Evaluation Under Context-Dependent Value

Coverage, Scarce Confirmation, and the Screening–Confirmation Firewall

Abstract

Evaluation systems often screen candidates before the conditions in which those candidates will succeed are known. When screening and downstream matching are separated, channel-marginal scoring can reject candidates whose value would become high after later information reveals where they fit. This paper studies the design of such staged systems. Blackwell's comparison of experiments implies that more informative downstream signals weakly increase both attainable continuation value and the welfare hidden by an upstream marginal screen. A *coverage frontier* then links prediction to search: approximate niche rankings suffice, with at most $2k\epsilon$ coverage regret under uniform weight error and exact recovery when the top-k boundary is separated. With scarce confirmation, however, reusing the original plausibility signal can make a suspected blind spot operate twice. We define a *screening–confirmation firewall* that limits this reuse and derive an ex ante sensitivity condition separating score error from the welfare cost of missed candidates. Partial randomized confirmation provides the stronger ex post result: with known inclusion probabilities, it identifies the realized firewall-versus-trusted policy contrast conditional on the queue actually formed. Randomization therefore supplies both robustness and institutional self-measurement, but only under adequate precision and outcome measurement. We show that zero differential leakage is insufficient: policy learning on a proxy welfare measure also requires approximately affine measurement, and upstream signal-dependent queue formation can attenuate the measured benefit of the firewall. A firewall intensity λ makes robustness, learning, and efficiency one policy continuum; maintaining $\lambda > 0$ can serve as an ongoing audit when evaluator performance drifts.

1. Introduction

Institutions rarely evaluate candidates in the conditions under which those candidates will ultimately succeed or fail. A grant panel sees a proposal rather than every experimental setting in which it might prove useful. A hiring committee sees a person before knowing which team or problem will make that person unusually valuable. A journal reviewer sees a manuscript before observing which later research programs it may unlock. A model evaluator observes behavior on a finite benchmark rather than across every deployment context.

The conventional response is to estimate expected value under a distribution of plausible conditions. This is often exactly right. If realization conditions are assigned independently of anything learned after screening, channel-marginal expected value is the relevant quantity. But many institutions are organizationally staged. One actor screens a candidate using one information set; conditional on surviving that screen, another actor later acquires information about fit, searches among realization contexts, and confirms apparent successes. In such systems, an upstream screen can destroy an option that a downstream matcher could have used productively.

This paper develops a minimal framework for that problem. It deliberately avoids a general theory of novelty, creativity, or intelligence. The central objects are candidates, realization channels, information held at different institutional stages, exploration breadth, and scarce confirmation capacity.

The first result is a two-stage restatement of Blackwell's comparison of experiments (Blackwell, 1953). An upstream evaluator screens using information available at admission. A downstream matcher, conditional on admission, may receive an additional signal about candidate–channel fit. If one downstream signal Blackwell-dominates another, the value attainable after admission weakly increases. The resulting *matching blind spot* is therefore not generated by a decision maker who knowingly discards information. It is generated by organizational separation, scoring mandates, bounded continuation-value calculation, or information that becomes available only after admission.

These sources of the blind spot imply different first-best remedies. If the screener simply lacks information that can cheaply be moved upstream, provide it. If the difficulty is computational, improve continuation-value estimation. If a mandate requires channel-marginal scoring, change the mandate. The firewall studied later is not a substitute for those remedies. It addresses a different problem: what to do when the original plausibility signal is deliberately bypassed by exploration and would otherwise be reused at the scarce confirmation stage.

The second result concerns implementation. Prediction and optionality act on one object: a *coverage frontier*. Given a signal, realization channels are ordered by the probability that they contain a productive niche; optionality determines which channels are available to search. Exact probabilities are unnecessary. If estimated niche weights differ from their true values by at most ϵ, using the estimated top k channels sacrifices at most $2k\epsilon$ coverage. A gap-dependent corollary is stronger: if the true top-k boundary exceeds 2ϵ, the selected set is exactly correct, and more generally only near-ties contribute to regret.

The third result concerns scarce confirmation. Exploration and confirmation should be separate stages. Exploratory hits can be queued while search continues, rather than allowing a noisy first hit both to terminate exploration and trigger promotion. Once those stages are separated, breadth is not intrinsically pathological. What remains is multiplicity and scarcity: broader exploration creates more true and false hits, all of which compete for a finite confirmatory budget.

The natural allocation rule is to use the original evaluator's plausibility score as a prior and confirm the hits most likely to be genuine. If the evaluator is correctly calibrated, this is efficient. If the same evaluator is systematically pessimistic on precisely the candidates exploration was intended to rescue, however, its reuse at confirmation can make the original screen operate twice. We therefore define a *screening–confirmation firewall*: conditional on entry into the confirmation queue, the original candidate-level plausibility signal cannot determine either access to confirmation or the confirmatory evidentiary threshold.

A firewall is not free. Under a well-specified evaluator it discards useful information. We therefore derive an ex ante misspecification-sensitivity condition rather than a universal recommendation. The actual leverage comes from the decision boundary: small errors in score near a hard cutoff can reverse selection even when the downstream welfare consequence of the reversal is arbitrarily large. Score error and welfare error need not share a scale.

