AUDITED PARTIAL IDENTIFICATION FOR OFF-POLICY EVALUATION UNDER NONIGNORABLE FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

When reward recording depends on the unseen reward itself, even unlimited logged data may not identify policy values. We study off-policy evaluation in this setting and ask which missing rewards are most informative to recover through targeted audits. Under observed transitions, target coverage, and a shared rectangular reward model, we characterize the sharp joint set of policy values and quantify how common uncertainty cancels in policy comparisons. At the population level, identifying a cell's reward mean reduces comparison width by exactly the product of the policies' absolute visitation difference and the cell's reward-mean ambiguity, independently of the recovered mean. Minimizing the remaining width therefore reduces to top-K audit selection for equal costs or a knapsack problem for unequal costs. For a fixed observation design, half this width lower-bounds the worst-case expected absolute estimation error. An independent design–confirmation procedure provides asymptotically valid outer confidence intervals, including when policy visitation differences are zero or near zero. Compared with separate marginal bounds, shared-law comparison yields 19.3-fold narrower population bounds in a controlled MDP and 4.4-fold narrower plug-in bounds in a semi-synthetic KuaiRec study. Two real follow-up analyses further demonstrate the importance of outcome stability and accounting for incomplete reward recovery.

1 INTRODUCTION

Motivation. Off-policy evaluation (OPE) estimates a target policy's value from another policy's data. With normalized discounted occupancy d^π and state–action reward mean $r_x = \mathbb{E}[R \mid X = x]$, its value under initial law ν is

$$V^\pi = \sum_{x \in \mathcal{X}} d^\pi(x) r_x. \tag{1}$$

Dividing by $1-\gamma$ gives the usual return, or the state value $v^\pi(s)$ when $\nu = \delta_s$. IPW, doubly robust, FQE, and other value-function methods address distribution shift and sampling error for this target (Precup et al., 2000; Jiang & Li, 2016; Thomas et al., 2015; Shi et al., 2022). Response weighting can preserve point identification under missing at random. If recording instead depends on the unseen reward, even an infinite passive log need not identify r_x, V^π, or which policy is better. Delayed conversion tracking and outcome-dependent follow-up can have this property. We ask which missing rewards can still change the policy decision, and which to recover with a limited audit budget.

Main research questions. For fixed policies and missing-not-at-random (MNAR) rewards:

- **Q1.** What is the sharp identified set for policy values and contrasts?
- **Q2.** Which cells should a limited randomized audit activate before their outcomes are known?
- **Q3.** How can we estimate this set and report uncertainty when occupancy contrasts may be zero?

Main result: a decision-relevant interval. A *sharp identified interval* contains exactly the values compatible with the population observation law and model; its width may persist with unlimited data. Suppose transitions are observed, the log covers the target-visited cells $\mathcal{X}_\Pi$, and all policies share a reward vector with admissible means $\prod_x [\ell_x, u_x]$. Write $\Delta_x = u_x - \ell_x$ and $c_x = d^{\pi_1}(x) - d^{\pi_0}(x)$.

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.

Positive, reward-independent recovery identifies the original missing-reward means on audited cells $\mathcal{C}$. For compatible passive and audit laws, Theorem 4 gives the remaining width

$$W_{1:0}(\mathcal{C}) = \sum_{x \in \mathcal{X}_\Pi \setminus \mathcal{C}} |c_x| \Delta_x. \tag{2}$$

Each term is the width removed by auditing that cell, regardless of the recovered mean; the mean determines the interval's location.

Corollary 1 (The cost of separate policy bounds)**.** *Under the conditions of Theorem 4, subtracting the two marginal value intervals gives a valid outer interval of width* $W_{\text{sep}} = \sum_{x \in \mathcal{X}_\Pi \setminus \mathcal{C}} (d^{\pi_1}(x) + d^{\pi_0}(x)) \Delta_x$*. It has the same center as the sharp shared-law interval, and*

$$W_{\text{sep}} - W_{1:0} = 2 \sum_{x \in \mathcal{X}_\Pi \setminus \mathcal{C}} \min\{d^{\pi_1}(x), d^{\pi_0}(x)\} \Delta_x. \tag{3}$$

Corollary 2 (An irreducible estimation error)**.** *Under the same conditions, let* $T(P)$ *be a policy value or fixed linear policy contrast, with sharp interval width* $W(\mathcal{C})$*. For fixed sensitivity level* Γ*, over compatible full-data laws* $\mathfrak{P}_\Gamma(\mathcal{C})$ *and observed-data estimators,*

$$\inf_{\widehat{T}} \sup_{P \in \mathfrak{P}_\Gamma(\mathcal{C})} \mathbb{E}_P |\widehat{T} - T(P)| \geq \tfrac{1}{2} W(\mathcal{C}). \tag{4}$$

For $T = V^{\pi_1} - V^{\pi_0}$*, point identification holds exactly when every unaudited cell has* $|c_x| \Delta_x = 0$*.*

Thus overlapping state–action occupancy cancels shared uncertainty; disjoint supports do not. The remaining width limits every estimator under the unchanged observation design. Both proofs are in Appendix B.

Cancellation by hand: two cells. Let $d^{\pi_1} = (3/4, 1/4)$, $d^{\pi_0} = (1/4, 3/4)$, and $r_1, r_2 \in [1/4, 3/4]$, with every pair feasible. Each policy value ranges over $[1/4, 3/4]$, so separate subtraction gives $[-1/2, 1/2]$. Using the *same* rewards in both values instead gives

$$V^{\pi_1} - V^{\pi_0} = \tfrac{1}{2}(r_1 - r_2) \in [-1/4, 1/4].$$

The common contribution $(r_1 + r_2)/4$ cancels. If an audit identifies $r_1 = 1/2$, we obtain $[-1/8, 1/8]$: auditing either cell removes $|c_x| \Delta_x = (1/2)(1/2) = 1/4$ of width. Before auditing, the absolute-error floor is $W/2 = 1/4$. A confidence interval must additionally cover estimation error.

Why sequential visitation matters. Two policies can choose the same action at a state yet value its missing reward differently because earlier decisions determine whether it is reached. Suppose π_1 reaches z at time one and terminates after its sole action, whereas π_0 terminates immediately. Both prescribe the same action at z, but auditing its reward removes $(1 - \gamma)\gamma \Delta_{za}$ of contrast width (Appendix B.1). Local action disagreement assigns this cell zero value; the occupancy contrast captures its downstream importance. Conversely, uncertainty cancels in cells visited equally by both policies. Audit priorities therefore reflect which missing rewards can affect the comparison.

Contribution and procedure. Building on shared-uncertainty comparison (Kallus & Zhou, 2018; Guerdan et al., 2024), our contribution connects sharp MNAR reward completion, each cell's exact *pre-audit* value in a sequential environment, and inference for policy comparisons. A design split estimates visitation and reward bounds, selects cells, and freezes the audit plan. Fresh confirmation rewards estimate missing means; an independent inference split supplies endpoints and uncertainty. Retaining both endpoint branches at uncertain occupancy-contrast signs supports an outer confidence interval. We report an ordering only when it excludes zero (Algorithm 1). With equal costs and simultaneous score errors bounded by ε, the estimated top-K support loses at most $2K\varepsilon$ in width reduction relative to the population optimum (Proposition 8). This quantifies the cost of learning audit priorities from data.

Figure 1 outlines this procedure. Proposition 3 and Theorem 4 give its identification results; Sections 5 and 6 give estimation and evidence.

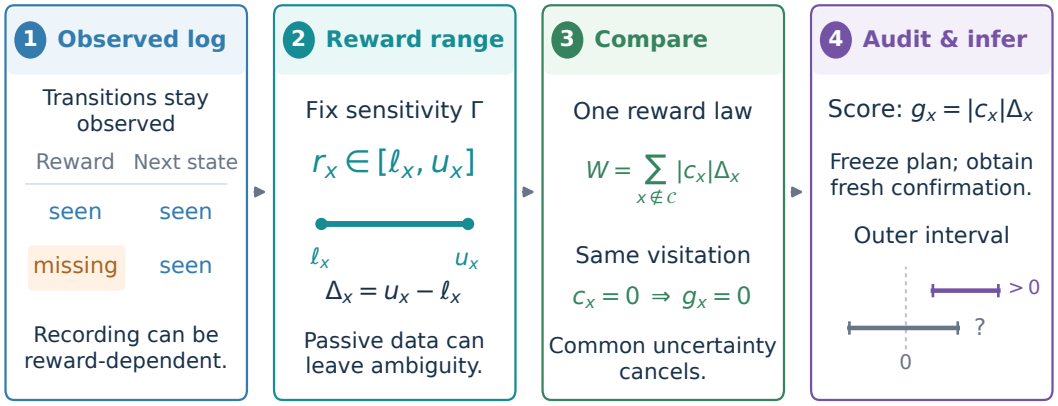


Figure 1: **From selective feedback to a policy decision.** The visitation contrast $c_x = d^{\pi_1}(x) - d^{\pi_0}(x)$ and reward ambiguity Δ_x determine $W = W_{1:0}(\mathcal{C})$. Valid audits remove selected terms under shared rectangular completion. A frozen design, fresh confirmation, and independent passive inference yield an outer interval. Schematic final intervals above zero favor π_1; those crossing zero leave the ordering unresolved.

2 Related Work

OPE with missing outcomes. IPW, doubly robust, and value-function methods estimate identified policy values (Jiang & Li, 2016; Shi et al., 2022; Uehara et al., 2022). Under MNAR, response models or identifying variables can restore identification under additional restrictions (Wang et al., 2025; Wei et al., 2026). We retain sensitivity-indexed value sets using selection-odds and expectile tools (Robins et al., 2000; Rosenbaum, 2002; Tan, 2006; Bellini et al., 2014; Bellini & Di Bernardino, 2017).

Shared uncertainty and audit design. Shared uncertainty already supports robust policy improvement and ranking (Kallus & Zhou, 2018; Zhang et al., 2021). Guerdan et al. (2024, Theorem 4.3) explicitly derive uncertainty cancellation and narrower regret bounds under confounded selective labels. Here observed transitions and shared rectangular reward completion make sequential occupancy contrasts yield an exact pre-audit width reduction. Double sampling recovers missing outcomes (Coppock et al., 2017; Levis et al., 2024); Neyman allocation optimizes precision within support (Neyman, 1934). We select support to reduce identification width and provide whole-set inference at occupancy-sign kinks (Fang & Santos, 2019; Hong & Li, 2018). Appendix A details the methodological boundaries.

3 Setup and Assumptions

Environment and policy value. A behavior policy μ generates data in a finite discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \nu)$, with state–action cells $X_t = (S_t, A_t) \in \mathcal{X} = \mathcal{S} \times \mathcal{A}$, rewards $R_{t+1} \in [0, 1]$, $\gamma \in (0, 1)$, and $S_0 \sim \nu$. Next states remain observed even when rewards are missing; ν is known or identified from episode starts. We evaluate fixed policies $\Pi = \{\pi_1, \ldots, \pi_J\}$ in this same environment. Each target-visited cell must occur under logging: $d^{\pi_j}(x) > 0 \Rightarrow d^{\mu}(x) > 0$ (direct coverage). Here occupancy is discounted visitation frequency:

$$d^{\pi_j}(s, a) = (1 - \gamma)\mathbb{E}_{\nu, P, \pi_j}\left[\sum_{t \geq 0} \gamma^t \mathbb{I}\{S_t = s, A_t = a\}\right]. \tag{5}$$

Observed transitions and coverage identify these visitation weights. We retain only target-visited cells $\mathcal{X}_\Pi = \{x : \max_j d^{\pi_j}(x) > 0\}$. With reward means $r_x = \mathbb{E}[R \mid X = x]$, the policy value is

$$V^{\pi_j} = \sum_{x \in \mathcal{X}_\Pi} d^{\pi_j}(x) r_x \in [0, 1]. \tag{6}$$

This is the standard OPE target, normalized to $[0, 1]$: division by $1 - \gamma$ gives the usual return, including $v^{\pi_j}(s)$ when $\nu = \delta_s$. The remaining identification problem is therefore about r. Collecting

occupancies in $D_{jx} = d^{\pi_j}(x)$ gives $V_\Pi = Dr$; D_x denotes column x. We study a fixed contrast $T_a = a^\top V_\Pi$. For example, $a = e_j$ selects one policy value, while $a = (1, -1)$ compares the rows (π_1, π_0).

