} = 1/57.6$ and $\lambda_{\min} v > .001$. A strictly safe policy can have asymptotically zero certification probability under this quadratic statistic. The product statistic on the same grid behaves differently. Its smallest stake has log growth

$$.901 \log(1 + .1/57.6) + .099 \log(1 - .9/57.6) \approx 3.7925 \times 10^{-6} > 0, \tag{96}$$

whereas its quadratic counterpart has drift -9.5245×10^{-6}. The positive sign also has a rational check: the inequalities $\log(1 + x) \geq x - x^2/2$ for $x \geq 0$ and $\log(1 - y) \geq -y - y^2/[2(1 - y)]$ for $0 \leq y < 1$ give a lower bound $17309/4644864000 > 0$ at $x = 1/576$ and $y = 1/64$. The strong law makes that product component diverge, so its certification probability tends to one. The example identifies a loss from the quadratic approximation, rather than a failure of every product test on the grid.

On P_0, $d = \rho a \alpha$ and $v = \rho a^2 \alpha^2$, so $v/d^2 = 1/\rho$ and $c/d = (1 - \alpha)/(\rho a \alpha)$. These identities recover the inverse-exposure rate. More generally, risk margin, exposure, and its outcome-dependent second moment all affect power. Equation 88 bounds the power of the exposure test. A complete contract also requires the count and capacity tests to pass and the fixed sequence to reach the policy. The proposition follows by combining standard Bernstein concentration with the fixed mixture.

G.9 General thinning power and the KL hard family

The second-moment result explains direct betting, whereas thinning depends only on the retained-unit law. We now prove the general thinning bound and match it with a KL-based hard family, thereby proving Theorem 13. Throughout, condition on the same frozen development object and fixed policy.

Let $0 < \mu \le \rho \le 1$, $0 \le r < \alpha = \alpha_W < 1$, and $\gamma, \beta \in (0,1)$. Count and capacity requirements are separate from the exposure component analyzed here.

For $T_i = \mathbf{1}_{A_i}\mathbf{1}\{U_i < w_i\}$, independence of the uniform coins gives

$$\Pr(T_i = 1) = \mu, \qquad \Pr(T_i = 1, Y_i = 1) = \mu r. \tag{97}$$

The pairs (T_i, Y_i) are i.i.d. Thus the retained sample count is $N_T \sim \mathrm{Bin}(n, \mu)$ and the conditional default count is $K_T \mid N_T = m \sim \mathrm{Bin}(m, r)$. With the minimum of five retained observations, the exact power is

$$\Pr(p_T \le \gamma) = \sum_{m=5}^{n} \Pr\{\mathrm{Bin}(n,\mu) = m\} \Pr\{F_{m,\alpha}(\mathrm{Bin}(m,r)) \le \gamma\}. \tag{98}$$

Consequently this test's power depends on $(n, \mu, r, \alpha, \gamma)$, not on further features of the exposure distribution. At $r = 0$, the exact retained-count threshold is m_T in equation 66. The midpoint argument below gives the sufficient threshold $m_\star$; the two thresholds analyze the same test at different levels of sharpness.

Write $\mathrm{kl}(u, v) = u \log(u/v) + (1-u)\log((1-u)/(1-v))$ with continuous endpoint conventions. Put $s = (r+\alpha)/2$ and use $m_\star$ from Equation 69. For $m \ge m_\star$, the event $K_T/m \le s < \alpha$ implies

$$F_{m,\alpha}(K_T) \le e^{-m\,\mathrm{kl}(K_T/m,\alpha)} \le e^{-m\,\mathrm{kl}(s,\alpha)} \le \gamma. \tag{99}$$

The first inequality is the binomial Chernoff bound (Hoeffding, 1963); the second uses that $\mathrm{kl}(u, \alpha)$ decreases for $u < \alpha$. When $r > 0$, the upper-tail Chernoff bound under the true retained parameter gives

$$\Pr(K_T/m > s \mid N_T = m) \le e^{-m\,\mathrm{kl}(s,r)} \le \beta/2. \tag{100}$$

When $r = 0$, $K_T = 0$ almost surely and this failure event is absent. Here $\mathrm{kl}(s, 0) = +\infty$ explains the zero convention in Equation 69. Finally, $n\mu \ge 2m_\star$ implies

$$\Pr(N_T < m_\star) \le e^{-n\mu/8} \le \beta/2. \tag{101}$$

A union bound over the two failure events proves the stated power, regardless of their dependence.

The fixed approval region admits a unit-exposure type and, if needed, a zero-exposure type. A reject type is required only when $\rho < 1$. Define P_r as follows:

Type	Probability	Exposure	Conditional label
Approved unit exposure	μ	1	$\mathrm{Bernoulli}(r)$
Approved zero exposure	$\rho - \mu$	0	0
Rejected	$1 - \rho$	fixed	0

Rows of zero probability need not be present. For each $q \in (\alpha, 1)$, let P_q change only the first row's label parameter to q. Directly,

$$R_W(P_r) = r, \qquad R_W(P_q) = q, \qquad \mathrm{KL}(P_r \| P_q) = \mu\,\mathrm{kl}(r, q). \tag{102}$$

Suppose a certificate event C uses n fresh observations and independent internal randomness, with $\Pr_{P_r}(C) \ge 1 - \beta$ and $\Pr_{P_q}(C) \le \delta$ for every $q > \alpha$, conditional on the same development object. The internal randomness has the same law under both populations, so the relative entropy between the full inputs is $n\mu\,\mathrm{kl}(r, q)$. Mapping the input to $\mathbf{1}_C$ gives

$$n\mu\,\mathrm{kl}(r,q) \ge \mathrm{kl}(\Pr_{P_r}(C), \Pr_{P_q}(C)) \ge \mathrm{kl}(1-\beta, \delta). \tag{103}$$

For completeness, the first inequality follows by splitting the KL integral over C and its complement and applying Jensen to each normalized integral (the log-sum inequality). The second follows because, for $p > q$, binary KL increases in p and decreases in q; here $\Pr_{P_r}(C) \ge 1 - \beta > \delta \ge \Pr_{P_q}(C)$. Zero or unit event probabilities use the continuous or infinite conventions. We obtain Equation 70 in the limit $q \downarrow \alpha$.