The paper's strongest identification result follows from asking how an institution could learn this efficiency cost. Under deterministic status-quo selection, outcomes for rejected candidates are missing by design. Partial randomized confirmation changes that. With known positive inclusion probabilities, the institution can estimate the realized value difference between a randomized firewall policy and the trusted rule directly. The ex ante decomposition into a calibrated baseline and hidden gains remains

useful before randomization or for extrapolation to new queues, but it is not required to determine which policy performed better in the randomized queue itself.

This makes randomization institutional self-measurement rather than pure sacrifice. Research funding provides concrete precedents. The Health Research Council of New Zealand has used a modified lottery for Explorer Grants since 2013, creating randomized variation that has subsequently been used to estimate the effect of funding on researcher outputs (Liu et al., 2020; Barnett et al., 2024). The Volkswagen Foundation has used partial randomization in its *Experiment!* program since 2017 and has explicitly evaluated the procedure as an alternative to finer-grained jury ranking among eligible proposals (VolkswagenStiftung, 2024).

The contribution is thus not another empirical claim that unconventional candidates are undervalued. Evidence on novelty and evaluation is heterogeneous (Boudreau et al., 2016; Ayoubi et al., 2021; Teplitskiy et al., 2022). The paper instead asks a design question: if an institution regards context-dependent value and evaluator misspecification as plausible, how should screening, exploration, confirmation, and randomization be arranged so that the same evaluator does not silently erase the option twice, and how can the institution learn whether that protection is worth its cost?

2. A Two-Stage Setup: Screening, Matching, and the Blind Spot

Let $z \in Z$ denote a candidate and $r \in R$ a *realization channel*: a context in which the candidate can be tried or deployed. A channel may represent an experimental setting, team, market, scientific subfield, task, organizational role, or deployment condition. Let $\omega \in \Omega$ be the latent state governing performance across channels, and let

$$m(z, r, \omega) \in R$$

be the payoff from realizing candidate z in channel r under state ω.

The institution has two stages. An upstream *screener* E observes information X and decides whether the candidate advances. A downstream *matcher* M, conditional on advancement, may observe an additional signal S about ω and candidate–channel fit.

Suppose the screener evaluates the candidate under a baseline channel distribution v using only X:

$$V_E(z \mid X) = E_{\omega \mid X} E_{r \sim v}[m(z, r, \omega)).$$

Equation (1) is the relevant quantity when the screener's mandate is channel-marginal evaluation or when downstream matching possibilities are not available to it at the admission decision.

Conditional on admission, the matcher observes $S = s$ and may choose a channel distribution q from a feasible set Q. Define downstream attainable value as

$$V_M(z \mid X, S) = E_s\left[\max_{q \in Q} E_{\omega \mid X, s} E_{r \sim q}[m(z, r, \omega))\right).$$

The feasible set may encode single-channel assignment, a limited search budget, capacity constraints, or other restrictions. The matcher is free to ignore S.

For a screening threshold τ, define the *matching blind-spot set*

$$B_S(\tau) = \{z : V_E(z \mid X) < \tau \le V_M(z \mid X, S)\}.$$

These are candidates rejected by the upstream screen even though downstream information and context-sensitive matching would make them worth retaining.

The two-stage structure is essential. If one integrated decision maker knew the downstream information structure, could value its option value correctly, and faced no institutional scoring constraint, it should incorporate that continuation value at screening. The blind spot instead arises when information or authority is separated across stages, when the downstream signal is generated only after admission, when continuation value is difficult to compute, or when the screening rule is constrained to score candidates on a channel-marginal basis.

Proposition 1. Blackwell monotonicity of continuation value and blind-spot welfare

Suppose downstream signal S_2 Blackwell-dominates S_1: S_1 can be generated from S_2 by a stochastic garbling. If both signals have the same acquisition cost and feasible downstream action set, then for every candidate z,

$$V_M(z \mid X, S_2) \geq V_M(z \mid X, S_1).$$

Consequently,

$$B_{S_1}(\tau) \subseteq B_{S_2}(\tau).$$

Moreover, for any nonnegative measure μ over candidates, define

$$g_S(z;\tau) = 1\{V_E(z \mid X) < \tau\} \max\{V_M(z \mid X, S) - \tau, 0\}.$$

The *blind-spot welfare mass* is then

$$L_S(\tau) = E_{z\sim\mu}[g_S(z;\tau)].$$

Then

$$L_{S_2}(\tau) \geq L_{S_1}(\tau).$$

Proof sketch

Any downstream decision rule implementable after observing S_1 can be replicated after observing S_2 by first garbling S_2 into S_1 and then following the same rule. The feasible set after S_2 therefore contains a strategy achieving the value under S_1, so its optimum cannot be lower. The upstream screening value and threshold are unchanged, which yields (5). Pointwise monotonicity of V_M implies pointwise monotonicity of the nonnegative integrand in (6), giving (7). ▫

Proposition 1 is a direct consequence of Blackwell (1953), not a new information-ordering theorem. Its contribution is institutional: downstream information that increases realizable continuation value can be rendered irrelevant by an earlier stage that never admits the candidate.

3. The Coverage Frontier

Equation (2) is general but too abstract for institutional design. A useful special case is a candidate with one productive niche. Conditional on the candidate being genuine and on signal S, let

$$w_r(S) = P(r^\star = r \mid H, S)$$

be the posterior probability that channel r is the productive niche, where H denotes the event that such a niche exists.