Policies depend only on logged states. Recovering rewards leaves their occupancies unchanged. Appendix F treats jointly missing (R, S'), where this separation fails; Appendix D discusses state aggregation.

Nonignorable feedback. Let $O \in \{0, 1\}$ indicate whether the realized reward is present in the routine log. We always observe (X, O) and observe R only if $O = 1$. Write

$$\begin{aligned} \rho_x &= \mathbb{P}(O = 1 \mid X = x), \\ F_{ox} &= \mathcal{L}(R \mid O = o, X = x), \qquad \mu_{ox} = \mathbb{E}[R \mid O = o, X = x], \quad o \in \{0, 1\}, \\ q_x(r) &= \mathbb{P}(O = 1 \mid R = r, X = x). \end{aligned} \tag{7}$$

The log identifies the response rate ρ_x and respondent law F_{1x}, but leaves the nonrespondent law F_{0x} unknown. The value depends on both: $r_x = \rho_x \mu_{1x} + (1 - \rho_x)\mu_{0x}$. Our sensitivity model bounds how strongly recording can depend on the reward. It conditions on $X = (S, A)$, averaging over observed S'; conditioning also on S' would impose a different restriction. For episode-level sharpness, we additionally assume a time-homogeneous joint kernel for (R_{t+1}, S_{t+1}, O_t) given X_t under fixed μ. Without this condition, sharpness concerns the observed one-step law.

Assumption 1 (Bounded selection odds and response positivity)**.** *For every partially observed cell visited by at least one policy in* Π*, let* $F_x = \mathcal{L}(R \mid X = x)$*. When* $0 < \rho_x < 1$*, we assume* $0 < q_x(r) < 1$ *for* F_x*-almost every* r *and, for a user-selected* $\Gamma \geq 1$*,*

$$\frac{1}{\Gamma} \leq \frac{\{1 - q_x(r)\}/q_x(r)}{\{1 - q_x(r')\}/q_x(r')} \leq \Gamma \quad \text{for } F_x \otimes F_x\text{-a.e. } (r, r'). \tag{8}$$

For example, $\Gamma = 2$ lets the odds of missingness differ by at most a factor of two between rewards in the same cell. This standard sensitivity restriction (Robins et al., 2000; Rosenbaum, 2002; Tan, 2006; Zhao et al., 2019) permits otherwise unrestricted response. At $\Gamma = 1$, $O \perp R \mid X$; larger Γ permits stronger selection. Positivity requires $F_{0x} \sim F_{1x}$, so even the limit $\Gamma \to \infty$ is restricted to respondent support. The boundary cells are handled separately: $\rho_x = 1$ gives $\ell_x = u_x = \mu_{1x}$, while $\rho_x = 0$ gives $[\ell_x, u_x] = [0, 1]$ absent valid auditing.

Which audits identify missing rewards? An audit recovers initially missing rewards. Let $Z = 1$ indicate successful recovery among $O = 0$ records. The activated cells form $\mathcal{C} \subseteq \mathcal{X}_\Pi$; each $x \in \mathcal{C}$ satisfies

$$\begin{aligned} \mathbb{P}(Z = 1 \mid O = 0, X = x) > 0, \\ Z \perp R \mid (O = 0, X = x). \end{aligned} \tag{9}$$

Set $Z = 0$ outside $\mathcal{C}$. The recovered measurement must target the original outcome at the original time. Under (9), any positive recovery probability identifies F_{0x} in the population; complete recovery is unnecessary. The number of recovered outcomes controls precision. Reward-selective recovery instead leaves residual ambiguity, handled by Proposition 5.

Throughout, the audit and passive laws must be compatible with the chosen Γ. Otherwise the model is empty. A pre-specified valid test can reject compatibility; disjoint empirical supports alone do not establish rejection.

One feasible reward law for all policies. Let $\mathfrak{P}_\Gamma(\mathcal{C})$ contain full-data laws matching the passive and audited laws and satisfying Assumption 1. Each such law provides a possible value for the standard target (6).

Assumption 2 (Shared rectangular reward completion)**.** *Across* $x \in \mathcal{X}_\Pi$*, every collection of cellwise-admissible reward completions is jointly attainable by one full-data law, all policies share its single reward vector* r*, and the operative mean set is* $\prod_x [\ell_x(\mathcal{C}), u_x(\mathcal{C})]$*.*

Sharing and rectangularity play different roles. Sharing requires a cell to have the same reward law under every policy. Rectangularity lets admissible choices in different cells be combined freely. This

makes cellwise optima jointly attainable and yields exact additive audit gains. Appendix G discusses cross-cell restrictions. The sharp joint value set is

$$\mathcal{Z}_\Gamma^\Pi(\mathcal{C}) = \{V_\Pi(P) : P \in \mathfrak{P}_\Gamma(\mathcal{C})\} \tag{10}$$

and its directional projection $\mathcal{I}_{\Gamma,a}(\mathcal{C}) = \{a^\top V_\Pi(P) : P \in \mathfrak{P}_\Gamma(\mathcal{C})\}$. These sets contain the possible values of the standard OPE targets. Their widths measure the ambiguity that remains after observing the log and audits.

4 OPE UNDER NONIGNORABLE FEEDBACK

We first bound each missing-reward mean, then combine those bounds using policy occupancies. Comparisons require the same reward law for both policies.

4.1 REWARD BOUNDS AND SINGLE-POLICY VALUES

An expectile balances deviations above and below a candidate mean with unequal weights. For a bounded law F, its τ-expectile $e_\tau(F)$ is the unique solution of

$$\tau\mathbb{E}_F[(R-e)_+] = (1-\tau)\mathbb{E}_F[(e-R)_+]. \tag{11}$$

Proposition 3 (Sharp cellwise MNAR bounds)**.** *For a cell with $0 < \rho_x < 1$, under Assumption 1, the sharp missing-reward mean interval is*

$$\left[e_{\tau_-}(F_{1x}), e_{\tau_+}(F_{1x})\right], \qquad \tau_- = \frac{1}{1+\Gamma}, \quad \tau_+ = \frac{\Gamma}{1+\Gamma}. \tag{12}$$

The sharp full-reward endpoints in an unaudited cell are

$$\begin{aligned} \ell_x &= \rho_x\mu_{1x} + (1-\rho_x)e_{\tau_-}(F_{1x}), \\ u_x &= \rho_x\mu_{1x} + (1-\rho_x)e_{\tau_+}(F_{1x}), \end{aligned} \tag{13}$$

where $\mu_{1x} = \mathbb{E}_{F_{1x}}R$.

Bayes' rule turns the odds bound into restrictions on the weights dF_{0x}/dF_{1x}; their extreme means give these expectiles. The estimator replaces F_{1x} and ρ_x by their empirical counterparts. For binary rewards, Appendix B.1 gives the closed-form endpoints.

For $x \in \mathcal{C}$, random auditing identifies $r_x^{\text{aud}} = \rho_x\mu_{1x} + (1-\rho_x)\mu_{0x}$ and sets $\ell_x(\mathcal{C}) = u_x(\mathcal{C}) = r_x^{\text{aud}}$; otherwise $(\ell_x(\mathcal{C}), u_x(\mathcal{C})) = (\ell_x, u_x)$. Define

$$\begin{aligned} m_x(\mathcal{C}) &= \frac{\ell_x(\mathcal{C}) + u_x(\mathcal{C})}{2}, \\ h_x(\mathcal{C}) &= \frac{u_x(\mathcal{C}) - \ell_x(\mathcal{C})}{2}, \qquad \Delta_x = u_x - \ell_x. \end{aligned} \tag{14}$$

Single-policy OPE. Policy occupancies are nonnegative, so the smallest value uses every lower reward endpoint and the largest uses every upper endpoint. Under Assumption 2 and compatibility, both choices are feasible:

$$\mathcal{I}_\Gamma^{\pi_j}(\mathcal{C}) = \left[\sum_{x\in\mathcal{X}_\Pi} d^{\pi_j}(x)\ell_x(\mathcal{C}), \sum_{x\in\mathcal{X}_\Pi} d^{\pi_j}(x)u_x(\mathcal{C})\right], \tag{15}$$

with exact width $\sum_{x\in\mathcal{X}_\Pi\setminus\mathcal{C}} d^{\pi_j}(x)\Delta_x$. The value is point identified exactly when every unaudited cell with $d^{\pi_j}(x) > 0$ has $\Delta_x = 0$. Auditing all relevant cells is one way to achieve this; at $\Gamma = 1$, positive routine response in every remaining cell also suffices.

4.2 POLICY COMPARISON UNDER A SHARED REWARD LAW

Each uncertain cell contributes the same reward mean to both policies, weighted by their respective visitation frequencies. Equal weights cancel; different weights leave only their difference. The joint geometry is a *zonotope*: a translated sum of line segments, one per uncertain cell.

Theorem 4 (Joint policy values and exact contrast width)**.** *Suppose the audited and passive laws are compatible, so $\mathfrak{P}_\Gamma(\mathcal{C}) \neq \varnothing$. Under direct occupancy coverage and Assumption 2, the sharp joint policy-value set is*

$$\mathcal{Z}_\Gamma^\Pi(\mathcal{C}) = Dm(\mathcal{C}) + \sum_{x \in \mathcal{X}_\Pi} [-h_x(\mathcal{C}), h_x(\mathcal{C})] D_x. \tag{16}$$

For any contrast $a \in \mathbb{R}^J$, write $c_x(a) = a^\top D_x$. Its sharp interval $[L_a(\mathcal{C}), U_a(\mathcal{C})]$ has endpoints

$$\begin{aligned} L_a(\mathcal{C}) &= \sum_{\substack{x \in \mathcal{X}_\Pi \\ c_x \geq 0}} c_x \ell_x(\mathcal{C}) + \sum_{\substack{x \in \mathcal{X}_\Pi \\ c_x < 0}} c_x u_x(\mathcal{C}), \\ U_a(\mathcal{C}) &= \sum_{\substack{x \in \mathcal{X}_\Pi \\ c_x \geq 0}} c_x u_x(\mathcal{C}) + \sum_{\substack{x \in \mathcal{X}_\Pi \\ c_x < 0}} c_x \ell_x(\mathcal{C}), \end{aligned} \tag{17}$$

and exact width

$$W_a(\mathcal{C}) := U_a(\mathcal{C}) - L_a(\mathcal{C}) = \sum_{x \in \mathcal{X}_\Pi \setminus \mathcal{C}} |c_x(a)| \Delta_x. \tag{18}$$

Consequently, activating any $\mathcal{A} \subseteq \mathcal{X}_\Pi \setminus \mathcal{C}$ contracts width by

$$W_a(\mathcal{C}) - W_a(\mathcal{C} \cup \mathcal{A}) = \sum_{x \in \mathcal{A}} |c_x(a)| \Delta_x, \tag{19}$$

independently of the as-yet unseen audited outcomes, on any maintained post-audit model that remains compatible.

Identifying a cell's reward mean removes its uncertainty segment. For two policies this gives (2) and Corollaries 1–2. Cancellation operates through overlapping *state–action* visitation. Sequential occupancies also capture differences that local action scores miss: earlier decisions change which states are reached, even when later actions agree. Appendix B.1 illustrates this sequential effect; Appendix B proves attainability.

5 Proposed Estimation and Audit Design

The design has two decisions: *which cells to audit* determines the remaining identification width; *how many rewards to obtain* determines sampling precision. Algorithm 1 separates design from confirmation so that the final inference uses a frozen audit plan.

Algorithm 1 Audited OPE with independent design and confirmation

Require: A log with observed transitions; fixed policies, contrast a, cells, and Γ
Require: Independent design and inference units; costs, budgets, success minima, and confidence level
Ensure: A plug-in value interval and an outer confidence band
1: Estimate policy occupancies and passive reward bounds on the design split
2: Use a design-only pilot, if needed, to estimate audit success and reward variances
3: Select cells and allocate attempts, reserving confirmation floors; freeze the plan
4: Collect randomized confirmation rewards targeting the original outcomes
5: Estimate occupancies and reward endpoints on the inference split; exclude all pilot rewards
6: Keep passive bounds in cells below the confirmation minimum; combine endpoints using contrast signs
7: Add the pre-specified confidence expansion; declare an ordering only if the band excludes zero

The primary procedure assumes outcome-independent confirmation and compatible passive and audit laws. Selective success uses the residual bounds in Proposition 5 and requires separately justified inference. Any pre-specified compatibility rejection is reported separately; coverage is not claimed conditional on passing such a test.