On this family $R_A(P_q) = \mu q/\rho$ and $R_D(P_q) = 0$. If $\mu q/\rho \le \alpha_A$, only the exposure contract is violated. This extra condition isolates the exposure constraint within a joint contract; it is not needed for the exposure-component lower bound.

Fix $\alpha, \beta \in (0,1)$, set $\gamma = \delta$, and require $0 < \delta \le (1-\beta)/2$. As $r \uparrow \alpha$,

$$\mathrm{kl}(r,\alpha) \sim \frac{(\alpha-r)^2}{2\alpha(1-\alpha)}, \qquad \mathrm{kl}(s,\alpha) \sim \mathrm{kl}(s,r) \sim \frac{(\alpha-r)^2}{8\alpha(1-\alpha)}. \tag{104}$$

Moreover, $\mathrm{kl}(1-\beta,\delta)$ is comparable to $\log(1/\delta)$ over the stated separated range, with constants depending on fixed β. This follows by continuity away from $\delta = 1-\beta$ and the positive limit of their ratio as $\delta \downarrow 0$. The sufficient condition and hard-family lower bound therefore yield

$$\Theta\left(\frac{\log(1/\delta)}{\mu(\alpha-r)^2}\right) \tag{105}$$

in this local worst-case sense. The constants can depend on the fixed α and β, and the expansion applies away from the endpoints $\alpha = 0$ and 1. The finite-sample relative-entropy bounds state the result without this local approximation. At $r = 0$, the lower denominator is $-\mu\log(1-\alpha)$, distinct from the infinite reverse-direction KL in the sufficient threshold. The hard support here differs from the low-exposure contamination construction in Theorem 11. Restricted supports and favorable second moments can admit better distribution-specific power, as analyzed in Proposition 18.

G.10 When amount information changes the target

The power results above treat the exposure target as distinct from count risk. We close the proof section by explaining when amount information truly changes the estimand and when the two risks coincide. Two distinctions help interpret the lower bound and the real-data results. First, if a population class is known to satisfy $w = a > 0$ on every approved application, then $R_W = R_A$ throughout that class. The count test already certifies exposure, with no inverse-a penalty. The lower-bound class instead allows a unit-exposure approved type that happens not to appear under P_0. Small observed weights do not rule out this possibility.

Second, a weighted risk need not exceed its count counterpart:

$$R_W = R_A + \frac{\mathrm{Cov}(w, Y \mid A)}{\mathbb{E}[w \mid A]}. \tag{106}$$

Under the unsafe P_1, this covariance is $(1-a)r(1-r)$ for $r = R_A(P_1)$, strictly positive when $0 < a < 1$ and zero when $a = 1$. It is also zero on the safe P_0. A credit cohort can have negative covariance in the selected portfolio, producing $\widehat{R}_W < \widehat{R}_A$. Both patterns are consistent with the target definition. The controlled construction establishes that the count contract alone is insufficient over the population class; the cohort study measures the additional certificate's cost in its observed setting.

H Candidate Construction and Testing Algorithms

To attribute empirical differences to certification, the learned policies must be held fixed across methods. This section describes how those candidates are produced and how independent development data determine the frontier and testing order.

H.1 Scorers and routing features

The comparison requires a diverse but fixed family of candidate policies. We first describe how the prediction scores are fitted and converted into the routing features used to build that family. We use balanced and unweighted logistic regression, a random forest, and a histogram gradient booster, with settings fixed across methods. Numeric and categorical preprocessing is fitted only on the training block, and early validation for the gradient booster also uses only that block. A fifth scorer averages the four raw predicted probabilities, which we calibrate separately. The exact estimator settings are retained in the supplementary configurations. These fixed scorers let us compare certification methods at the same prediction quality.

On the independent score-development block, the implementation fits both isotonic and logistic probability calibration, selecting the smaller in-sample Brier score (isotonic on ties or a single-class

block). It then sorts calibrated scores and creates $\max\{1, \min(20, \lfloor n_{\text{score}}/80 \rfloor)\}$ approximately equal-size bins. For a bin with k adverse labels out of m observations, the routing band uses the .05 and .95 quantiles of $\text{Beta}(k+1, m-k+1)$. Cumulative maxima enforce nondecreasing lower and upper edges. These are uniform-prior beta bands used as routing features, not frequentist simultaneous confidence bounds. Their statistical quality affects candidate usefulness, while the separate risk sample determines certification validity.

For each scorer, fixed quantile grids define approval, rejection, and band- width thresholds. Duplicate quantiles and invalid pairs with $\tau_A \geq \tau_R$ are removed. The precise grids are recorded in the supplementary configuration.

Point routing approves if $p < \tau_A$, rejects if $p > \tau_R$, and defers otherwise. Interval routing substitutes the upper edge for approval and the lower edge for rejection. Hybrid routing approves if the upper edge is below τ_A or if both the band is sufficiently narrow and $p < \tau_A$; the analogous lower-edge or narrow-band rule determines rejection. The policy defers overlapping and otherwise unmatched cases. We assign a stable identifier to each combination of model, routing mode, and thresholds. We count grid candidates after removing duplicate quantiles and invalid threshold pairs.