Suppose only a subset $R_b \subseteq R$ is available, where b indexes investment in maintaining realization options. With equal exploration costs, define the *coverage frontier*

$$q_{b,S}(k) = \max_{\substack{A \subseteq R_b \\ |A| \le k}} \sum_{r \in A} w_r(S).$$

$q_{b,S}(k)$ is the maximum posterior probability that the productive niche lies among the first k channels the institution can search.

Prediction and optionality operate on the same frontier. Better prediction changes the ordering and concentration of $w_r(S)$, moving probability mass toward earlier ranks. Greater optionality changes R_b, the channels available at all, and can raise the frontier's ceiling.

Institutions rarely know the true $w_r(S)$. The useful question is therefore how much is lost under approximate ordering.

Proposition 2. Coverage regret from approximate channel ordering

Let w_r be the true posterior niche weights and $\hat{w}_r$ estimated weights satisfying

$$\| \hat{w} - w \|_\infty \le \epsilon.$$

Let $A_k^\star$ be the true top-k set and $\hat{A}_k$ the top-k set under $\hat{w}$. Then

$$\sum_{r \in A_k^\star} w_r - \sum_{r \in \hat{A}_k} w_r \le 2k\epsilon.$$

Proof

Add and subtract the estimated top-k sums for $A_k^\star$ and $\hat{A}_k$. The difference decomposes into two estimation-error terms and one ranking term. The ranking term is nonpositive because $\hat{A}_k$ maximizes the estimated top-k sum. Each estimation-error term is at most $k\epsilon$ in absolute value, giving total regret at most $2k\epsilon$. ▫

Corollary 2.1. Gap-dependent stability

Let $w_{(1)} \ge \cdots \ge w_{(R)}$ denote the ordered true weights. If

$$w_{(k)} - w_{(k+1)} > 2\epsilon,$$

then

$$\hat{A}_k = A_k^\star$$

and coverage regret is zero.

More generally, if the estimated top-k set differs from the true top-k set in exactly d memberships, then

$$q^{\star}(k) - q^{\hat{w}}(k) \le 2d\epsilon.$$

Only channels within a 2ϵ neighborhood of the selection boundary can switch membership. The corollary is therefore often more useful than the generic $2k\epsilon$ bound: approximate ordering is exact when the boundary is well separated, and errors are charged only to near-ties.

Worked example: a crude grant-panel ordering

Suppose a funding program has twenty plausible pilot contexts for a proposal but can initially explore only four. Domain experts produce a rough posterior ordering. If a validation exercise supports $\epsilon = 0.015$, then the generic bound limits coverage loss to

$$2(4)(0.015) = 0.12.$$

If the estimated fourth- and fifth-ranked contexts are separated by more than 0.03, the top-four set is stable under the same error bound. The panel does not need precise posterior probabilities. A coarse but approximately correct ordering can capture most or all of the available matching value.

For unequal exploration costs, (8) becomes a budgeted coverage problem rather than a top-k problem. The equal-cost formulation is retained because it isolates the paper's design point.

4. Exploration and Confirmation Are Different Stages

The coverage frontier describes where to look. It does not determine what evidence should count as success.

A common but problematic architecture is

$$\text{explore} \rightarrow \text{first apparent hit} \rightarrow \text{stop and promote}.$$

Under this rule, noise can both terminate exploration and trigger promotion. The pathology belongs to the architecture, not to breadth itself. A better design is

$$\text{explore} \rightarrow \text{queue apparent hits} \rightarrow \text{confirm independently},$$

while exploration continues according to the search policy.

Once stages are separated, exploratory false positives need not prevent genuine niches from being reached. What remains is multiplicity and queue scarcity. Let E_r denote the event that channel r ultimately produces a false *promotion* after the complete screening-and-confirmation pipeline, and suppose

$$P(E_r) \le \acute{\beta}$$

for each of k explored channels. Without assumptions about dependence across channels,

$$P\left(\bigcup_{r=1}^{k} E_r\right) \le k\acute{\beta}$$

by the union bound. To keep the probability of at least one false promotion below δ, it is sufficient that

$$\acute{\beta} \le \frac{\delta}{k}.$$

The bound is a worst-case guarantee. Its attraction is that it does not require cross-channel independence. If confirmatory error falls exponentially with accumulated log-evidence, $\beta(t) \approx e^{-ct}$, then (13) implies evidence requirements growing only logarithmically with breadth.

The more consequential dependence problem often lies *within* the screening-and-confirmation pair. If both stages reuse the same flawed dataset, reviewer misconception, measurement protocol, or model artifact, the combined false-promotion probability can be much larger than a naive product of nominal stage-specific error rates. This creates an independent rationale for the firewall introduced next: removing the original evaluator signal from confirmation is a direct decorrelation step. It does not by itself create independent confirmation if other data or methods remain shared.

5. Scarce Confirmation and the Screening–Confirmation Firewall

If confirmation were unlimited, every exploratory hit could be investigated and candidate-level plausibility would play little role in rationing follow-up. Real institutions have finite confirmation capacity. Suppose exploration yields N candidate hits but only $K < N$ can receive confirmation.

Let S_i^{ext} denote the original candidate-level plausibility signal, E_i post-entry exploratory evidence, and X_i structural information such as domain, cost, safety tier, or stage. Let the trusted rule have confirmation probabilities t_i, with $\sum_i t_i = K$; in the simplest deterministic case, $t_i \in \{0,1\}$.