5.1 Choosing audit cells and sample sizes

The value of activating a cell is its exact contribution to removed width:

$$g_x(a) = |c_x(a)| \Delta_x, \qquad c_x(a) = \sum_{j=1}^J a_j d^{\pi_j}(x). \tag{20}$$

At the population level, equal activation costs give top-K selection; unequal costs give the knapsack program (43). For two policies, prioritize cells with both large visitation differences and wide reward bounds. This rule minimizes remaining identification width. The recovered outcomes and final confidence band then determine whether an ordering can be certified.

Within the chosen support, use the cost-adjusted Neyman allocation

$$(n_x^{\mathrm{aud}})^* \propto \frac{|c_x(a)|(1-\rho_x)\sigma_{0x}}{\sqrt{c_x^{\mathrm{unit}}}}, \tag{21}$$

where σ_{0x} is the missing-reward standard deviation and c_x^{unit} is cost per successful audit. For positive-weight cells, this minimizes the audit variance contribution given a larger passive log (Appendix B.7). A pilot estimates σ_{0x}; without one, $\sigma_{0x} \le 1/2$ supplies conservative planning weights for rewards in $[0,1]$.

A finite-budget plan must first reserve enough attempts to meet confirmation minima. Appendix G combines these floors with activation and total attempt costs. Remaining resources can follow (21) after these reservations. Cells below the success minimum retain passive bounds. With reward-selective success, the gain becomes $|c_x(a)|(\Delta_x - \Delta_x^A)$; Proposition 5 and Appendix G extend planning to residual ambiguity and uncertain recovery costs.

5.2 From observations to the reported interval

Once the audit plan is frozen, five steps turn the inference data into a policy-value interval. The first four estimate identification endpoints; the fifth accounts for sampling error.

1. Estimate where each policy visits. Estimate $\widehat{P}(s' \mid s,a) = N(s,a,s')/N(s,a)$ from transition counts, then compute

$$\widehat{P}^{\pi}(s,s') = \sum_a \pi(a \mid s)\widehat{P}(s' \mid s,a),$$
$$(\widehat{d}_S^{\pi})^{\top} = (1-\gamma)\widehat{\nu}^{\top}(I - \gamma\widehat{P}^{\pi})^{-1}, \qquad \widehat{d}^{\pi}(s,a) = \widehat{d}_S^{\pi}(s)\pi(a \mid s). \tag{22}$$

A target-required row with $N(s,a) = 0$ requires a pre-specified merge or a report of insufficient coverage; inserting a zero transition row is invalid.

2. Estimate passive reward bounds. Let $\widehat{\rho}_x$, $\widehat{\mu}_{1x}$, and $\widehat{F}_{1x}$ be the empirical response rate, respondent mean, and respondent reward law. Compute the two empirical expectiles $\widehat{e}_{\tau_\pm,x} = e_{\tau_\pm}(\widehat{F}_{1x})$ and set

$$\widehat{\ell}_x^{\mathrm{pas}} = \widehat{\rho}_x\widehat{\mu}_{1x} + (1-\widehat{\rho}_x)\widehat{e}_{\tau_-,x},$$
$$\widehat{u}_x^{\mathrm{pas}} = \widehat{\rho}_x\widehat{\mu}_{1x} + (1-\widehat{\rho}_x)\widehat{e}_{\tau_+,x}. \tag{23}$$

Boundary conventions apply only when $\rho_x \in \{0,1\}$ is known by design. Incidental empirical boundaries require pre-specified merging or primitive projection.

3. Update cells with enough confirmation rewards. Estimate the missing-reward mean $\widehat{\mu}_{0x}$ using the Hájek mean (45), which weights each confirmation reward by its inverse frozen inclusion probability. Equal probabilities give the sample mean. On cells meeting the pre-specified success minimum, replace both endpoints by $\widehat{r}_x^{\mathrm{aud}} = \widehat{\rho}_x\widehat{\mu}_{1x} + (1-\widehat{\rho}_x)\widehat{\mu}_{0x}$; otherwise keep the passive bounds. Write these final endpoints as $\widehat{\ell}_x, \widehat{u}_x$ and their effective audit support as $\widehat{\mathcal{C}}_{\mathrm{act}}$; the exact rule is (46).

4. Combine cells with the correct signs. Set $\widehat{c}_x(a) = \sum_j a_j \widehat{d}^{\pi_j}(x)$ and report

$$\widehat{L}_a(\widehat{\mathcal{C}}_{\mathrm{act}}) = \sum_x \min\{\widehat{c}_x(a)\widehat{\ell}_x, \widehat{c}_x(a)\widehat{u}_x\},$$
$$\widehat{U}_a(\widehat{\mathcal{C}}_{\mathrm{act}}) = \sum_x \max\{\widehat{c}_x(a)\widehat{\ell}_x, \widehat{c}_x(a)\widehat{u}_x\}. \tag{24}$$

A negative coefficient reverses the reward endpoint used in the lower bound. For $a = e_j$, this estimates the single-policy interval (15); for $a = (1,-1)$, it estimates the shared-law policy difference.

Table 1: Complete sign-envelope pipeline at nominal 95% confidence: 200 replications per reward law. Coverage and certification are percentages; MCSEs are in parentheses. Top-20 and all-80 use 2,000 and 8,000 confirmation rewards, respectively. Tables 7–9 give the full design and boundary controls.

True contrast	Support	Set cover (MCSE)	CI width	Certify > 0	Certify < 0
0.0145	None	100.0 (0.0)	0.0304	0.0	0.0
0.0145	Learned top 20	99.0 (0.7)	0.0154	100.0	0.0
0.0145	All 80	93.0 (1.8)	0.0015	100.0	0.0
0.0000	All 80	95.5 (1.5)	0.0016	2.0	2.5
-0.0030	All 80	95.0 (1.5)	0.0016	0.0	100.0
0.0030	All 80	95.5 (1.5)	0.0015	100.0	0.0

5. Add uncertainty from finite samples. The plug-in interval estimates the range of compatible values. We expand it into an outer band for the *whole range*, at nominal confidence $1-\alpha$. Near equal visitation, an estimated coefficient can change sign. The screened sign-envelope therefore keeps both endpoint branches when $|\widehat{c}_x| \le a_n = N_{\mathrm{pas}}^{-2/5}$ and one branch otherwise. For each multiplier draw it takes the larger outward perturbation over the retained branches, including both occupancy and reward-endpoint noise. The upper $(1-\alpha/2)$ quantiles of the two outward perturbations give

$$\widehat{\mathcal{I}}^{\mathrm{dir}}_{a,1-\alpha} = \left[\widehat{L}_a - \frac{(q^*_{L,n})_+}{r_n}, \widehat{U}_a + \frac{(q^*_{U,n})_+}{r_n}\right]. \tag{25}$$

Here N_{pas} is the passive inference size and $r_n = \sqrt{\min_k n_k}$ for independent inference-source sizes n_k. Appendix C defines the multiplier maps and quantiles. Keeping both branches protects coverage near ties; Appendix C quantifies the resulting conservative envelope. A separately predeclared primitive-projection interval (47) provides an alternative.

Inference guarantee. Corollary 10 establishes consistent endpoints and uniform asymptotic whole-set coverage under independent bounded sampling units, positive cell probabilities, regular expectiles, and nondegenerate endpoint variance. The guarantee allows zero or near-zero occupancy contrasts and unequal source sizes, including vanishing audit/passive ratios when all sources grow and the stated conditions hold. Appendix C details the regularity conditions and their boundaries.

Reading the final interval. For $V^{\pi_1} - V^{\pi_0}$, a positive lower confidence endpoint supports π_1; a negative upper endpoint supports π_0. A band containing zero leaves the ordering unresolved, even if it is narrow. Use level α/Q for each of Q pre-specified contrasts. The reporting procedure is fixed before inference; it is never chosen because it yields a desired ordering.

6 Evidence

We separate three questions: how much ambiguity shared rewards remove, which audits reduce it, and how accurately the full procedure reports uncertainty. The controlled MDP addresses all three; KuaiRec and the follow-up studies examine distinct practical limits. Full designs are in Appendix E.

Audit support and sample size. In the controlled MDP, the single-policy value is 0.6154; complete-case and MAR-IPW biases are 0.1321 and 0.1331 (Table 6). In Figure 2, 20 targeted cells leave identification width 0.1531, versus 0.1962 for random support. More samples within these cells improve precision but cannot reveal rewards in the other 60. Full support removes the remaining ambiguity.

Shared uncertainty and audit targeting. Shared-law comparison makes bounds 19.3 times narrower in the MDP and 4.4 times narrower in KuaiRec than separate marginal subtraction. These are gains from sharing the reward law. To isolate audit targeting, a 30-case MDP grid holds that law and activated-cell count fixed (Table 12). At $K = 20/40$, our score reduces median remaining width by $0.8\%/5.4\%$ relative to local-action ranking and $6.6\%/9.5\%$ relative to occupancy-difference-only ranking. The score therefore adapts the audit to both sequential visitation and reward ambiguity; Table 12 reports the full distribution of gains across environments.

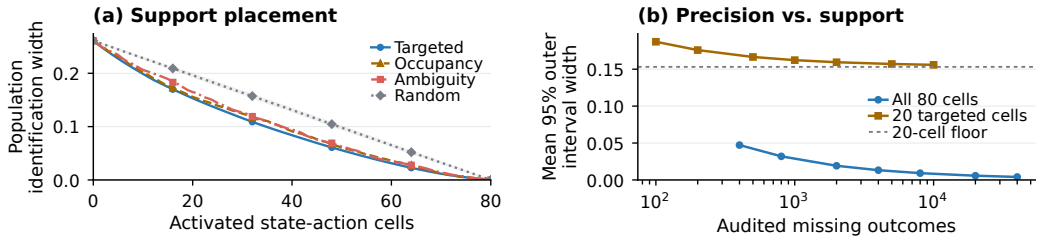


Figure 2: **Support controls identification; sample size controls precision.** (a) Targeted support reduces a single policy's population identification width. (b) More outcomes improve precision within fixed support: 20 cells leave a nonzero identification floor (dashed), whereas all 80 allow convergence to a point. These panels condition on the population and support to isolate audit sampling. Figure 4 reports conditional coverage; Table 1 evaluates the complete comparison pipeline.

The complete procedure and calibration. Table 1 evaluates the complete learned-design pipeline, including positive, zero, and near-zero contrasts. Learned top-20 support covers the positive-effect identified set in 99.0% of the original 200 runs and certifies the ordering in every run. A fresh sample-size scan increases both passive and audit samples without changing the method (Table 8). At the original sample size, 1,000 new repetitions give 95.2% positive-effect coverage (95% Monte Carlo interval $[93.7, 96.4]\%$), while null coverage is 95.6%. Near occupancy ties, the screened band attains 94.5–100% whole-set coverage across the local-to-zero diagnostic, with only 14.1–18.6% of fixed-scale projection's confidence expansion (Table 14). It thus retains informative intervals while accommodating uncertain signs. Appendix E reports all calibration and wrong-sign diagnostics with Monte Carlo uncertainty.

Sequential data with dependent observations. A separate pointwise experiment uses 400 independent length-75 episodes per split and reweights whole episodes in the bootstrap (Table 15). Under fixed fast- and slow-mixing kernels, learned top-20 auditing gives 96.0% and 95.5% whole-set coverage. It certifies the positive ordering in all 200 runs, compared with 2% and 1% without audits. These results demonstrate the value of reward recovery with within-episode dependence.

Matching the reward budgets. A paired comparison gives top-20 and all-80 the same confirmation budget ($B = 2{,}000$ or $8{,}000$) and pilot budget ($B/5$). All-80 produces narrower bands in all four conditions (Table 16), illustrating the value of removing the identification floor when every cell can be audited. This complements support targeting under an activation limit. Coverage varies across conditions; Appendix E.2 reports every result and its Monte Carlo uncertainty. These budgets exclude activation fees and failed attempts.