H.2 An independently ordered Pareto baseline

Once the candidates are fixed, development data determine which policies are worth testing and in what order. This subsection defines that ordering while keeping all certification labels out of the search. For each fixed cap, exposure test, and utility definition, we compute every candidate's original empirical approval risk and deferral rate on the score-development block. We add exposure risk for an exposure contract and negative development utility as the final coordinate. We retain the nondominated vectors, minimizing all coordinates. Our implementation uses weak dominance in every coordinate with strict improvement in at least one, retaining one deterministic representative of exact duplicate vectors. We sort that frontier by the maximum development component p-value, ascending, then by development utility and stable candidate index to break ties. HB and other exposure tests share the original risk coordinates but can induce different orders. On independent risk-calibration data, we test the ordered policy p-values at δ, retain the passing prefix, and stop at the first $p > \delta$. We select its maximum calibration-utility policy, breaking ties by smaller empirical deferral and stable index. This follows the development-frontier and independent fixed-sequence structure of Laufer-Goldshtein et al. (2023). Its public single-risk example uses frontier coordinates $(-\text{accuracy}, \text{cost})$ and descending accuracy. That example's code uses strict $p < \delta$ and a strict-all-coordinates dominance check; our stated boundary and dominance conventions specify the adaptation. Either test-boundary convention is valid for super-uniform p-values. The first failing policy is never included in the returned set.

The safety-ordered fixed sequence uses the same development maximum-p order without Pareto filtering. The utility-ordered sequence sorts by decreasing development utility. Event-wise Bonferroni tests every component at $\delta/(KM)$, with $K = 2$ or 3; policy Bonferroni-IUT uses δ/M on the maximum p-value. The latter isolates the gain from IUT instead of attributing that gain to Holm.

We distinguish planned sequence length from the number of hypotheses actually reached, including the first failure. Output agreement includes common empty outputs; agreement on a selected policy is evaluated only when both methods return one.

I Datasets, Calendar Blocks, and Utilities

The lending results can only be interpreted after fixing which outcomes are observed and when calibration takes place. This section defines the mature-loan cohort, calendar splits, comparison datasets, and utility measures that give all methods a common evaluation target.

I.1 Mature Lending Club cohorts

We begin with the primary lending cohort, where the timing and resolution of loan outcomes define which rows can enter the analysis. The filtering below keeps the target population explicit rather than treating unresolved loans as observed outcomes. The full accepted-loan archive contains 2,260,701 rows. Within 2012–2015 36-month originations, 506,760 Fully Paid and 82,728 Charged Off loans

Table 20: Original rolling blocks. Each training window spans 24 months and is subsampled to 60,000 loans using a fixed split. Score and risk blocks each span three complete months. All model repetitions share the row membership of a window.

Test quarter	Score n	Previous-risk n	Test n	Training months
2014 Q4	37,881	40,595	50,020	2012-04–2014-03
2015 Q1	40,595	50,020	56,568	2012-07–2014-06
2015 Q2	50,020	56,568	64,222	2012-10–2014-09
2015 Q3	56,568	64,222	73,567	2013-01–2014-12
2015 Q4	64,222	73,567	88,669	2013-04–2015-03

give the 589,488-row resolved cohort. We exclude 72 Current, 61 Late (31–120 days), 9 In Grace Period, and 5 Late (16–30 days) loans. These 147 unresolved rows are about .025% of this restricted origination pool. Other issue years, loan terms, and statuses are outside the defined cohort rather than relabeled as default/nondefault.

We select predictors from the loader's explicit list of origination fields. We exclude outcome and status fields, later payment totals, recoveries, collection fees, last-payment dates, and source row identifiers. Imputation, category encoding, utility normalization, and model fitting use the training role only. Issuance year/month are observable cohort features. This design excludes recorded post-outcome features, while the time at which final labels became available remains unknown.

I.2 Size-matched calibration comparisons

The rolling design changes both calendar time and calibration sample size. To isolate the calendar effect, we construct a paired comparison with equal calibration counts and identical held-out test applications. The paired study reuses every fitted model, score-development statistic, candidate policy, and training-derived utility specification from Table 20. A fixed permutation, independent of all labels, splits each original test quarter into risk-calibration and held-out test halves. Their respective sizes are $(25010, 25010)$, $(28284, 28284)$, $(32111, 32111)$, $(36783, 36784)$, and $(44334, 44335)$. The size-matched previous-quarter comparator subsamples exactly the same calibration count from the preceding risk block using a separate fixed permutation. Both procedures evaluate the identical held-out test rows. Approval and retained counts can differ across the two calibration populations despite equal total sample size.

Thinning uses separate fixed random streams for calibration and held-out diagnostic statistics. These permutations and streams are shared across the three model seeds rather than treated as extra independent replications. Membership archives permit explicit disjointness and test-identity checks. This design isolates the calibration cohort at fixed model, candidate family, evaluation data, and total calibration size. The same-quarter comparison provides a retrospective reference under exchangeability; it assumes that the mature labels are available.

I.3 Historical replay and additional public datasets

The mature-cohort analysis is the primary application, but it does not show how the procedures behave under the earlier pipeline or with much smaller samples. We therefore add one historical replay and three public-data comparisons with clearly separated roles. The historical Lending replay preserves the loader's first-50,000-row setting and prior target definition for a correction-only comparison. After filtering it has 44,290 rows: 26,574 training, 4,429 score, 4,429 risk, and 8,858 test observations. The training role spans November and December 2015; score, risk, and test all contain December 2015 observations. The split preserves issue-month order but leaves the order within each month unspecified. Ten fixed model seeds reuse this one membership split; their values are recorded in the supplementary configuration. Its label definition includes delinquent adverse statuses, unlike the resolved-cohort main experiment. This replay is not pooled with the mature-cohort study.