A *full firewall* is an allocation rule with confirmation probabilities f_i satisfying

$$\sum_i f_i = K$$

and whose allocation and confirmatory thresholds depend on E_i, allowed structural variables X_i, and randomization ξ, but not on S_i^{ext} or its evaluative proxies.

For the value accounting below, a candidate-specific confirmatory threshold can be folded into f_i as an *effective confirmation probability*: raising the evidentiary threshold lowers the probability that a queued candidate reaches confirmed promotion and outcome observation. Thus the same notation covers both rationing of confirmation slots and plausibility-dependent confirmatory thresholds, provided the induced effective probabilities are known or randomized by design. This abstraction does not assume the two mechanisms have identical error profiles; it only places their effect on expected policy value on the same probability scale.

The restriction is stage-local. It does not require the institution to ignore priors when deciding where or how deeply to explore. A low-plausibility candidate can therefore receive shallow exploration and never reach the queue. The firewall protects the confirmation stage, not the candidate end to end.

5.1 Firewall intensity

Practical systems need not jump discontinuously from full trust to full exclusion. Define an *allocation-firewall intensity* $\lambda \in [0,1)$ by

$$p_i(\lambda) = (1-\lambda)t_i + \lambda f_i.$$

At $\lambda = 0$, confirmation follows the trusted rule. At $\lambda = 1$, allocation is fully firewalled from S^{ext}. Intermediate values deliberately retain some influence of the original plausibility signal while assigning

positive probability through the firewall component. The confirmatory evidentiary threshold can still be fully firewalled once a candidate is selected for confirmation.

If $f_i > 0$ for every candidate in the declared eligible queue, then any $\lambda > 0$ gives every candidate positive confirmation probability. This makes partial randomization an explicit point on the same policy axis as the full firewall rather than an exception introduced only for identification.

5.2 Reference classes must be firewall-safe in effect

Class-specific confirmation rules may be justified by structural differences in cost, risk, or measurement quality. But the class system can recreate the forbidden plausibility signal at a coarser resolution.

Let $C = g(X^{struct})$ be the reference class and let $p(C) \in [p_{min}, p_{max})$ denote the resulting confirmation propensity. Define *class leakage*

$$\ell_C = \max_{s,s'} TV\left(P(C \mid S^{ext} = s), P(C \mid S^{ext} = s')\right).$$

Then

$$\left|E[p(C) \mid S^{ext} = s) - E[p(C) \mid S^{ext} = s')\right| \leq \ell_C (p_{max} - p_{min}).$$

Firewall safety is therefore an auditable property, not merely an intention. Structural classes should be preregistered from variables such as domain, stage, cost, or safety rather than reviewer plausibility or obvious proxies, and institutions should report how strongly the resulting partition predicts the signal it is supposed to exclude.

5.3 Scarcity is why the firewall is costly

With $K = N$, all queued hits can be confirmed and little is gained by using candidate-level plausibility to ration follow-up. With $K \ll N$, the plausibility signal is valuable precisely because it selects among scarce confirmation slots. The firewall therefore replaces an efficient but potentially misspecified allocator with another rule: a lottery, queue order, structural quotas, or allocation based on post-entry evidence alone.

This is the normative tension of the paper. The firewall's robustness is purchased by giving up some calibrated allocative efficiency.

6. Ex Ante Sensitivity: Cutoff Leverage and Welfare Loss

Before an institution has randomized, it may still want to reason about when a firewall could be worthwhile. The following decomposition is an ex ante sensitivity device; Section 7 shows that once a randomized confirmation policy has actually run, the realized policy difference can be identified directly without decomposing outcomes into a calibrated baseline and hidden gains.

Let T denote the trusted rule. In a calibrated baseline state θ_0, suppose T selects set T, $|T| = K$, and is optimal given the available signal. Let s_i denote the evaluator score and $s_{(K)}$ the trusted cutoff.

Suppose evaluator misspecification on the *score scale* is pessimistic and bounded by e. Only candidates sufficiently close to the cutoff can have their selection status reversed. Define the vulnerable band

$$V_e = \{i \notin T : 0 < s_{(K)} - s_i \leq e\},$$

with $H_e = |V_e|$. If misspecification affects at most fraction η of the queue, the number of candidates whose allocation can be reversed is bounded by

$$m(\eta, e) = min\{\lfloor \eta N \rfloor, H_e\}.$$

Cutoff-leverage principle

The score error e and welfare error g are different objects. e governs *which allocation decisions can flip*. Let g_i denote the downstream welfare gain of a hidden candidate relative to the calibrated baseline or candidate it displaces. There is no requirement that $g_i \leq e$. Indeed, the economically important regime can be

$$e \text{ small}, g_i \text{ large}.$$

A small score error can place a candidate on the wrong side of a hard capacity cutoff while the welfare consequence of that discrete decision is arbitrarily larger than the score perturbation. This distinction matters whenever the evaluator's score is a ranking device or proxy rather than a cardinal measure of social welfare.

Let $V_\lambda(\theta_0) = \sum_i p_i(\lambda) v_i^0$ be the expected value of firewall intensity λ in the calibrated baseline, and define its calibrated efficiency cost

$$\Delta_\lambda^0 = V_T(\theta_0) - V_\lambda(\theta_0).$$

Suppose the evaluator is pessimistically misspecified on a hidden set $B \subseteq T^c$, and candidate $i \in B$ has true value $v_i = v_i^0 + g_i$, $g_i > 0$, while other values remain unchanged.