Evidence from real outcomes. KuaiRec demonstrates the comparison on real watch outcomes under a simulated response mask. Its constructed policies share the within-cell recommendation distribution and vary cell visitation, matching the shared-law model behind the 4.4-fold plug-in bound reduction. Appendix E.5 further quantifies sensitivity to within-cell policy differences.

The horizon-one YOP and Coppock applications examine audit validity and outcome recovery. In Uganda's YOP, 362 recovered outcomes contract bounds by 25.2% under temporal outcome stability (Fiala et al., 2025; Rose, 2026). In Coppock, explicitly bounding the 28 failures among 100 invitees still yields a 71.9% contraction under outcome stability (Coppock et al., 2017; 2016). Appendix E.1 reports the remaining intervals, temporal drift, and incentive-sensitivity analyses; these studies assess follow-up assumptions separately from the sequential MDP experiments.

Conclusion. Shared-law comparison locates the missing rewards that matter for a policy decision. Audit support reduces identification ambiguity; sample size improves precision. Ordering requires an outer confidence interval to exclude zero.

ETHICS STATEMENT

This work is intended to make uncertainty from selective feedback visible rather than to certify a policy from assumptions alone. Sensitivity parameters and audit designs can still be misspecified, and small audited subgroups can expose sensitive outcomes. A real audit requires consent, IRB approval, or another documented legal basis as appropriate; least-privilege access, encryption, retention and deletion limits; and monitoring that allocation does not disproportionately burden vulnerable groups. Deployments should report the full sensitivity curve, pre-specify stopping rules, and avoid interpreting interval exclusion as evidence of fairness or safety. Our YOP and Coppock analyses use de-identified public CC0 replication files and involve no new contact. KuaiRec likewise uses an existing public research dataset; its missingness and audits are simulated.

REPRODUCIBILITY STATEMENT

The supplement gives complete proofs, estimands, audit mappings, frame decisions, drift diagnostics, and inference conditions. The companion archive supplies public CC0 YOP and Coppock sources, derived data, code, bootstrap draws, seeded simulation outputs, tables, figures, and validation scripts. Its `Makefile` runs experiments and rebuilds displays; `data_manifest/` records sources, licenses, checksums, counts, exclusions, and split hashes. Result metadata records seeds, and `code/requirements.txt` pins dependencies. Markov, local-kink, and operational experiments need no external data. KuaiRec code, hashes, and derived results are included; its separately licensed raw archive must be obtained from the official source after license acceptance. Fixed seeds and numerical checks regenerate reported quantities; validators check scientific identities, exports, and source checksums. The archive README gives the full reproduction checklist.

AI USE STATEMENT

Generative AI tools assisted with developing and writing mathematical proofs, polishing the manuscript, writing code used to generate synthetic datasets for the simulation experiments, and finding related work. The authors proposed the original research ideas and independently sourced the real-world datasets used in the two empirical studies. The complete reproducibility package provides the data and code needed to reproduce all experimental results. The authors verified every detail of the AI-assisted mathematical proofs and take full responsibility for the results and the final manuscript.

REFERENCES

Mohammad Beigi, Ming Jin, Junshan Zhang, Qifan Wang, and Lifu Huang. Adversarial reward auditing for active detection and mitigation of reward hacking, 2026.

Fabio Bellini and Elena Di Bernardino. Risk management with expectiles. *The European Journal of Finance*, 23(6):487–506, 2017. doi: 10.1080/1351847X.2015.1052150.

Fabio Bellini, Bernhard Klar, Alfred Müller, and Emanuela Rosazza Gianin. Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54:41–48, 2014. doi: 10.1016/j.insmatheco.2013.10.015.

Alexis Bellot and Silvia Chiappa. Towards estimating bounds on the effect of policies under unobserved confounding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3321.

Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Operations Research*, 72(3):1071–1086, 2024. doi: 10.1287/opre.2021.0781.

Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1999–2007, 2021.

Andrew Bennett, Nathan Kallus, Miruna Oprescu, Wen Sun, and Kaiwen Wang. Efficient and sharp off-policy evaluation in robust markov decision processes. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3590.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1–3):33–57, 1996.

David A. Bruns-Smith. Model-free and model-based policy evaluation when causality is uncertain. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1116–1126, 2021.

Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S. Thomas. Universal off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shreyas Chaudhari, Renhao Zhang, Philip S. Thomas, and Bruno Castro da Silva. Which rewards matter? reward selection for reinforcement learning under limited feedback, 2025.

Hua Chen, Zhi Geng, and Xiao-Hua Zhou. Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics*, 65 (3):675–682, 2009.

Tong Chen and Thomas Lumley. Optimal sampling for design-based estimators of regression models. *Statistics in Medicine*, 41(8):1482–1497, 2022. doi: 10.1002/sim.9300.

Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737, 2013.

Pierre Clavier, Emmanuel Rachelson, Erwan Le Pennec, and Matthieu Geist. Bootstrapping expectiles in reinforcement learning, 2024.

Alexander Coppock, Alan S. Gerber, Donald P. Green, and Holger L. Kern. Replication data for: Combining double sampling and bounds to address non-ignorable missing outcomes in randomized experiments. Harvard Dataverse, Version 1, 2016.

Alexander Coppock, Alan S. Gerber, Donald P. Green, and Holger L. Kern. Combining double sampling and bounds to address nonignorable missing outcomes in randomized experiments. *Political Analysis*, 25(2):188–206, 2017. doi: 10.1017/pan.2016.6.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric J. Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024. doi: 10.1080/01621459.2023.2191817.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in Neural Information Processing Systems*, 33:9398–9411, 2020.

Lin Dong, Eric Laber, Yair Goldberg, Rui Song, and Shu Yang. Ascertaining properties of weighting in the estimation of optimal treatment regimes under monotone missingness. *Statistics in Medicine*, 39(25):3503–3520, 2020.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023. doi: 10.1080/01621459.2022.2069572.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *Review of Economic Studies*, 86(1):377–412, 2019. doi: 10.1093/restud/rdy049.

Nathan Fiala, Julian Rose, Filder Aryemo, and Jörg Ankel-Peters. Timing matters: The (very) long-run impacts of cash grants during a crisis. *Journal of Development Economics*, 175:103447, 2025. doi: 10.1016/j.jdeveco.2024.103447.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 540–550, 2022. doi: 10.1145/3511808.3557220.

Peter B. Gilbert, Xuesong Yu, and Andrea Rotnitzky. Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Statistics in Medicine*, 33(6):901–917, 2014. doi: 10.1002/sim.6006.

Yair Goldberg and Michael R Kosorok. Q-learning with censored data. *Annals of Statistics*, 40(1): 529–560, 2012.

Luke Guerdan, Amanda Lee Coston, Ken Holstein, and Steven Wu. Predictive performance comparison of decision policies under confounding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16673–16705. PMLR, 2024. URL https://proceedings.mlr.press/v235/guerdan24a.html.

Josiah P. Hanna, Yash Chandak, Philip S. Thomas, Martha White, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. *Journal of Machine Learning Research*, 25(313):1–58, 2024.

Tobias Hatt, Daniel Tschernutter, and Stefan Feuerriegel. Generalizing off-policy learning under sample selection bias. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 769–779, 2022.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Efficient and sharp off-policy learning under unobserved confounding. In *International Conference on Learning Representations*, 2026.

Han Hong and Jessie Li. The numerical delta method. *Journal of Econometrics*, 206(2):379–394, 2018. doi: 10.1016/j.jeconom.2018.06.007.

Joel L. Horowitz and Charles F. Manski. Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics*, 84(1):37–58, 1998. doi: 10.1016/S0304-4076(97)00077-8.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5301–5310, 2021.

Chinmaya Kausik, Yangyi Lu, Kevin Tan, Maggie Makar, Yixin Wang, and Ambuj Tewari. Offline policy evaluation and optimization under confounding. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1459–1467. PMLR, 2024.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23734–23757. PMLR, 2024.

Dingwen Kong and Lin Yang. Provably feedback-efficient reinforcement learning via active reward learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

David Krueger, Jan Leike, Owain Evans, and John Salvatier. Active reinforcement learning: Observing rewards at a cost, 2020.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003.

Alexander W. Levis, Rajarshi Mukherjee, Rui Wang, Heidi Fischer, and Sebastien Haneuse. Double sampling for informatively missing data in electronic health record-based comparative effectiveness research. *Statistics in Medicine*, 43(30):6086–6098, 2024. doi: 10.1002/sim.10298.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Myrl G. Marmarelis, Ali Hasan, Kamyar Azizzadenesheli, R. Michael Alvarez, and Anima Anandkumar. Off-policy predictive control with causal sensitivity analysis. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 2958–2972. PMLR, 2025.

Rui Miao, Zhengling Qi, and Xiaoke Zhang. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 593–606, 2022.

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683, 2016.

Wang Miao, Lan Liu, Yilin Li, Eric J Tchetgen Tchetgen, and Zhi Geng. Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/IMS Journal of Data Science*, 1(2):1–23, 2024. doi: 10.1145/3592389.

Daniel Molitor. Anytime-valid optimal policy identification, 2026.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18819–18831, 2020.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934. doi: 10.2307/2342192.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann, 2000.

Kevin Regan and Craig Boutilier. Regret-based reward elicitation for Markov decision processes. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 444–451, 2009. URL https://arxiv.org/abs/1205.2619.

Kevin Regan and Craig Boutilier. Robust policy computation in reward-uncertain MDPs using nondominated policies. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pp. 1127–1133, 2010.

James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 1–94. Springer, 2000.

Julian Rose. Replication data for: Timing matters: The (very) long-run impacts of cash grants during a crisis. Harvard Dataverse, Version 1, 2026.

Paul R. Rosenbaum. *Observational Studies*. Springer, 2 edition, 2002.

Martin Saveski, Steven Jecmen, Nihar B. Shah, and Johan Ugander. Counterfactual evaluation of peer-review assignment policies. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 84(3):765–793, 2022. doi: 10.1111/rssb.12465.

Chengchun Shi, Jin Zhu, Ye Shen, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, 119(545):273–284, 2024. doi: 10.1080/01621459.2022.2110878. URL https://doi.org/10.1080/01621459.2022.2110878.

Jorg Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4): 1299–1315, 2009.

Tatsuki Takahashi, Chihiro Maru, and Hiroko Shoji. Off-policy evaluation for recommendations with missing-not-at-random rewards. In *International Symposium on Affective Science and Engineering ISASE2025*, pp. 1–4. Japan Society of Kansei Engineering, 2025. doi: 10.5057/isase.2025-C000003. URL https://www.jstage.jst.go.jp/article/isase/ISASE2025/0/ISASE2025_1_19/_article/-char/en.

Rikiya Takehi, Masahiro Asami, Kosuke Kawakami, and Yuta Saito. A general framework for off-policy learning with partially-observed reward. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mUbYof5MKp.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006. doi: 10.1198/016214506000000023.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning, 2022. URL https://arxiv.org/abs/2212.06355.

Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22491–22511, 2022.

Han Wang, Yang Xu, Wenbin Lu, and Rui Song. Off-policy evaluation under nonignorable missing data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65020–65058. PMLR, 2025.

Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 24(3):1097–1116, 2014. doi: 10.5705/ss.2012.074.

Zixiao Wang, Amiremad Ghassami, and Ilya Shpitser. Identification and estimation for nonignorable missing data: A data fusion approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50467–50488. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24t.html.

Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/IMS Journal of Data Science*, 1(3):1–42, 2024. doi: 10.1145/3643693.

Ziheng Wei, Annie Qu, and Rui Miao. Off-policy evaluation for missingness-aware policies in MDPs with rewards missing not at random. *arXiv preprint arXiv:2606.20206*, 2026. URL https://arxiv.org/abs/2606.20206.

Xiangkun Wu, Ting Li, Gholamali Aminian, Armin Behnamnia, Hamid R. Rabiee, and Chengchun Shi. Pessimistic data integration for policy evaluation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3581. URL https://papers.nips.cc/paper_files/paper/2025/hash/9a3e2737cbb40d41b4a8efe33dbf511b-Abstract-Conference.html.

Yang Xu, Jin Zhu, Chengchun Shi, Shikai Luo, and Rui Song. An instrumental variable approach to confounded off-policy evaluation. In *International Conference on Machine Learning*, pp. 38848–38880. PMLR, 2023.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587–2615, 2022. doi: 10.1214/22-AOS2195.