Home Credit uses the full labeled application table and available historical-table aggregates. Its application-level adverse-outcome target supports a count contract; the provided historical tables enrich features but do not supply the outcome-timed rolling protocol used for Lending Club. HELOC uses the supplied public risk-performance labels. Credit Approval is a small historical benchmark

Table 21: Exact total observations required for 80% exposure-test power. The necessary bound is a real-valued lower bound from Theorem 11; each feasible sample size is an integer.

a	Necessary bound	Hoeffding	Betting	Thinning
1	23.54	314	210	64
0.5	45.73	1,229	402	131
0.2	112.28	7,563	976	332
0.1	223.19	30,104	1,926	666
0.05	445.00	120,121	3,820	1,334
0.02	1,110.42	749,662	9,485	3,339
0.01	2,219.46	2,997,191	18,911	6,679

whose labels describe approval decisions rather than subsequent repayment defaults. We use this task to examine certification with few observations.

Each additional dataset uses five fixed repetitions of a shuffled 60/10/10/20 training/score/risk/test split and estimator randomness. These are joint split/model repetitions, not five model seeds on one fixed split. The certification sizes are 30,751, 1,046, and 69 for Home Credit, HELOC, and Credit Approval. Home caps are $\{.04, .05, .06\}$; HELOC and Credit Approval caps are $\{.20, .30, .40\}$. All use the four capacity values in the main protocol, $\delta = .05$, and constant utility. We exclude source-row metadata from model features. Older Home Credit archives identify rows by their deterministic position in the loaded table rather than the original application ID.

I.4 Utility definitions

Risk certification determines which policies are admissible; utility determines which certified policy is preferred. We use two prespecified proxies to distinguish a simple decision value from an amount-sensitive lending value. Constant utility awards $+1$ for an approved nondefault, -5 for an approved default, 0 for rejection, and $-.2$ for deferral. Its empirical value is $(N_A - 6K_A - .2K_D)/n$. A negative value is a loss under this specified proxy, even when a policy meets its statistical contract.

The Lending amount/interest proxy is normalized using training data only. Let c be the training 99th percentile of positive loan amount and r its training median. Define $\widetilde{L} = \min(L, c)/r$, annual interest i clipped to $[0, .5]$, and term T clipped to $[12, 60]$ months. Missing amount, interest, or term uses its training fallback. Then

$$u = \begin{cases} \widetilde{L}\, iT/12, & A = 1, Y = 0, \\ -.6\widetilde{L}, & A = 1, Y = 1, \\ -.01, & D = 1, \\ 0, & \text{reject.} \end{cases} \tag{107}$$

The utility proxy assumes a loss-given-default of .6 and a simple-interest return. It omits amortization, recoveries, prepayment, funding costs, and review accuracy, so it measures a specified sensitivity scenario rather than net profit. The training quantile c in utility is separate from the fixed exposure normalizer 40,000 in R_W. The main cohort's saved statistics use unclipped exposure weights. We report both calibration and held-out test utility and select policies solely with calibration utility.

J Controlled Power and Error Experiments

Before reintroducing candidate search, we examine the binary-loss lending model in controlled populations where the source of difficulty is known. Exact calculations quantify the data cost, and null simulations check the implemented tests at unsafe and boundary risks.

The exact-power study fixes $\rho = .5$, $\alpha_W = .1$, $\delta = .05$, $\beta = .2$, and $a \in \{1, .5, .2, .1, .05, .02, .01\}$. It computes the smallest integer sample size for power at least .8 using exact binomial survival functions. We find the required approval count from the full mixture's monotone log wealth on the zero-default family. We save the attained power and the power at the preceding integer to verify that the sample size is minimal on this family.

Table 22: Monte Carlo rejection of at least one true policy null, at $\delta = .05$. Entries are maxima across the prespecified boundary and mixed-null settings, with the corresponding pointwise 95% CP interval. Monte Carlo checks complement the proofs; the intervals do not adjust for selecting the maximum across settings.

Method	Replications per setting	Maximum rate	Pointwise interval
Bonferroni-IUT	20,000	0.04020	[0.03752, 0.04301]
Event-wise Bonf.	20,000	0.02635	[0.02417, 0.02866]
Fixed sequence	20,000	0.05005	[0.04707, 0.05316]
Holm-IUT	20,000	0.04020	[0.03752, 0.04301]

Table 23: General thinning power: rounded necessary size n_L, sufficient size n_U, and simulated power at each. Each entry uses 20,000 replications of the exact multinomial/binomial count laws.

μ	r	n_L	n_U	Power at n_L	Power at n_U
0.5	0	37	720	0.0002	1.0000
0.5	0.05	229	3,184	0.5582	1.0000
0.5	0.08	1,610	20,920	0.5974	1.0000
0.5	0.09	6,656	84,988	0.6127	1.0000
0.1	0	181	3,600	0.0068	1.0000
0.1	0.05	1,142	15,920	0.5533	1.0000
0.1	0.08	8,048	104,600	0.5989	1.0000
0.1	0.09	33,277	424,940	0.6058	1.0000
0.02	0	905	18,000	0.0111	1.0000
0.02	0.05	5,706	79,600	0.5541	1.0000
0.02	0.08	40,236	523,000	0.6008	1.0000
0.02	0.09	166,384	2,124,700	0.6147	1.0000

We additionally draw 20,000 independent samples per setting from the contamination family, with $a \in \{1, .1, .01\}$, $n \in \{100, 1000, 10000\}$, and target exposure risks $q \in \{0, \alpha_W, 2\alpha_W\}$. The three outcome types are low-exposure good approvals, unit-exposure defaulting approvals, and rejections. Their multinomial counts exactly generate the required sufficient statistics; low-exposure approvals are then binomially thinned. The experiments test the complete count/exposure/capacity contract at one-policy level $\delta = .05$, with $\alpha_A = B = .1$. We include count-only testing to show how a certificate for another target can miss exposure violations. We save every raw count vector, p-value, and pass indicator.