Proposition 3. Ex ante firewall sensitivity at a declared misspecification state

At this common state, firewall intensity λ has lower regret than the trusted rule if and only if

$$\sum_{i \in B} p_i(\lambda) g_i > \Delta_\lambda^0.$$

Proof

At any common state θ,

$$R_T(\theta) - R_\lambda(\theta) = V_\lambda(\theta) - V_T(\theta),$$

because the oracle value cancels. Relative to θ_0, the trusted rule receives none of the gains on B because $B \cap T = \varnothing$. The randomized firewall receives candidate i with probability $p_i(\lambda)$, increasing its expected value by $\sum_{i \in B} p_i(\lambda) g_i$. The sign gives the result. ▫

Proposition 3 is intentionally simple. Its substantive content comes from the cutoff geometry in (17)-(18) and the fact that score misspecification and welfare loss need not share a scale. It is a sensitivity statement before randomization, not a minimax theorem.

For the full uniform K-of-N lottery, $f_i = K/N$ and

$$\Delta_{lot}^0 = K\left(\dot{v}_T^0 - \dot{v}^0\right).$$

If the vulnerable band is not binding and $m = \eta N$, the full-lottery break-even condition reduces to

$$\eta g > \acute{\nu}_T^0 - \acute{\nu}^0.$$

Equation (21) is transparent but generally not identified from deterministic historical selection. That is the motivation for the next section.

7. Randomization as Institutional Self-Measurement

The ex ante decomposition in Section 6 is useful before experimentation, but randomized confirmation identifies a different and more direct object. Let Q denote the realized confirmation queue. Importantly, Q may itself depend on the upstream plausibility signal because exploration depth remains stage-local. All quantities in this section are therefore conditional on the queue that actually formed.

For candidate $i \in Q$, let $D_i \in \{0,1\}$ indicate outcome observation under firewall intensity λ, with known first-order inclusion probability

$$p_i(\lambda) = P\left(D_i = 1 \mid Q, \text{pre-randomization information}\right), \; p_i(\lambda) > 0.$$

Let Y_i be the measured outcome. Initially suppose Y_i is an unbiased affine measurement of the target value ν_i; Section 7.3 relaxes this assumption.

The trusted rule has target value

$$V_T(Q) = \sum_{i \in Q} t_i \nu_i,$$

while firewall intensity λ has target value

$$V_\lambda(Q) = \sum_{i \in Q} p_i(\lambda) \nu_i.$$

Define the queue-conditional realized policy contrast

$$D_\lambda(Q) = V_\lambda(Q) - V_T(Q).$$

Proposition 4. Randomized confirmation identifies the queue-conditional policy contrast

The estimator

$$\widehat{D}_\lambda = \sum_{i \in Q} \frac{p_i(\lambda) - t_i}{p_i(\lambda)} D_i Y_i$$

is design-unbiased for $D_\lambda(Q)$ when Y_i is unbiased for ν_i on the target scale.

Proof

Because $E\left[D_i Y_i \mid Q\right] = p_i(\lambda)\nu_i$ under randomized inclusion and unbiased outcome measurement,

$$E\left[\widehat{D}_\lambda \mid Q\right] = \sum_{i \in Q} \left[p_i(\lambda) - t_i\right]\nu_i = V_\lambda(Q) - V_T(Q). \; \square$$

This is stronger than estimating the calibrated efficiency cost Δ_λ^0 from historical data. Under deterministic status-quo confirmation, outcomes are missing exactly for the candidates needed to estimate the counterfactual queue mean. Once positive-probability randomized confirmation has run, the institution can estimate the realized rule difference directly without identifying a hidden set B or decomposing values into $v_i^0 + g_i$.

The baseline/misspecification decomposition remains useful *ex ante*: it helps an institution decide whether randomization is worth trying, interpret why a difference arose, and extrapolate to queues with different compositions. It is not needed to determine which policy performed better in the randomized queue itself.

The conditioning on Q is substantive. If exploratory depth depends on S^{ext}, pessimistically scored candidates can be filtered out before randomization. Proposition 4 then identifies the firewall contrast among candidates that survived that earlier process, not the end-to-end contrast from the original submission population. If the candidates most likely to benefit from the firewall are also least likely to enter Q, the queue-conditional contrast is attenuated toward the incumbent rule.

7.1 Partial randomization estimates the full-firewall contrast

Because (14) is a convex mixture,

$$V_\lambda(Q) = (1-\lambda) V_T(Q) + \lambda V_F(Q),$$

where $V_F(Q)$ is the value of the full firewall allocation f on the same queue. Therefore

$$D_\lambda(Q) = \lambda \left[V_F(Q) - V_T(Q) \right).$$

For any $\lambda > 0$,

$$\frac{\widehat{D}_\lambda}{\lambda}$$

is therefore an unbiased estimator of the realized full-firewall contrast on that queue. A partial lottery can learn about a full firewall without fully adopting it.

The division by λ has a precision cost. Partial randomization is not a free identification trick.