Jasper B. Yang, Bryan E. Shepherd, Thomas Lumley, and Pamela A. Shaw. Optimal two-phase sampling designs for generalized raking estimators with multiple parameters of interest, 2025.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2020.

Bo Zhang, Jordan Weiss, Dylan S. Small, and Qingyuan Zhao. Selecting and ranking individualized treatment rules with unmeasured confounding. *Journal of the American Statistical Association*, 116(533):295–308, 2021. doi: 10.1080/01621459.2020.1736083.

Junzhe Zhang and Elias Bareinboim. Causal eligibility traces for confounding robust off-policy evaluation. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4933–4942. PMLR, 2025.

Jiwei Zhao and Yanyuan Ma. A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association*, 117(540):1916–1930, 2022.

Qingyuan Zhao, Dylan S. Small, and Bhaswar B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B*, 81(4):735–761, 2019. doi: 10.1111/rssb.12327.

A Closest Methodological Boundaries

OPE and missing outcomes. Classical OPE methods address action shift and sampling error around point-identified values (Bradtke & Barto, 1996; Lagoudakis & Parr, 2003; Precup et al., 2000; Liu et al., 2018; Nachum et al., 2019; Dudík et al., 2011; Jiang & Li, 2016; Kallus & Uehara, 2022; Uehara et al., 2020; Yin & Wang, 2020; Thomas et al., 2015; Dai et al., 2020; Shi et al., 2022; Chandak et al., 2021; Uehara et al., 2022). Ignorable missingness permits inverse-weighted or semiparametric correction (Goldberg & Kosorok, 2012; Dong et al., 2020; Tsiatis, 2007). Takehi et al. (2025) point-identify contextual-bandit policy gradients using $p(o \mid x)$ and fully observed secondary rewards; we instead retain a sensitivity-indexed MNAR set. Under MNAR rewards, Wang et al. (2025) study monotone missingness and use inverse-probability weighting for point-value estimation under an explicitly modeled response mechanism, Takahashi et al. (2025) combine reward and logging weights, and Wei et al. (2026) use future-state shadows and completeness. Selection models, instruments, shadows, proximal bridges, and data fusion supply other point-identifying restrictions (Chen et al., 2009; Wang et al., 2014; Miao et al., 2016; Zhao & Ma, 2022; Miao et al., 2024; Bennett et al., 2021; Bennett & Kallus, 2024; Cui et al., 2024; Miao et al., 2022; Wang et al., 2024). Separately, Wu et al. (2025) combine historical controls with experimental data to improve A/B testing under distribution shift.

Partial identification and shared uncertainty. Missing-outcome and bounded-odds models are classical (Manski, 1990; Horowitz & Manski, 1998; Robins et al., 2000; Rosenbaum, 2002; Tan, 2006; Zhao et al., 2019), as is expectile density-envelope duality (Bellini et al., 2014; Bellini & Di Bernardino, 2017). Closely related odds-envelope work derives sharp causal bounds and regular estimators under unmeasured treatment confounding (Yadlowsky et al., 2022; Dorn & Guo, 2023). Those models perturb treatment assignment rather than outcome recording, and therefore do not imply the observed-law completion or audit-support geometry studied here. Proposition 3 applies the known duality to the outcome-response law and supplies the reverse construction needed for sharpness. OPE bounds for action-support holes instead use monotonicity or smoothness (Saveski et al., 2023; Khan et al., 2024). Inference for partially identified parameters distinguishes parameter coverage from whole-set coverage (Imbens & Manski, 2004; Stoye, 2009); intersection-bound methods provide related inference tools (Chernozhukov et al., 2013). Our whole-set requirement also covers occupancy-sign kinks, using directional-bootstrap theory (Fang & Santos, 2019; Hong & Li, 2018).

Shared-uncertainty policy improvement and ranking are established (Kallus & Zhou, 2018; Zhang et al., 2021). Guerdan et al. (2024, Theorem 4.3 and Appendix E) derive cancellation, its gain over separate subtraction, and the identification–sampling distinction for predictive utility under confounded selective labels; we claim none of these general principles as new. Here observed MDP transitions and rectangular MNAR reward completion give the pre-audit objective (2): valid recovery removes selected terms, followed by confirmation allocation and whole-set inference at occupancy-sign kinks. Equation (27) shows why local action agreement misses sequential visitation differences. Related confounding bounds include Bellot & Chiappa (2024) and Hess et al. (2026); the latter give sharp one-policy bounds, efficient estimators, and improvement guarantees from an upper regret bound. Regan & Boutilier (2009) elicit reward-bound constraints through regret-based queries, and Regan & Boutilier (2010) share rewards across policies. Our audits instead recover selectively missing realized outcomes, with completion restricted by the response law and audit design.

Adjacent designs and boundaries. Auxiliary-variable, proxy, instrumental, sensitivity, and robust-MDP methods restrict actions, states, populations, or transitions (Shi et al., 2024; Namkoong et al., 2020; Kallus & Zhou, 2020; Bruns-Smith, 2021; Hatt et al., 2022; Kausik et al., 2024; Zhang & Bareinboim, 2025; Marmarelis et al., 2025; Bennett et al., 2021; Xu et al., 2023; Iyengar, 2005; Nilim & El Ghaoui, 2005; Bennett et al., 2024). Our occupancies use observed transitions; Appendix F treats missing reward–transition packets separately. Classical double sampling (Coppock et al., 2017; Levis et al., 2024) motivates recovery, while Neyman and two-phase designs minimize variance (Neyman, 1934; Gilbert et al., 2014; Chen & Lumley, 2022; Yang et al., 2025). We separate support activation for identification from contrast-weighted allocation (21). Acquisition and exploration obtain labels, actions, or trajectories (Krueger et al., 2020; Kong & Yang, 2022; Chaudhari et al., 2025; Wan et al., 2022; Hanna et al., 2024); our audits recover rewards from completed actions. Reward-hacking audits (Beigi et al., 2026) address a different target. Our inference is batch, unlike

time-uniform OPE (Karampatziakis et al., 2021; Waudby-Smith et al., 2024; Molitor, 2026). Table 2 summarizes the closest boundaries.

Table 2: Selected methodological boundaries. Cancellation and shared uncertainty predate this paper. The audit column concerns acquisition of initially missing outcomes, rather than statistical assessment alone. "Not studied" describes the cited formulation, not an impossibility.

Work	Main uncertainty	Output	Shared-law comparison	Outcome acquisition	Inference
Wang et al. (2025)	MNAR monotone dropout	Point value	Not studied	No	Point CI
Guerdan et al. (2024)	Confounded selective labels	Regret interval	Yes; cancellation	No	Endpoint estimation
Hess et al. (2026)	Unmeasured treatment confounding	Sharp value bounds; regret upper bound	Regret upper bound	No	Efficient estimation
Yadlowsky et al. (2022); Dorn & Guo (2023)	Unmeasured treatment confounding	Sharp effect bounds	Not the target	No	Regular / orthogonal
Khan et al. (2024)	Action-support gaps	Sharp value set	Not studied	No	Bound estimation
Coppock et al. (2017); Levis et al. (2024)	Outcome nonresponse	Bounds / point effect	Not studied	Random follow-up	Bound CI / EIF
This paper	MNAR rewards, observed S'	Sharp joint value set	Occupancy contrasts	Targeted cells	Whole-set; conditional kink band

B Proofs

B.1 Binary rewards and consequences for policy comparison

For binary respondent success probability p_x, the formulas reduce to

$$e_{\tau_-} = \frac{p_x}{p_x + \Gamma(1-p_x)}, \qquad e_{\tau_+} = \frac{\Gamma p_x}{\Gamma p_x + 1 - p_x}. \tag{26}$$

Why visitation, rather than local action choice? For a contextual bandit, $d^{\pi_j}_{\mathrm{CB}}(s,a) = p(s)\pi_j(a \mid s)$, so (2) becomes $\sum_{(s,a)\in\mathcal{X}_\Pi\setminus\mathcal{C}} p(s)|\pi_1(a \mid s) - \pi_0(a \mid s)|\Delta_{sa}$ (a one-decision discounted embedding adds a factor $1-\gamma$). Sequential decisions differ: suppose π_1 reaches z at $t=1$ and terminates after its sole action, while π_0 terminates at $t=0$. Both prescribe the same sole action at z, yet

$$|d^{\pi_1}(z,a) - d^{\pi_0}(z,a)|\Delta_{za} = (1-\gamma)\gamma\Delta_{za} > 0. \tag{27}$$

Local action agreement misses this uncertainty because earlier decisions change whether z is visited.

B.2 Proof of Proposition 3

For any probability law F, define the density-weight class

$$\mathcal{W}_\Gamma(F) = \left\{ w > 0 : \begin{array}{l} \mathbb{E}_F w = 1, \quad 0 < \operatorname{ess\,inf} w, \\ \operatorname{ess\,sup} w \le \Gamma \operatorname{ess\,inf} w \end{array} \right\}. \tag{28}$$

We first show that $dF_{0x}/dF_{1x} \in \mathcal{W}_\Gamma(F_{1x})$, and that every member of this class is induced by a response function satisfying (8). This supplies both the bounds and their attainability.

Fix a cell with $0 < \rho_x < 1$ and suppress subscripts. Write

$$F = \rho F_1 + (1-\rho)F_0, \qquad w := \frac{dF_0}{dF_1} = \frac{\rho}{1-\rho}\frac{1-q}{q}.$$

Bayes' rule and Assumption 1 imply

$$\mathbb{E}_{F_1} w = 1, \qquad \frac{\operatorname{ess\,sup} w}{\operatorname{ess\,inf} w} \le \Gamma, \qquad w \in \mathcal{W}_\Gamma(F_1).$$

Conversely, for any $w \in \mathcal{W}_\Gamma(F_1)$, define

$$F_w = \rho F_1 + (1-\rho)wF_1, \qquad q_w(r) = \frac{\rho}{\rho + (1-\rho)w(r)}.$$

Then

$$q_w\, dF_w = \rho\, dF_1, \qquad (1-q_w)\, dF_w = (1-\rho) w\, dF_1,$$

so the constructed law reproduces both observed response strata.

To preserve the always-observed next state, retain the respondent joint law of (R, S') and couple wF_1 with the observed nonrespondent S' marginal. This fixes every observed cell kernel; the time-homogeneous-kernel condition then fixes the observed episode law by induction.

Proposition 8 of Bellini et al. (2014), applied at $\tau_- = 1/(1+\Gamma)$ and $\tau_+ = \Gamma/(1+\Gamma)$, gives

$$\inf_{w \in \mathcal{W}_\Gamma(F_1)} \int r w\, dF_1 = e_{\tau_-}(F_1), \qquad \sup_{w \in \mathcal{W}_\Gamma(F_1)} \int r w\, dF_1 = e_{\tau_+}(F_1),$$

including endpoint attainment. Moreover, for $w_t = (1-t)w_0 + t w_1$,

$$\operatorname{ess\,sup} w_t \le (1-t) \operatorname{ess\,sup} w_0 + t \operatorname{ess\,sup} w_1 \le \Gamma\{(1-t) \operatorname{ess\,inf} w_0 + t \operatorname{ess\,inf} w_1\} \le \Gamma \operatorname{ess\,inf} w_t,$$

and $\mathbb{E}_{F_1} w_t = 1$. Thus $\mathcal{W}_\Gamma(F_1)$ is convex and its mean image is the whole interval between the attained extrema. Hence

$$\left\{ \rho\mu_1 + (1-\rho) \int r\, dF_0 : F_0 = wF_1,\ w \in \mathcal{W}_\Gamma(F_1) \right\}$$
$$= \rho\mu_1 + (1-\rho)[e_{\tau_-}(F_1), e_{\tau_+}(F_1)] = [\ell_x, u_x].$$

This proves (12)–(13); direct substitution of a Bernoulli law into (11) gives (26). □

B.3 Selective success of an audit

For an activated cell, write $\eta_x = \mathbb{P}(Z=1 \mid O=0, X=x) \in (0,1)$, $G_{1x} = \mathcal{L}(R \mid O=0, Z=1, X=x)$, and $q_x^A(r) = \mathbb{P}(Z=1 \mid O=0, R=r, X=x)$. Suppose audit response is positive and, for a user-selected $\Gamma_A \ge 1$,

$$0 < q_x^A(r) < 1 \quad F_{0x}\text{-a.e.}, \qquad \frac{1}{\Gamma_A} \le \frac{\{1-q_x^A(r)\}/q_x^A(r)}{\{1-q_x^A(r')\}/q_x^A(r')} \le \Gamma_A \quad F_{0x} \otimes F_{0x}\text{-a.e.} \tag{29}$$

Proposition 5 (Selective audit success)**.** *For a cell with $0 < \rho_x < 1$, under Assumption 1, (29), and both response-positivity conditions, the exact class for $F_{0x} = \mathcal{L}(R \mid O=0, X=x)$ is*

$$\mathcal{H}_x = \left\{ \eta_x G_{1x} + (1-\eta_x) w_A G_{1x} :\ w_A \in \mathcal{W}_{\Gamma_A}(G_{1x}),\ \frac{d\{\eta_x G_{1x} + (1-\eta_x) w_A G_{1x}\}}{dF_{1x}} \in \mathcal{W}_\Gamma(F_{1x}) \right\}. \tag{30}$$

If $\mathcal{H}_x \neq \varnothing$, define

$$\mu_{0x}^{A,-} = \inf_{H \in \mathcal{H}_x} \int r\, dH(r), \qquad \mu_{0x}^{A,+} = \sup_{H \in \mathcal{H}_x} \int r\, dH(r),$$
$$\ell_x^A = \rho_x \mu_{1x} + (1-\rho_x)\mu_{0x}^{A,-}, \qquad u_x^A = \rho_x \mu_{1x} + (1-\rho_x)\mu_{0x}^{A,+}. \tag{31}$$

Then $[\ell_x^A, u_x^A]$ is the sharp full-reward mean interval and is nested in $[\ell_x, u_x]$. If $\Gamma_A = 1$, it collapses to the audited full-reward mean, provided the routine and audit laws are compatible. The same conclusion holds in the separately handled complete-ascertainment boundary $\eta_x = 1$.