The separate multiple-testing stress study generates exact binomial marginals coupled by a Gaussian copula, with $M \in \{20, 100\}$ and latent correlation in $\{0, .8\}$. It uses 1,000 approval trials, 2,000 capacity trials, and 500 independent development trials for each component. The four settings are all approval nulls at the boundary, all safe, half unsafe in approval, and half unsafe in capacity. Safe approval/capacity probabilities are .065/.04; the unsafe half uses .13 in its affected component. There are 20,000 replications per setting. This construction isolates dependence among valid binomial tests; it does not model the full joint distribution of overlapping credit policies.

J.1 Nonzero risk and exposure-mass complexity

The zero-default experiment isolates exposure mass, but practical policies may operate close to the risk cap. We therefore vary both exposure mass and the risk margin to check the full scaling predicted by Theorem 13. We use its explicit hard family with $\rho = .5$, $\alpha = .1$, $\delta = .05$, and target power $1 - \beta = .8$. We cross $\mu \in \{.5, .1, .02\}$ with $r \in \{0, .05, .08, .09\}$ and simulate 20,000 independent count samples both at the rounded necessary bound and at the sufficient bound. The necessary and sufficient bounds define the two reference sample sizes. We report all 24 settings. Table 23 shows inverse-μ scaling and the growing sample cost as the risk margin decreases. At every sufficient size, all simulations pass; the pointwise 95% Clopper–Pearson lower bound is .9998. The necessary and sufficient sizes have matching rates but differ substantially at finite sample sizes.

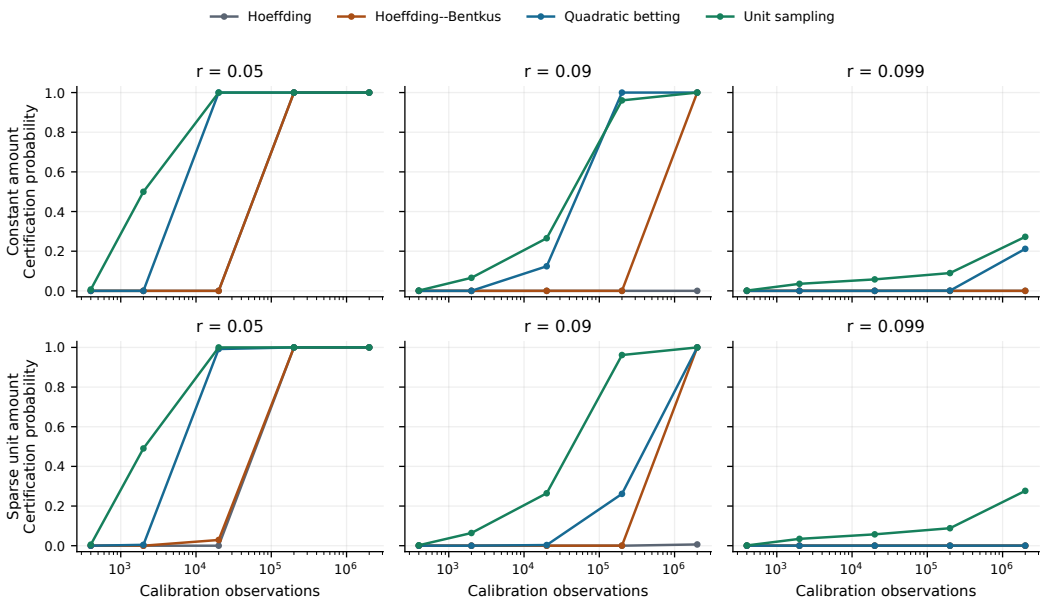

Figure 11: Fixed-policy mechanism comparison at $\rho = .5$, $a = .1$, $\mu = .05$, and $\alpha = .1$. Top: constant exposure; bottom: sparse unit exposure, with ten times the second moment at the same risk. Each point uses 20,000 simulations. Thinning has the same theoretical power in the two rows; finite simulation draws are independent. The fixed quadratic betting grid has no positive-drift member in the bottom-right population. We report each of the four tests separately.

J.2 Fixed-policy tests at equal mass and risk

The previous diagnostic varies the two quantities that govern thinning. Direct tests can additionally depend on the distribution of exposure, so we next hold mass and risk fixed while changing the second moment. To do so, consider two populations with the same ρ, $\mu = \rho a$, and $R_W = r$. In the constant-exposure population every approved application has $w = a$ and label $\mathrm{Bernoulli}(r)$. In the sparse-unit population, approved unit-weight applications have mass ρa and the same label distribution; zero-weight, non-defaulting approvals have mass $\rho(1 - a)$. The remaining mass is rejected in both cases. Writing $b(r) = (1 - r)\alpha^2 + r(1 - \alpha)^2$, their second moments are

$$v_{\text{constant}} = \rho a^2 b(r), \qquad v_{\text{sparse}} = \rho a b(r). \tag{108}$$

Thinning therefore has the same power in both populations, whereas the quadratic betting wealth differs in drift and variability.

The frozen diagnostic crosses two populations, three amounts $a \in \{1, .1, .01\}$, six risks $r \in \{0, .05, .09, .099, .10, .12\}$, and five expected retained counts $n\mu \in \{20, 100, 1000, 10000, 100000\}$, with $\rho = .5$, $\alpha = .1$, $\delta = .05$, and 20,000 replications per setting. All four tests use the same generated counts within each setting, with the model, policy, and independent order held fixed. The 180 settings include the boundary null and strictly unsafe risk, alongside the safe cases.