7.2 Identification does not imply useful precision

Under independent Bernoulli confirmation for illustration, and allowing conditional outcome variance $\sigma_i^2 = Var(Y_i \mid i)$, the variance of (24) is

$$Var\left(\widehat{D}_\lambda\right) = \sum_i \left[p_i(\lambda) - t_i \right)^2 \left[\frac{1 - p_i(\lambda)}{p_i(\lambda)} v_i^2 + \frac{\sigma_i^2}{p_i(\lambda)} \right).$$

Estimating the full-firewall contrast by $\widehat{D}_\lambda / \lambda$ amplifies this variance by $1/\lambda^2$. Rare inclusion, heavy-tailed values, and noisy outcomes can therefore make a design-unbiased estimator practically weak. A single annual funding program with a few randomized awards and decade-long outcomes may be unable to resolve a modest policy difference. Pooling across cohorts or programs, increasing λ, using preregistered intermediate outcomes, or targeting a coarser estimand may be necessary.

Equation (26) uses independent Bernoulli assignment only for transparency. An institution with a hard budget of exactly K confirmation studies should use a fixed-size unequal-probability design–for example conditional-Poisson/rejective or systematic sampling–that preserves the chosen first-order probabilities. Proposition 4 continues to hold because it uses only first-order inclusion expectations; valid variance estimation then requires the corresponding joint inclusion probabilities.

A mechanism can make a quantity identifiable long before it makes that quantity statistically learnable at institutionally useful precision.

7.3 Outcome validity requires both non-leakage and affine calibration

Randomization solves selection into outcome observation; it does not guarantee that the observed outcome is a cardinal measure of the welfare the institution wants to compare. Two distinct requirements matter.

First, the outcome measure should not reintroduce the plausibility signal at fixed welfare. Let W_i denote the target welfare construct and Y_i the observed proxy. Define *outcome-measure leakage* conditional on true welfare and allowed structural variables by

$$\ell_Y = \max_{w,x,s,s'} TV\left(P\left(Y \mid W=w, X=x, S^{ext}=s\right), P\left(Y \mid W=w, X=x, S^{ext}=s'\right)\right).$$

If Y is bounded in an interval of width R_Y, then at fixed W and X,

$$\left|E\left[Y \mid W, X, S^{ext}=s\right) - E\left[Y \mid W, X, S^{ext}=s'\right)\right| \le R_Y \ell_Y.$$

Second, even $\ell_Y = 0$ is not sufficient for welfare identification. Let

$$h(w) = E\left[Y \mid W=w\right)$$

after conditioning or residualizing on allowed structural variables. If

$$h(w) = a + b\,w, \; b>0,$$

then constants cancel because both policies allocate K expected slots, and the randomized contrast on Y equals b times the welfare contrast. If h is nonlinear, randomization identifies a contrast in $h(W)$, not in W.

Define the worst affine approximation error

$$\delta_{aff} = \min_{a,b>0} \max_{w} \left|h(w) - (a+b\,w)\right).$$

For the rescaled full-firewall contrast,

$$\left|\frac{D_\lambda^Y}{\lambda} - b\left[V_F^W(Q) - V_T^W(Q)\right)\right| \le 2K\,\delta_{aff}$$

under zero leakage; with bounded differential leakage, a conservative additional term of order $2K\,R_Y\,\ell_Y$ applies. Thus the self-measurement argument requires approximate *affine* measurement, not merely non-differential measurement.

The distinction is substantive rather than cosmetic. Raw citations, log citations, binary follow-on indicators, and expert welfare ratings are different cardinalizations. A convex mapping h can magnify

differences in the upper range where the trusted allocation is concentrated and thereby make the firewall look worse relative to the underlying welfare scale; a concave mapping can attenuate that advantage. Barnett et al. (2024) estimate publications and citations, which are informative empirical endpoints, but translating those outcomes into scientific welfare requires an additional measurement argument.

7.4 Why a near-zero estimate is weak evidence against the firewall

Three features of the design can make a genuinely useful firewall look unhelpful.

First, Proposition 4 is queue-conditional. If low-S^{ext} candidates receive shallower exploration, likely beneficiaries can be removed before randomization, attenuating the contrast toward zero. Second, when the trusted rule concentrates on higher-W candidates, a convex proxy transformation can magnify the trusted side even with $\ell_Y = 0$. Third, low λ and heavy-tailed outcomes can produce confidence intervals so wide that an institution observes no statistically detectable difference for many cohorts. The third mechanism is not estimator bias, but it creates a predictable interpretive failure if ``not significant'' is read as evidence that protection has no value.

Accordingly, $\widehat{D}_\lambda \approx 0$ is strong evidence against the firewall only when queue coverage of intended beneficiaries, approximate affine outcome validity, and adequate statistical precision have themselves been established. Otherwise a null-looking result is compatible with a valuable firewall.

7.5 Worked partial-lottery design and ongoing audit

Suppose $N = 100$ eligible grant proposals compete for $K = 10$ confirmation studies. The trusted rule deterministically chooses ten. A full firewall is a uniform lottery, $f_i = 0.1$.