Proof. Let $G_{0x} = \mathcal{L}(R \mid O=0, Z=0, X=x)$. Conditional Bayes' rule gives

$$\frac{dG_{0x}}{dG_{1x}}(r) = \frac{\eta_x}{1-\eta_x} \frac{1-q_x^A(r)}{q_x^A(r)}.$$

Thus, by Proposition 3's Bayes correspondence,

$$(29) \iff w_A := \frac{dG_{0x}}{dG_{1x}} \in \mathcal{W}_{\Gamma_A}(G_{1x}), \qquad F_{0x} = \eta_x G_{1x} + (1-\eta_x) w_A G_{1x},$$

and routine response requires

$$(8) \iff \frac{dF_{0x}}{dF_{1x}} \in \mathcal{W}_\Gamma(F_{1x}).$$

These are precisely the two constraints defining $\mathcal{H}_x$. Conversely, every member is generated by

$$q_x^A(r) = \frac{\eta_x}{\eta_x + (1-\eta_x)w_A(r)}$$

and the routine response function in Proposition 3. Applying that proof's coupling within the observed (O, Z) strata preserves the joint (R, S') laws when observed, every unobserved-reward S' marginal, and hence the transition and episode laws.

The class is convex. For continuous rewards, put $\lambda = F_{1x} + G_{1x}$, $p = dF_{1x}/d\lambda$, $g = dG_{1x}/d\lambda$, and $f = dH/d\lambda$. The ratio restrictions are equivalent to the existence of $t \in [1/\Gamma, 1]$ and $t_A \in [1/\Gamma_A, 1]$ with

$$tp \le f \le \Gamma tp, \qquad \{\eta_x + (1-\eta_x)t_A\}g \le f \le \{\eta_x + (1-\eta_x)\Gamma_A t_A\}g, \qquad \int f\, d\lambda = 1.$$

These constraints define a weak-* closed, bounded subset of $L^\infty(\lambda)$ times two compact intervals. Its projection is compact, and $f \mapsto \int rf\, d\lambda$ is weak-* continuous. Thus the extreme means are attained; convexity fills every intermediate mean. Moreover,

$$\mathcal{H}_x \subseteq \{wF_{1x} : w \in \mathcal{W}_\Gamma(F_{1x})\} \quad \Longrightarrow \quad [\ell_x^A, u_x^A] \subseteq [\ell_x, u_x].$$

At $\Gamma_A = 1$, $w_A = 1$ and $F_{0x} = G_{1x}$; when $\eta_x = 1$, complete ascertainment gives the same result directly. □

Applying Proposition 3 to audit response alone gives

$$\begin{aligned} b_{0x}^- &= \eta_x \mu_{A1,x} + (1-\eta_x) e_{1/(1+\Gamma_A)}(G_{1x}), \\ b_{0x}^+ &= \eta_x \mu_{A1,x} + (1-\eta_x) e_{\Gamma_A/(1+\Gamma_A)}(G_{1x}). \end{aligned} \tag{32}$$

Intersecting $[b_{0x}^-, b_{0x}^+]$ with the passive missing-mean interval is valid but need not be sharp: projection does not commute with intersection. On common finite support, the sharp endpoints follow from the linear program obtained by imposing both density-ratio constraints in (30). For example, on $\{0, 1/2, 1\}$ take $F_1 = (14, 15, 11)/40$, $G_1 = (4, 3, 1)/8$, $\Gamma = 3$, $\Gamma_A = 3/2$, and $\eta = 1/2$. The separately projected intervals intersect to $[9/32, 11/32]$, whereas (30) has the strictly smaller sharp interval $[49/160, 11/32]$; endpoint witnesses are $(21/40, 27/80, 11/80)$ and $(9/20, 33/80, 11/80)$.

Corollary 6. *For activated cells use $[\ell_x^A, u_x^A]$ from (31); elsewhere retain $[\ell_x, u_x]$. If every local class is nonempty and Assumption 2 holds for these modified cellwise sets, Theorem 4 holds with the modified endpoints. Writing $\Delta_x^A = u_x^A - \ell_x^A$, activation reduces directional width by*

$$|c_x(a)|(\Delta_x - \Delta_x^A), \tag{33}$$

rather than necessarily by $|c_x(a)|\Delta_x$. This follows directly from Proposition 5 and Theorem 4.

Planning when audit success is unknown. Let $\mathcal{A}_x$ be a simultaneous pilot confidence region for the successful-audit law, η_x, and any user-specified Γ_A. Solving (30) over each $A \in \mathcal{A}_x$ gives a worst-case residual width $\overline{\Delta}_x^A$ and hence guaranteed activation gain

$$\underline{g}_x^A = |c_x(a)| \max\{0, \Delta_x - \overline{\Delta}_x^A\}. \tag{34}$$

Equation (34) conditions on the population $c_x(a)$ and Δ_x; when they are estimated, include them in the simultaneous design region and minimize the full gain over that region as in Proposition 8. If an attempt costs c_x^{attempt} and succeeds with probability η_x, its first-order cost per successful ascertainment is $c_x^{\text{attempt}}/\eta_x$; uncertainty in η_x is handled by its pilot lower confidence bound. Without historical or randomized pilot information, the passive law permits $\Delta_x^A = \Delta_x$, so no positive guaranteed pre-audit gain is available. A two-wave protocol can then randomize a small pre-specified exploration support, construct $\mathcal{A}_x$ there, and use disjoint confirmatory data for re-optimized support and endpoint inference. Transporting the pilot to unvisited cells requires an explicit homogeneity or pooling model.

Computation. Empirical expectiles cost $O(\sum_x n_x \log n_x)$ by sorting, and J dense occupancy solves cost $O(J|\mathcal{S}|^3)$, with standard sparse improvements. Equal-cost support is one sort. General 0–1 knapsack is NP-hard; integer-scaled costs admit $O(|\mathcal{X}_\Pi|B_{\text{on}})$ dynamic programming and profit scaling gives an FPTAS. On finite outcome support, each selective-success endpoint is a cellwise linear program.

B.4 Proof of Theorem 4

By Proposition 3, audit validity, and Assumption 2, the sharp reward-mean set is

$$\mathcal{R}(\mathcal{C}) = \prod_{x\in\mathcal{X}_\Pi} [\ell_x(\mathcal{C}), u_x(\mathcal{C})] = m(\mathcal{C}) + \prod_{x\in\mathcal{X}_\Pi} [-h_x(\mathcal{C}), h_x(\mathcal{C})].$$

Observed transitions make D common to every completion, hence

$$D\mathcal{R}(\mathcal{C}) = Dm(\mathcal{C}) + \sum_{x\in\mathcal{X}_\Pi} [-h_x(\mathcal{C}), h_x(\mathcal{C})]D_x,$$

which proves (16). With $c_x = a^\top D_x$, the maximum and minimum of $a^\top v$ over this set are

$$a^\top Dm(\mathcal{C}) \pm \sum_{x\in\mathcal{X}_\Pi} h_x(\mathcal{C})|c_x|.$$

The endpoint witnesses

$$r_x^U = m_x(\mathcal{C}) + h_x(\mathcal{C})\,\text{sign}(c_x), \qquad r_x^L = m_x(\mathcal{C}) - h_x(\mathcal{C})\,\text{sign}(c_x)$$

are jointly attainable by Assumption 2 (take either value when $c_x = 0$). Convexity of $\mathcal{R}(\mathcal{C})$ fills the scalar interval between them. Expanding the support values gives (17) and (18); setting $h_x = 0$ on audited cells gives (19). □

B.5 Proof of Corollary 1

Linearity gives the common center. Since $r + s - |r - s| = 2\min\{r, s\}$ for $r, s \geq 0$, substitute $r = d^{\pi_1}(x)$ and $s = d^{\pi_0}(x)$, multiply by Δ_x, and sum over $x \in \mathcal{X}_\Pi \setminus \mathcal{C}$ to obtain (3). □

B.6 Proof of Corollary 2

Let P_- and P_+ be the sign-aware endpoint completions from Theorem 4. Their passive-and-supported-audit laws coincide, while $T_a(P_+) - T_a(P_-) = W_a(\mathcal{C})$. Thus every available-data statistic T has the same distribution Q under both laws. Pointwise,

$$|T - T_a(P_-)| + |T - T_a(P_+)| \geq W_a(\mathcal{C}). \tag{35}$$

Taking Q-expectations shows that one endpoint risk is at least half the width. Equation (18) vanishes exactly when no unaudited cell has $|c_x(a)|\Delta_x > 0$. □

B.7 Within-support allocation and vanishing audits

Proposition 7. *Let n_x^{aud} count successfully ascertained missing rewards. With effective per-success cost c_x^{unit} and fixed total sampling cost $\sum_x c_x^{\text{unit}} n_x^{\text{aud}} = B_{\text{unit}}$, the continuous variance minimizer over cells with positive first-order weight is given by Equation (21). If the frozen support stabilizes to a nonrandom $\mathcal{C}$ containing every cell with $|c_x(a)|\Delta_x > 0$, give each positive-weight cell a diverging count and a nonrandom positive limiting audit share. If $n_{\text{aud}} := \sum_x n_x^{\text{aud}} = o(n)$ and the passive estimator from n observations is root-n consistent, the frozen-design and conditional-CLT conditions in Appendix B yield asymptotic normality of the full plug-in estimator.*

Cells below pre-specified confirmation minima retain passive width; diverging relevant passive and successful-audit counts are required even when the audit fraction vanishes. For $J = 1, a = 1$, (21) is classical Neyman allocation.

Proof. Let $\sigma_{0x}^2 = \mathrm{Var}(R \mid O = 0, X = x)$, $b_x = |c_x(a)|(1-\rho_x)\sigma_{0x}$, and $\mathcal{X}_+^{\mathcal{C}}(a) = \{x \in \mathcal{C} : b_x > 0\}$. For $b_x > 0$ under $\sum_x c_x^{\mathrm{unit}} n_x^{\mathrm{aud}} = B_{\mathrm{unit}}$, Cauchy–Schwarz gives

$$\left(\sum_x \frac{b_x^2}{n_x^{\mathrm{aud}}}\right)\left(\sum_x c_x^{\mathrm{unit}} n_x^{\mathrm{aud}}\right) \geq \left(\sum_x b_x \sqrt{c_x^{\mathrm{unit}}}\right)^2,$$

with equality exactly when $n_x^{\mathrm{aud}} \propto b_x/\sqrt{c_x^{\mathrm{unit}}}$, proving (21).