At $r = .09, n = 200,000$, quadratic betting power is 1.0000 under constant exposure and .2614 under sparse-unit exposure; thinning gives .9606 and .9617. This difference isolates the second moment at equal exposure mass and risk. HB has zero observed passes in the former setting and .00035 in the latter; at $n = 2,000,000$ its powers are 1.0000 and .99995. At $r = .099$, every quadratic betting grid component has negative drift in the sparse-unit family, illustrating the failure mechanism proved in Appendix G.8. The same-grid product has positive log growth at that risk, so this failure belongs to the quadratic approximation. The relative performance of the tests therefore depends on the population. The maximum observed boundary/unsafe rejection rate for thinning is .0514; its pointwise interval includes .05. These simulations check the implementation alongside the validity proofs.

Table 24: Full HB power on P_0 at the exact sample sizes where quadratic betting or thinning first reaches 80% power. Other parameters match Figure 10. HB can outperform quadratic betting at larger exposure and require far more evidence at smaller exposure.

a	n_E	HB power at n_E	n_T	HB power at n_T
1	210	1.0000	64	0.8091
0.5	402	0.9978	131	$< 10^{-10}$
0.2	976	$< 10^{-10}$	332	$< 10^{-10}$
0.1	1,926	$< 10^{-10}$	666	$< 10^{-10}$
0.05	3,820	$< 10^{-10}$	1,334	$< 10^{-10}$
0.02	9,485	$< 10^{-10}$	3,339	$< 10^{-10}$
0.01	18,911	$< 10^{-10}$	6,679	$< 10^{-10}$

Table 24 adds the complete HB baseline to the sample sizes in Figure 10. We evaluate its actual rejection indicator over the binomial approval-count law, including the five-count guard. The omitted tail probability is below 2.1×10^{-13} in every row. The ceiling in the Bentkus branch can make the set of rejection counts nonmonotone. We therefore sum its rejection indicator directly rather than assume a single count threshold.

K Extended Empirical Results

The main lending table combines several design choices. The results below separate the effects of the certification test, candidate order, multiplicity correction, and calendar transfer by holding the remaining pieces fixed in turn.

K.1 Every selection procedure at the reference contract

We begin with the broadest comparison: every selection rule is applied to the same candidates and reference contract. This establishes which differences arise before any component of the design is held fixed. Table 25 separates the effects of the single-policy test and the selection procedure. Empty outputs remain in the return-count denominator; utility, risk, and empirical exceedances are defined only for runs returning a policy. The machine-readable tables retain the full risk/capacity grid, both utility definitions, and every model seed. Repeated methods and caps on one test cohort are not independent statistical replications.

K.2 Crossing fixed development orders and certification tests

The preceding table changes the test and the search rule together. To separate those effects, we cross each certification test with each frozen development order on the same runs. We freeze the original-risk/negative-utility Pareto frontier for each of the fifteen same-quarter runs at $\alpha_A = \alpha_W = .09$, $B = .10$, $\delta = .05$, and amount/interest utility. Using only development data, we compute three orders from HB, betting, and thinning maximum component p-values. We evaluate each order with all three certification tests, giving 135 records, and save the orders before loading certification statistics. All comparisons use the same models, candidates, thinning draws, and evaluation observations.

All three tests return 14/15 under the HB order and 15/15 under either other order. The order therefore explains the return-count difference in this setting. Under the fixed HB, betting, and thinning orders, respectively, betting improves test utility over HB by .00748, .00912, and .00975, averaging the five quarter-level paired means. Thinning further improves over betting by .00520, .00445, and .00488. We report common empty outputs and unpaired returns separately. The fifteen model runs share five quarters, which remain the units for temporal comparisons.

K.3 Calendar effects at fixed calibration size

The fixed-order comparison removes search as an explanation. We next hold the calibration size and test set fixed to examine whether using the same or previous calendar quarter changes the selected policies.

Table 25: All selection procedures on the same-quarter reference, $\alpha_A = \alpha_W = .09$, $B = .10$, $\delta = .05$, and amount/interest utility. "Ex." counts returned runs whose empirical held-out contract exceeds a cap, not a population FWER estimate. The count row omits the exposure constraint. Risk and utility means are conditional on return.