Choose $\lambda = 0.2$. Then a trusted top-ten proposal has

$$p_i(0.2) = 0.8 + 0.2(0.1) = 0.82,$$

while every other eligible proposal has

$$p_i(0.2) = 0.02.$$

The expected number of confirmation slots remains ten, while 1.8 expected slots per cohort are reallocated away from deterministic trust. Every candidate has positive inclusion probability, so the queue-conditional firewall contrast is identified. Because $D_{0.2} = 0.2(V_F - V_T)$, the observed policy contrast can be rescaled to estimate the full-firewall contrast, but the standard error is multiplied by five. In a heavy-tailed setting, the program should expect to pool repeated cohorts before the result is decision-useful. With a hard ten-study budget, the same first-order probabilities can be implemented through a fixed-size design rather than independent Bernoulli draws.

Randomization also need not be a temporary learning phase. Panels change, fields move, and evaluator calibration can drift. If λ is ever set permanently to zero, the institution loses positivity for detecting future deterioration. A standing floor $\lambda_{min} > 0$ therefore has a distinct interpretation: an ongoing audit expense that preserves the ability to measure policy drift. The appropriate floor trades recurring efficiency cost against monitoring speed and precision.

This is the practical continuum omitted by a binary firewall/trust comparison. More randomization buys more robustness and more information per cohort, but also moves the institution farther from the trusted allocation while the evaluator is correct.

8. Related Work and Empirical Anchors

The framework combines familiar components but places them in a particular institutional sequence.

Blackwell (1953) supplies the information-ordering result underlying Proposition 1. Weitzman (1979) provides a canonical account of costly search and why information about alternatives rationally affects search order. Multiple-testing procedures such as Holm (1979) formalize the need to tighten evidence when many hypotheses are examined. Manski (2021) emphasizes decision making when one probabilistic specification should not simply be trusted.

The screening-confirmation problem also has close relatives in work on repeated use of evaluative signals. Kleinberg and Raghavan (2021) show that algorithmic monoculture can reduce social welfare because decision makers make correlated errors when they rely on the same ranking system. The present problem is the sequential analogue: reuse of one plausibility signal at screening and confirmation can propagate one misspecification across stages. Coate and Loury (1993) provide an older statistical-discrimination tradition in which noisy beliefs and downstream assignment interact. Neither paper yields the firewall result here, but both show why stage structure matters for the social consequences of an evaluative signal.

Scientific evaluation provides a concrete motivating setting. Boudreau et al. (2016), using randomized evaluator-proposal assignments, found systematic relationships between intellectual distance, novelty, and resource allocation. Ayoubi et al. (2021) report selectivity patterns disfavouring more novel work by their measures. Teplitskiy et al. (2022) emphasize that novelty and peer-review outcomes vary across contexts and that publication-conditioned samples can obscure the underlying selection process. These findings motivate possible misspecification; they do not establish that a firewall is warranted in a particular institution.

Partial funding lotteries demonstrate both feasibility and identification value. The Health Research Council of New Zealand began using a modified lottery for Explorer Grants in 2013; eligible applications are screened and then funded at random until the budget is exhausted (Liu et al., 2020; Barnett et al., 2024). Barnett et al. exploit the resulting randomized assignment to estimate the effect of funding on subsequent publications and citations. The Volkswagen Foundation has used partial randomization in the *Experiment!* initiative since 2017 after scientific review of eligible high-risk proposals (VolkswagenStiftung, 2024). These cases are not direct tests of the screening-confirmation firewall, but they show that randomization can be inserted after eligibility screening and can generate institutional knowledge unavailable under deterministic ranking.

9. Scope Conditions and Strategic Response

Four boundaries should be explicit.

First, the firewall is stage-local. The plausibility signal may still influence exploratory depth. An implausible candidate can receive too little exploration to generate a hit and therefore never reach the protected queue. This is not a contradiction: priors are often valuable in search. It does mean that Proposition 4 identifies a contrast conditional on a queue whose composition may already reflect the signal being audited. If systematic pessimism suppresses exactly the candidates most likely to benefit from the firewall, the estimated queue-conditional contrast is attenuated relative to an end-to-end intervention. Robustness at exploration would require a separate mechanism such as minimum exploratory depth or randomized exploration, with a separate efficiency cost.

Second, candidate submission and effort are treated as exogenous. A known lottery or eligibility threshold can change behavior: applicants may optimize to enter the protected queue, change the riskiness of proposals, or alter effort devoted to eligibility dimensions. The protected margin and the gameable margin can therefore coincide. The HRC applicant survey found that most respondents reported little change in application effort under the lottery, but this is not a general mechanism-design result (Liu et al., 2020). Endogenous entry and strategic manipulation are important extensions.

Third, pilots can be both information and intervention. A downstream experiment may consume a niche, reveal an idea, change organizational receptivity, train the candidate, or otherwise alter the payoff environment. In such cases S is not merely an observation of a fixed state. The clean Blackwell decomposition becomes a sequential control problem in which experimentation both learns about and changes realization conditions.

Fourth, randomization identifies policy performance only relative to its declared outcome metric and eligible queue. Class leakage ℓ_C and outcome leakage ℓ_Y provide partial audits of two ways the excluded plausibility signal can re-enter the system. Neither can certify the absence of every proxy pathway.

10. Discussion: What the Firewall Does and Does Not Solve

The framework supports three design claims.

Preserve continuation value when information arrives later. A channel-marginal screen can destroy candidates whose downstream value depends on information acquired only after admission. The first-best remedy depends on why the information is missing: move it earlier when possible, improve continuation-value calculation when computation is the bottleneck, and change the mandate when marginal scoring is imposed administratively.