For the limit, let $\mathcal{D}_n$ be generated by the independent design split and design-only pilot, and define

$$\mathcal{R}_a = \{x : |c_x(a)|\Delta_x > 0\}, \qquad \mathcal{Z}_a = \{x \notin \mathcal{C} : c_x(a) = 0,\ \Delta_x > 0\}.$$

Let $n_{\mathrm{aud}} = \sum_{x \in \mathcal{C}} n_x^{\mathrm{aud}}$. Assume that the frozen support converges in probability to a nonrandom $\mathcal{C}$ and, conditionally on $\mathcal{D}_n$, pilot outcomes are excluded, $\mathcal{R}_a \subseteq \mathcal{C}$, $n_{\mathrm{aud}}/n \to 0$, and, for nonrandom constants λ_x,

$$\frac{n_x^{\mathrm{aud}}}{n_{\mathrm{aud}}} \to_p \lambda_x \in (0,1] \quad (x \in \mathcal{X}_+^{\mathcal{C}}(a)), \qquad \sum_{x \in \mathcal{X}_+^{\mathcal{C}}(a)} \lambda_x = 1. \tag{36}$$

Each activated cell also receives a diverging $o(n_{\mathrm{aud}})$ floor. Assume the passive estimator is root-n and the confirmation means satisfy the conditional Lindeberg–Feller or martingale CLT. Because $\mathcal{R}_a \subseteq \mathcal{C}$,

$$\mathcal{X}_\Pi \setminus \mathcal{C} = \{x \notin \mathcal{C} : \Delta_x = 0\} \cup \mathcal{Z}_a.$$

On $\mathcal{Z}_a$, choose any bounded measurable $\widehat{r}_x^{\circ} \in [\widehat{\ell}_x, \widehat{u}_x]$. These means are not point identified, but $c_x(a) = 0$; since $\mathcal{X}_\Pi$ is finite,

$$\sqrt{n_{\mathrm{aud}}} \sum_{x \in \mathcal{Z}_a} \{\widehat{c}_x(a)\widehat{r}_x^{\circ} - c_x(a) r_x\} = \sqrt{n_{\mathrm{aud}}} \sum_{x \in \mathcal{Z}_a} \widehat{c}_x(a)\widehat{r}_x^{\circ} = O_p\left(\sqrt{\frac{n_{\mathrm{aud}}}{n}}\right) = o_p(1).$$

The $\Delta_x = 0$ cells and all passive occupancy, response-rate, and respondent-law terms likewise contribute $O_p(\sqrt{n_{\mathrm{aud}}/n}) = o_p(1)$. With frozen $\mathcal{C}, n_x^{\mathrm{aud}}$, pilot outcomes excluded, and the stated conditional confirmation CLT, conditionally on $\mathcal{D}_n$,

$$\begin{aligned} \sqrt{n_{\mathrm{aud}}}(\widehat{T}_a - T_a) &= \sum_{x \in \mathcal{X}_+^{\mathcal{C}}(a)} \frac{c_x(a)(1-\rho_x)}{\sqrt{\lambda_x}} \sqrt{n_x^{\mathrm{aud}}}(\widehat{\mu}_{0x} - \mu_{0x}) + o_p(1) \\ &\Rightarrow N\left(0, \sum_{x \in \mathcal{X}_+^{\mathcal{C}}(a)} \frac{c_x(a)^2(1-\rho_x)^2\sigma_{0x}^2}{\lambda_x}\right). \end{aligned} \tag{37}$$

Zero-variance cells contribute zero after their diverging floors; Slutsky's theorem gives the same limit unconditionally. □

Attempt budgets. If n_x confirmation attempts have outcome-independent success rate η_x, then $c_x^{\mathrm{unit}} = c_x^{\mathrm{attempt}}/\eta_x$ and (21) is equivalent to $n_x \propto b_x/\sqrt{\eta_x c_x^{\mathrm{attempt}}}$. Estimation and resampling use the two-phase score in (53), or its without-replacement analogue. Under selective success, (44) is replaced by the residual-endpoint influence function or a verified primitive bootstrap.

Proposition 8. *Let $\mathfrak{F}_B$ be the feasible support family in (43), let $\mathcal{X}_\Pi$ be fixed and finite, and let $\Theta_{1-\alpha_s}$ be a simultaneous confidence region for $\theta = (D, \ell, u)$. Write $\Theta = \Theta_{1-\alpha_s}$ and define*

$$g_x(\theta) = |a^\top D_x|(u_x - \ell_x), \quad \underline{g}_x = \inf_{\theta \in \Theta} g_x(\theta), \quad \overline{g}_x = \sup_{\theta \in \Theta} g_x(\theta), \tag{38}$$

after intersecting Θ with $u_x \geq \ell_x$ and the occupancy parameter space, and set

$$\varepsilon_x = \sup_{\theta \in \Theta} |g_x(\theta) - g_x(\widehat{\theta})|. \tag{39}$$

Let $G(S) = \sum_{x\in S} g_x(\theta_0)$. On $\{\theta_0 \in \Theta\}$, these intervals and radii hold simultaneously. If $\widehat{S}$ is the plug-in optimum and $S^\star$ the population optimum, then

$$0 \le G(S^\star) - G(\widehat{S}) \le \sum_{x\in S^\star} \varepsilon_x + \sum_{x\in \widehat{S}} \varepsilon_x \le \underbrace{\max_{S\in\mathfrak{F}_B} \sum_{x\in S} \varepsilon_x + \sum_{x\in\widehat{S}} \varepsilon_x}_{B_\varepsilon(\widehat{S})}. \tag{40}$$

Alternatively, the box-robust support

$$S^{\mathrm{LCB}} \in \arg\max_{S\in\mathfrak{F}_B} \sum_{x\in S} \underline{g}_x \tag{41}$$

maximizes worst-case gain over the rectangular score outer hull and satisfies

$$0 \le G(S^\star) - G(S^{\mathrm{LCB}}) \le \underbrace{\max_{S\in\mathfrak{F}_B} \sum_{x\in S} \overline{g}_x - \sum_{x\in S^{\mathrm{LCB}}} \underline{g}_x}_{B_{\mathrm{LCB}}}. \tag{42}$$

If Θ has asymptotic coverage $1-\alpha_s$, both certificates hold with asymptotic probability at least $1-\alpha_s$. A uniform radius ε and maximum feasible size K_B give plug-in regret at most $2K_B\varepsilon$; under equal costs, $g_{(K)} - g_{(K+1)} > 2\varepsilon$ gives exact support recovery.

Proof. On $\{\theta_0 \in \Theta\}$, projection gives every interval and radius simultaneously, including when $a^\top D_x = 0$. Writing $\widehat{G}(S) = \sum_{x\in S} g_x(\widehat{\theta})$, plug-in optimality gives

$$G(S^\star) - G(\widehat{S}) \le \{G(S^\star) - \widehat{G}(S^\star)\} + \{\widehat{G}(\widehat{S}) - G(\widehat{S})\},$$

which is bounded by the two sums in (40). For the LCB rule,

$$G(S^\star) \le \max_{S\in\mathfrak{F}_B} \sum_{x\in S} \overline{g}_x, \qquad G(S^{\mathrm{LCB}}) \ge \sum_{x\in S^{\mathrm{LCB}}} \underline{g}_x,$$

which proves (42). Finally, $\inf_{g_x\in[\underline{g}_x,\overline{g}_x]} \sum_{x\in S} g_x = \sum_{x\in S} \underline{g}_x$, so (41) is exactly maximin for the rectangular outer box, not necessarily for the dependent joint image of Θ. If $\varepsilon_x \le \varepsilon$ and $|S| \le K_B$,

$$G(S^\star) - G(\widehat{S}) \le |S^\star|\varepsilon + |\widehat{S}|\varepsilon \le 2K_B\varepsilon.$$

Under equal costs and $g_{(K)} - g_{(K+1)} > 2\varepsilon$, every population top-$K$ cell x and excluded cell y satisfy

$$\widehat{g}_x - \widehat{g}_y \ge g_x - g_y - 2\varepsilon > 0,$$

which proves exact support recovery. □

For several pre-specified contrasts $a^{(q)}$ with nonnegative weights λ_q, weighted total width remains additive with cell score $\Delta_x \sum_q \lambda_q |a^{(q)\top} D_x|$. A minimax-over-contrasts objective is instead an integer epigraph problem; it is not a knapsack.

C Endpoint Estimation and Inference

C.1 Implementation formulas and reporting conventions

If $c_x^{\mathrm{on}} > 0$ is the fixed activation cost and B_{on} the activation budget, the population support problem is

$$\max_{z_x\in\{0,1\}} \sum_{x\in\mathcal{X}_\Pi} z_x g_x(a) \quad \text{s.t.} \quad \sum_{x\in\mathcal{X}_\Pi} c_x^{\mathrm{on}} z_x \le B_{\mathrm{on}}. \tag{43}$$

Support and allocation are frozen before disjoint confirmation. For cells with $1-\rho_x > 0$, let n_x^{aud} missing rewards be audited independently and let $\sigma_{0x}^2 = \mathrm{Var}(R \mid O = 0, X = x)$. Conditional on the larger passive log, the audit contribution to contrast variance is

$$\mathrm{AVar}(\widehat{T}_a) = \sum_{x\in\mathcal{C}} \frac{\{c_x(a)(1-\rho_x)\}^2 \sigma_{0x}^2}{n_x^{\mathrm{aud}}}. \tag{44}$$

Its cost-constrained minimizer is (21).

For planned support $\mathcal{C}$ satisfying (9), define the effective finite-sample support by $\widehat{\mathcal{C}}_{\mathrm{act}} = \{x \in \mathcal{C} : n_x^{\mathrm{succ}} \geq n_{x,\min}\}$, where the confirmation minima are pre-specified. The primary audit estimator is the design-weighted Hájek mean

$$\widehat{\mu}_{0x} = \frac{\sum_i \mathbb{I}\{X_i = x, O_i = 0\} Z_i^C R_i / \zeta_{ix}}{\sum_i \mathbb{I}\{X_i = x, O_i = 0\} Z_i^C / \zeta_{ix}}, \tag{45}$$

where ζ_{ix} is the frozen confirmation inclusion probability. A Horvitz–Thompson mean is a design-unbiased alternative. Set

$$\widehat{r}_x^{\mathrm{aud}} = \widehat{\rho}_x \widehat{\mu}_{1x} + (1 - \widehat{\rho}_x)\widehat{\mu}_{0x}, \qquad (\widehat{\ell}_x(\widehat{\mathcal{C}}_{\mathrm{act}}), \widehat{u}_x(\widehat{\mathcal{C}}_{\mathrm{act}})) = \begin{cases} (\widehat{r}_x^{\mathrm{aud}}, \widehat{r}_x^{\mathrm{aud}}), & x \in \widehat{\mathcal{C}}_{\mathrm{act}}, \\ (\widehat{\ell}_x^{\mathrm{pas}}, \widehat{u}_x^{\mathrm{pas}}), & x \notin \widehat{\mathcal{C}}_{\mathrm{act}}. \end{cases} \tag{46}$$

Thus cells below their minima retain passive endpoints. The body suppresses the support argument on these cell endpoints for readability. Fixed-support asymptotics require $\mathbb{P}(\widehat{\mathcal{C}}_{\mathrm{act}} = \mathcal{C}) \to 1$; Corollary 10 gives sufficient rate conditions. Boundary conventions apply only to cells known by design to have $\rho_x \in \{0, 1\}$; incidental empirical boundaries do not imply population point identification.

Proposition 9's sign-envelope retains endpoint noise and both branches at uncertain signs; singleton branches give ordinary linear inference.

A separately predeclared conservative alternative projects the fixed-scale primitive region (66):

$$\widehat{\mathcal{I}}_{a,1-\alpha}^{\mathrm{proj}} = \left[\inf_{\theta \in \Theta_{1-\alpha}} L_a(\theta), \sup_{\theta \in \Theta_{1-\alpha}} U_a(\theta) \right]. \tag{47}$$

We do not choose between intervals according to which certifies. Under Corollary 10's independent bounded source and variance conditions, at $r_n = \sqrt{\min_k n_k}$,

$$(\widehat{L}_a, \widehat{U}_a) \to_p (L_a(\mathcal{C}), U_a(\mathcal{C})), \qquad \liminf_n \inf_{P \in \mathcal{P}_n} \mathbb{P}_P\{[L_a(\mathcal{C}), U_a(\mathcal{C})] \subseteq \widehat{\mathcal{I}}_{a,1-\alpha}\} \geq 1 - \alpha. \tag{48}$$

All source sizes must grow; partial support and vanishing independent audit/passive ratios are allowed under the stated conditions. Ordering requires the outer set to exclude zero. For reward-selective success, use Proposition 5's residual endpoints; the primary bootstrap cannot be used without a valid extended primitive bootstrap. Any pre-specified valid audit-compatibility-test rejection is reported separately: no test-based conditional coverage is claimed.