Test	Selection	Return	Utility	$\widehat{R}_A$	$\widehat{R}_W$	$\widehat{R}_D$	Ex.
Count only	Event-wise Bonf.	15/15	0.1276	0.0794	0.0768	0.0572	0
Count only	Bonferroni-IUT	15/15	0.1281	0.0797	0.0771	0.0561	0
Count only	Holm-IUT	15/15	0.1281	0.0797	0.0771	0.0561	0
Count only	Fixed sequence	15/15	0.1326	0.0816	0.0788	0.0629	0
Count only	Utility-first FS	0/15	–	–	–	–	–
Count only	Pareto Testing	15/15	0.1362	0.0837	0.0808	0.0521	0
Hoeffding	Event-wise Bonf.	0/15	–	–	–	–	–
Hoeffding	Bonferroni-IUT	0/15	–	–	–	–	–
Hoeffding	Holm-IUT	0/15	–	–	–	–	–
Hoeffding	Fixed sequence	6/15	0.0869	0.0580	0.0566	0.0650	0
Hoeffding	Utility-first FS	0/15	–	–	–	–	–
Hoeffding	Pareto Testing	6/15	0.0918	0.0599	0.0585	0.0706	0
Hoeffding–Bentkus	Event-wise Bonf.	6/15	0.0774	0.0543	0.0531	0.0702	0
Hoeffding–Bentkus	Bonferroni-IUT	6/15	0.0834	0.0565	0.0547	0.0542	0
Hoeffding–Bentkus	Holm-IUT	6/15	0.0834	0.0565	0.0547	0.0542	0
Hoeffding–Bentkus	Fixed sequence	14/15	0.1134	0.0720	0.0698	0.0551	0
Hoeffding–Bentkus	Utility-first FS	0/15	–	–	–	–	–
Hoeffding–Bentkus	Pareto Testing	14/15	0.1175	0.0740	0.0718	0.0550	0
Quadratic betting	Event-wise Bonf.	13/15	0.1100	0.0716	0.0689	0.0576	0
Quadratic betting	Bonferroni-IUT	15/15	0.1118	0.0718	0.0693	0.0593	0
Quadratic betting	Holm-IUT	15/15	0.1118	0.0718	0.0693	0.0593	0
Quadratic betting	Fixed sequence	15/15	0.1243	0.0780	0.0750	0.0544	0
Quadratic betting	Utility-first FS	0/15	–	–	–	–	–
Quadratic betting	Pareto Testing	15/15	0.1279	0.0797	0.0768	0.0555	0
Thinning	Event-wise Bonf.	15/15	0.1110	0.0717	0.0693	0.0556	0
Thinning	Bonferroni-IUT	15/15	0.1130	0.0729	0.0705	0.0660	0
Thinning	Holm-IUT	15/15	0.1130	0.0729	0.0705	0.0660	0
Thinning	Fixed sequence	15/15	0.1283	0.0795	0.0768	0.0632	0
Thinning	Utility-first FS	0/15	–	–	–	–	–
Thinning	Pareto Testing	15/15	0.1313	0.0814	0.0783	0.0649	0

Table 26: The nine order/test crossings. Utility first averages the returned model runs within each quarter, then weights the five quarters equally. It can therefore differ from the pooled conditional mean in Table 18 when one quarter has a missing return. No crossing has an empirical contract exceedance.

Development order	Certification test	Return	Cal. utility	Test utility
betting	betting	15/15	0.1278	0.1279
betting	hb	15/15	0.1183	0.1187
betting	thinning	15/15	0.1322	0.1323
hb	betting	14/15	0.1242	0.1243
hb	hb	14/15	0.1167	0.1168
hb	thinning	14/15	0.1293	0.1295
thinning	betting	15/15	0.1262	0.1264
thinning	hb	15/15	0.1162	0.1167
thinning	thinning	15/15	0.1311	0.1313

Table 27: The paired study by quarter, using betting and Pareto selection at the same contract as Figure 12. Means are over returned model runs; "Ex." is an empirical contract-exceedance count.

Quarter	Calibration	Return	Utility	$\widehat{R}_A$	$\widehat{R}_W$	Ex.
2014 Q4	Same quarter	3/3	0.1384	0.0826	0.0804	0
2014 Q4	Previous quarter	3/3	0.1488	0.0868	0.0842	0
2015 Q1	Same quarter	3/3	0.1263	0.0803	0.0761	0
2015 Q1	Previous quarter	3/3	0.1369	0.0840	0.0803	0
2015 Q2	Same quarter	3/3	0.1166	0.0820	0.0790	0
2015 Q2	Previous quarter	3/3	0.1166	0.0820	0.0790	0
2015 Q3	Same quarter	3/3	0.1427	0.0798	0.0777	0
2015 Q3	Previous quarter	3/3	0.1248	0.0716	0.0700	0
2015 Q4	Same quarter	3/3	0.1154	0.0739	0.0708	0
2015 Q4	Previous quarter	3/3	0.1378	0.0844	0.0813	0

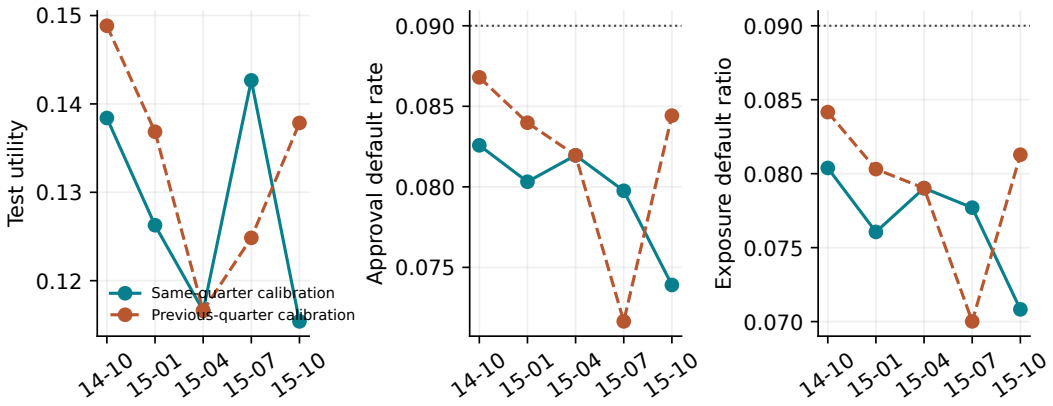

Figure 12: Quarter-level means over three model seeds for the betting exposure test with Pareto selection, $\alpha_A = \alpha_W = .09$ and $B = .10$. Each pair uses identical held-out test applications and equally many calibration applications. Dotted lines denote the risk caps. These five calendar cohorts, rather than fifteen model runs, are the distinct temporal evaluation units.

K.4 Effects of multiple-testing correction

We then return to the historical replay to isolate a different design choice: how multiplicity is allocated within and across policies. Because the candidates and data split are shared, the comparison focuses on the correction itself.