Separate exploration from confirmation. Exploratory breadth should not be conflated with evidentiary acceptance. Queue hits, continue search when appropriate, and confirm with a process calibrated to the breadth actually searched. Removing the original plausibility signal from confirmation also reduces one obvious source of repeated-error correlation.

If a signal is being bypassed, do not let it screen twice. If exploration is explicitly meant to circumvent a suspected blind spot in a plausibility signal, reusing that same signal to allocate scarce confirmation or set candidate-specific evidentiary thresholds can nullify the protection. The screening-confirmation firewall prevents that reuse after queue entry. It deliberately gives up information when the evaluator is right, so its value is empirical rather than axiomatic.

Section 6 supplies an ex ante sensitivity analysis for institutions deciding whether to randomize. Section 7 supplies the stronger ex post result: once partial randomized confirmation has run, the institution can estimate the realized policy contrast on the queue that actually formed. That same section states the limits of institutional self-measurement. Low-intensity randomization may identify but not precisely estimate the effect; upstream signal-dependent exploration can filter beneficiaries before the queue forms; and outcome proxies require both low leakage and approximately affine measurement if their contrast is to represent welfare. A permanently positive randomization floor can therefore serve as continuing audit capacity rather than a one-time experiment.

11. Conclusion

Evaluation under context-dependent value is a staged decision problem. An upstream screen acts before some information about candidate-channel fit exists or can be used. A downstream matcher may later obtain that information and condition realization on it. Blackwell's information order implies that stronger downstream signals increase attainable continuation value and weakly increase the welfare mass hidden by a channel-marginal screen.

The coverage frontier makes that insight operational. Prediction orders where to search, optionality determines what can be searched at all, and approximate rankings can suffice. Uniform weight error gives a $2k\epsilon$ regret bound, while a separated top-k boundary gives exact recovery.

Exploration and confirmation should then be separated. Broad exploration creates multiplicity and a larger confirmation queue, not an inherent reason to stop exploring at the first noisy hit. Scarce confirmation is where evaluator misspecification becomes an institutional design problem. Reusing the same plausibility signal to ration confirmation can make a suspected blind spot operate twice and can propagate the same failure mode across stages.

A screening-confirmation firewall prevents that double use, but it sacrifices useful information when the evaluator is calibrated. Before randomization, the relevant sensitivity condition compares this calibrated cost with the hidden welfare expected to be recovered. Small score errors near a hard cutoff can flip decisions whose welfare consequences are much larger, so score error and welfare error must remain separate quantities.

Once randomized confirmation has run, the comparison becomes simpler. Known positive inclusion probabilities identify the realized value difference between the randomized firewall policy and the trusted rule directly; the calibrated-baseline/hidden-gain decomposition is no longer needed for that queue. Partial randomization therefore serves two functions: it protects candidates against evaluator misspecification and allows the institution to measure the value of that protection.

That self-measurement is only as good as the population it randomizes over, its statistical power, and its outcome scale. Signal-dependent exploration can remove intended beneficiaries before the protected queue forms. Heavy-tailed outcomes and small randomized shares can make an unbiased estimate too noisy to guide policy for many cycles. And even a non-leaking proxy identifies welfare only when its conditional expectation is approximately affine in the target construct; nonlinear transformations can materially change the measured policy contrast. Institutions should therefore treat queue formation, randomization intensity, pooling horizon, reference-class leakage, outcome leakage, and affine outcome calibration as design choices rather than afterthoughts.

The paper's specific recommendation is consequently modest:

> **Use plausibility where it helps search. If exploration is meant to protect against a plausibility model's blind spots, firewall that model from scarce confirmation to the degree worth its robustness cost, and preserve enough randomization to keep auditing that cost.**

References

Ayoubi, C., Pezzoni, M., & Visentin, F. (2021). Does it pay to do novel science? The selectivity patterns in science funding. *Science and Public Policy, 48*(5), 635–648.

Barnett, A., Blakely, T., Liu, M., Garland, L., & Clarke, P. (2024). The impact of winning funding on researcher productivity, results from a randomized trial. *Science and Public Policy, 51*(6), 1042–1050. https://doi.org/10.1093/scipol/scae045

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics, 24*(2), 265–272. https://doi.org/10.1214/aoms/1177729032

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science, 62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Coate, S., & Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review, 83*(5), 1220–1240.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences, 118*(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Liu, M., Choy, V., Clarke, P., Barnett, A., Blakely, T., & Pomeroy, L. (2020). The acceptability of using a lottery to allocate research funding: A survey of applicants. *Research Integrity and Peer Review, 5*, 3. https://doi.org/10.1186/s41073-019-0089-z

Manski, C. F. (2021). Econometrics for decision making: Building foundations sketched by Haavelmo and Wald. *Econometrica, 89*(6), 2827–2853.

Teplitskiy, M., Peng, H., Blasco, A., & Lakhani, K. R. (2022). Is novel research worth doing? Evidence from peer review at 49 journals. *Proceedings of the National Academy of Sciences, 119*(47), e2118046119. https://doi.org/10.1073/pnas.2118046119

VolkswagenStiftung. (2024, May 30). Everyone is equal in the lottery drum. https://www.volkswagenstiftung.com/en/news/news/everyone-equal-lottery-drum

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica, 47*(3), 641–654. https://doi.org/10.2307/1910412