Table 28: Returned policies in ten model-seed runs on the one historical Lending split, with count contract, $B = .10$, $\delta = .05$, and constant utility. The IUT column removes the within-policy component correction; Holm and Bonferroni-IUT have the same nonempty event.

α_A	Event Bonferroni	Bonferroni-IUT	Holm-IUT	Pareto
0.08	0/10	0/10	0/10	8/10
0.09	0/10	5/10	5/10	9/10
0.10	7/10	8/10	8/10	10/10
0.12	10/10	10/10	10/10	10/10

The full candidate count varies from 5,514 to 6,919 across these model seeds, whereas the mature-cohort families range from 12,717 to 19,802. Table 28 isolates the correction applied within each policy. The improvement over event-wise Bonferroni comes from this change, while Holm and Bonferroni-IUT have the same nonempty event. The complete rejected sets show whether Holm certifies additional policies or changes the policy with the highest calibration utility. The paired raw tables condition selected-policy agreement on both methods returning one, separately recording common empty outputs.

K.5 Additional datasets and small-sample failures

Finally, we move beyond the main lending cohort to test whether the same selection patterns appear with different data sizes and targets. The small benchmarks are especially useful for showing how an independently ordered sequence can return a policy when a large simultaneous family cannot.

A policy can satisfy the risk contract and still have negative utility. The HELOC results illustrate this distinction under the constant-utility proxy. A contract that also requires positive population utility would need a valid lower confidence bound on that quantity. The occasional returns from 69 calibration observations reflect the allocation of the error budget. A fixed sequence tests a short, independently ordered prefix at a larger local level, whereas Bonferroni allocates a smaller level to each policy in a large family.

Table 29: Count-contract comparisons over five joint split/model repetitions. Constant utility is used. "Ex." counts empirical count/capacity exceedances among returned runs. Small-sample returns from independent ordering are retained even when a simultaneous Bonferroni family returns none.

Dataset	Selection	α_A	B	Return	Utility	$\widehat{R}_A$	$\widehat{R}_D$	Ex.
Home Credit	Bonferroni-IUT	0.05	0.05	5/5	0.4941	0.0349	0.0303	0
Home Credit	Holm-IUT	0.05	0.05	5/5	0.4941	0.0349	0.0303	0
Home Credit	Pareto Testing	0.05	0.05	5/5	0.4981	0.0347	0.0351	0
HELOC	Bonferroni-IUT	0.30	0.10	0/5	–	–	–	–
HELOC	Holm-IUT	0.30	0.10	0/5	–	–	–	–
HELOC	Pareto Testing	0.30	0.10	4/5	-0.0999	0.2116	0.0435	0
Credit Approval	Bonferroni-IUT	0.30	0.20	0/5	–	–	–	–
Credit Approval	Holm-IUT	0.30	0.20	0/5	–	–	–	–
Credit Approval	Pareto Testing	0.30	0.20	4/5	0.0707	0.1322	0.0362	0

L Reproducibility and Artifact Organization

We close the appendix by connecting the statistical workflow to the saved artifacts. The code follows the same separation between model fitting, policy development, and certification as the analysis, and the records below allow each reported result to be traced back to its inputs.

The experiment entry point is `scripts/run_contract_comparison.py` with `configs/review_revision.json`. It saves configurations, environment/package versions, source hashes, fitted estimator parameters, row membership, prediction arrays, candidate definitions, score/risk/test sufficient-statistic tables, component p-values, development orders, complete rejection masks, and selected-policy results. The runner writes a completion marker after all requested runs finish. The summary scripts verify that each record belongs to a completed experiment and each selected identifier belongs to its certified set.

The mature-cohort, historical, and additional-data artifacts are separate from `results_review_within_cohort` and `results_review_matched_temporal`. The latter two derive from frozen predictions through `scripts/run_cohort_reference.py` and `configs/review_cohort_reference.json`; no model is refitted. HB is added by `scripts/run_hb_comparison.py`, also without refitting. It preserves the original component count/capacity tests, reuses original-risk frontier membership, and recomputes the development order with HB p-values. Each derived run links its input files by source hash. We specified the reference and HB comparisons after inspecting the first rolling results and report them as subsequent diagnostics.

Exact-power and null experiments use `scripts/simulate_contract_certification.py` with `configs/review_theory.json`. The analysis entry point, `scripts/analyze_contract_revision.py`, generates the paper tables and figures from the completed archives and records their input hashes. It separates calibration from test utility, validates boolean fields explicitly, and reports missing utility for empty outputs rather than converting them to zero. The full grid is in `results_review_summary/summary.csv`; per-window and per-run versions are adjacent. Paired effects use the same run identifiers and are evaluated only where both compared procedures return a policy.

The later margin, moment, and fixed-order diagnostics use the `review_margin.json`, `review_mechanisms.json`, and `review_fixed_order.json` configurations. Their environment manifests identify the commands, source hashes, and package versions. The diagnostics save every count replication or fixed-order rejection array, and a separate table builder records input hashes. We designed these additional diagnostics in response to review of the manuscript.

The anonymous supplement contains code only. From its root directory, `configure.py` generates the frozen configurations used by the runners. The source code includes no datasets, prediction caches, model weights, or experimental output. Its documentation gives the commands for generating configurations, running tests and numerical checks, and launching the linear and nonlinear multilabel experiments. Input data must occupy the paths recorded in each configuration; the loader verifies the source files before fitting. Completed output directories are immutable. We also check archived predictions for finite scores, valid probability ranges, ordered band edges, and score variation, and retain numerical warnings with the run records.

Reproduction requires the input archives, preprocessing rules, policy definitions, normalization, randomization settings, software environment, and complete result records. We preserve source snapshots from before the guard corrections so that each result remains linked to the code that produced it.