C.2 Influence function of an expectile

For $\tau \in (0, 1)$, define

$$\psi_\tau(r, e) = \tau(r - e)_+ - (1 - \tau)(e - r)_+, \tag{49}$$

$$\kappa_\tau(e) = \tau \mathbb{P}(R > e) + (1 - \tau)\mathbb{P}(R < e). \tag{50}$$

Away from an atom at e_τ, the expectile functional has influence function

$$\mathrm{IF}_\tau(r; F) = \frac{\psi_\tau(r, e_\tau)}{\kappa_\tau(e_\tau)}. \tag{51}$$

Indeed, for $F_\varepsilon = (1 - \varepsilon)F + \varepsilon \delta_r$ and $e_\varepsilon = e_\tau(F_\varepsilon)$,

$$0 = \frac{d}{d\varepsilon} \mathbb{E}_{F_\varepsilon} \psi_\tau(R, e_\varepsilon) \Big|_{\varepsilon = 0} = \psi_\tau(r, e_\tau) - \kappa_\tau(e_\tau)\dot{e}_0,$$

which gives (51). In each unaudited cell, the endpoint ℓ_x or u_x combines the empirical response proportion, observed-reward mean, and respondent expectile. We aggregate their empirical influence functions with the occupancy influence function and cluster by trajectory (or user).

For completeness, let $Y^{\mathrm{pas}} = (X, O, OR)$, let $p_x = \mathbb{P}(X = x)$, and let $e_{\tau,x}$ denote the respondent expectile. The passive influence function of the unaudited cell endpoint $\theta_x = \rho_x \mu_{1x} + (1 - \rho_x) e_{\tau,x}$ is

$$\phi_{\theta_x}(Y^{\mathrm{pas}}) = \frac{\mathbb{I}\{X = x\}}{p_x} \left[(O - \rho_x)(\mu_{1x} - e_{\tau,x}) + O(R - \mu_{1x}) + \frac{1 - \rho_x}{\rho_x} O \ \mathrm{IF}_\tau(R; F_{1x}) \right]. \tag{52}$$

For a randomized confirmation audit within the inference frame, let Z^C denote confirmation ascertainment, let $Y^{2\mathrm{ph}} = (X, O, OR, Z^C, Z^C R)$, and let $\zeta_x = \mathbb{P}(Z^C = 1 \mid O = 0, X = x)$ be fixed before confirmation outcomes. Under (9), the audited cell-mean influence function is

$$\phi_{r_x}^{2\mathrm{ph}}(Y^{2\mathrm{ph}}) = \frac{\mathbb{I}\{X = x\}}{p_x}\left[(O - \rho_x)(\mu_{1x} - \mu_{0x}) + O(R - \mu_{1x}) + (1 - O)\frac{Z^C}{\zeta_x}(R - \mu_{0x})\right]. \tag{53}$$

For simple random sampling without replacement, the corresponding Horvitz–Thompson or Hájek design score and finite-population correction are used. If the audit sample is externally independent, its variance contribution may instead be added separately. Cluster resampling must preserve the two-phase design that generated the estimator. Equation (53) concerns outcome-independent ascertainment. Selective-success endpoints from Proposition 5 require a verified joint Gaussian primitive approximation satisfying Proposition 9, or separate program-specific inference.

For a contrast, let $\phi_{c,x} = \sum_j a_j \phi_{d,jx}$. If $c_x \neq 0$, define

$$q_x^L = \begin{cases} \ell_x, & c_x > 0, \\ u_x, & c_x < 0, \end{cases} \qquad q_x^U = \begin{cases} u_x, & c_x > 0, \\ \ell_x, & c_x < 0. \end{cases} \tag{54}$$

At an audited cell set $q_x^L = q_x^U = r_x^{\mathrm{aud}}$. The contrast-endpoint influence functions are

$$\phi_L = \sum_x \{q_x^L \phi_{c,x} + c_x \phi_{q^L,x}\}, \qquad \phi_U = \sum_x \{q_x^U \phi_{c,x} + c_x \phi_{q^U,x}\}. \tag{55}$$

Consistent cluster or trajectory variances for (55), using (53) on audited cells or the corresponding external-audit score, give the standard errors in (56).

For a fixed finite partition, fixed or independently selected support, and $|c_x|$ bounded away from zero whenever $\Delta_x > 0$, remove deterministic endpoints and handle them directly. Assuming every remaining endpoint asymptotic variance is positive, define the studentized endpoint-expanded confidence set

$$\widehat{\mathcal{I}}_{a,1-\alpha}^{\mathrm{out}} = \left[\widehat{L}_a - z_{1-\alpha/2}\widehat{s}_L, \widehat{U}_a + z_{1-\alpha/2}\widehat{s}_U\right]. \tag{56}$$

C.3 Cross-fitting and simultaneous policy comparison

For flexible state aggregation, independent trajectories or users may be divided into design and inference splits. Audit support is chosen on the design split; fixed-partition endpoint scores are evaluated on the inference split. In the finite tabular model, state occupancy is the smooth map $(1-\gamma)\nu^\top(I - \gamma P^\pi)^{-1}$ of the initial law (when estimated) and transition rows, followed by multiplication by the fixed action probabilities; its influence function therefore follows from the ordinary delta method. Cross-fitting learned nuisance functions can reduce overfitting, but without an orthogonal score and explicit nuisance-rate conditions it does not by itself imply the root-sample-size coverage result in (68).

For several contrasts with uniform sign separation—$|c_x(a)|$ bounded away from zero whenever $\Delta_x > 0$—a trajectory multiplier bootstrap may draw mean-zero, unit-variance multipliers and take the $(1-\alpha)$ quantile of the maximum studentized regular endpoint deviation. Expanding every lower endpoint downward and every upper endpoint upward by this common critical value gives asymptotic simultaneous outer coverage. This ordinary centered-endpoint bootstrap is not used when sign separation fails.

When $c_x = 0$ and $\Delta_x > 0$, the endpoint map is Hadamard directionally, but not ordinarily, differentiable. Write $m_x = (\ell_x + u_x)/2$, $h_x = (u_x - \ell_x)/2$, and

$$L_a(\theta) = \sum_x \{c_x m_x - |c_x| h_x\}, \qquad U_a(\theta) = \sum_x \{c_x m_x + |c_x| h_x\}.$$

For a primitive direction $z_x = (z_{c,x}, z_{\ell,x}, z_{u,x})$, set $z_{m,x} = (z_{\ell,x} + z_{u,x})/2$ and $z_{h,x} = (z_{u,x} - z_{\ell,x})/2$. Then

$$L'_{a,\theta}(z) = \sum_x \{m_x z_{c,x} + c_x z_{m,x} - h_x \dot{a}_{c_x}(z_{c,x}) - |c_x| z_{h,x}\},$$

$$U'_{a,\theta}(z) = \sum_x \{m_x z_{c,x} + c_x z_{m,x} + h_x \dot{a}_{c_x}(z_{c,x}) + |c_x| z_{h,x}\}, \tag{57}$$

where $\dot{a}_c(z) = \mathrm{sign}(c)z$ for $c \neq 0$ and $\dot{a}_0(z) = |z|$. Thus an ordinary centered n-out-of-n endpoint bootstrap need not be consistent at the kink (Fang & Santos, 2019).

Our primary inference procedure is the screened sign-envelope below. It retains uncertainty in both the occupancy coefficients and reward endpoints, including on cells whose signs are uncertain.

Proposition 9 (Screened sign-envelope outer band)**.** *For this subsection, write $\mathcal{X} = \mathcal{X}_\Pi$. Condition on an independent design/pilot sigma-field. Fix the finite cell set, policies, audit support, and confirmation allocation; use only the disjoint inference observations and frozen confirmation audits. Write $\theta = (c_x, \ell_x, u_x)_{x\in\mathcal{X}}$, retaining $\ell_x = u_x$ on audited cells, and suppose $0 \le \ell_x \le u_x \le 1$. Uniformly over a triangular array $P \in \mathcal{P}_n$, suppose that, for a conditional-deterministic $r_n \to \infty$, $Z_n = r_n(\widehat{\theta} - \theta_{n,P})$ and its conditional multiplier analogue Z_n^* admit a common moving Gaussian approximation $Z_{n,P} \sim N(0, \Sigma_{n,P})$, with uniformly bounded covariance. Precisely, writing d_{BL} for bounded-Lipschitz distance and $\mathcal{L}_P^*$ for conditional law given the inference observations, for every $\varepsilon > 0$,*

$$\sup_{P\in\mathcal{P}_n} d_{\mathrm{BL}}\{\mathcal{L}_P(Z_n), \mathcal{L}(Z_{n,P})\} \longrightarrow 0,$$
$$\sup_{P\in\mathcal{P}_n} \mathbb{P}_P[d_{\mathrm{BL}}\{\mathcal{L}_P^*(Z_n^*), \mathcal{L}(Z_{n,P})\} > \varepsilon] \longrightarrow 0. \tag{58}$$

Select the true lower branch $q_x^L = \ell_x$ if $c_x \ge 0$ and $q_x^L = u_x$ otherwise, and the true upper branch $q_x^U = u_x$ if $c_x \ge 0$ and $q_x^U = \ell_x$ otherwise. Define their outward linear maps by

$$H_L(z) = \sum_x (q_x^L z_{c,x} + c_x z_{q^L,x}), \qquad H_U(z) = -\sum_x (q_x^U z_{c,x} + c_x z_{q^U,x}).$$

Assume $\mathrm{Var}\{H_j(Z_{n,P})\} \ge \underline{\sigma}^2 > 0$ uniformly for $j = L, U$. Choose a pre-specified $a_n \downarrow 0$ such that

$$\sup_{P\in\mathcal{P}_n} \mathbb{P}_P(\|\widehat{c} - c_{n,P}\|_\infty \ge a_n) \longrightarrow 0. \tag{59}$$

For the lower endpoint retain the branch labels

$$\widehat{B}_{L,x} = \begin{cases} \{\ell\}, & \widehat{c}_x > a_n, \\ \{u\}, & \widehat{c}_x < -a_n, \\ \{\ell, u\}, & |\widehat{c}_x| \le a_n, \end{cases}$$

and define $\widehat{B}_{U,x}$ by exchanging the two singleton labels. For $b \in \{\ell, u\}$, let $\widehat{q}_{x,b}$ and $z_{q_b,x}$ denote the corresponding endpoint estimate and primitive direction. Set

$$G_L^*(z) = \sum_x \max_{b\in\widehat{B}_{L,x}} \{\widehat{q}_{x,b} z_{c,x} + \widehat{c}_x z_{q_b,x}\},$$
$$G_U^*(z) = \sum_x \max_{b\in\widehat{B}_{U,x}} \{-\widehat{q}_{x,b} z_{c,x} - \widehat{c}_x z_{q_b,x}\}. \tag{60}$$

Let $q_{L,n}^*, q_{U,n}^*$ be the conditional upper $(1-\alpha/2)$ quantiles of $G_L^*(Z_n^*)$ and $G_U^*(Z_n^*)$, respectively. Then

$$\widehat{\mathcal{I}}_{a,1-\alpha}^{\mathrm{dir}} = \left[\widehat{L}_a - \frac{(q_{L,n}^*)_+}{r_n}, \widehat{U}_a + \frac{(q_{U,n}^*)_+}{r_n}\right] \tag{61}$$

satisfies

$$\liminf_n \inf_{P\in\mathcal{P}_n} \mathbb{P}_P\Big\{[L_a(\theta_{n,P}), U_a(\theta_{n,P})] \subseteq \widehat{\mathcal{I}}_{a,1-\alpha}^{\mathrm{dir}}\Big\} \ge 1 - \alpha. \tag{62}$$

No separation from the screening threshold is required. Exact and local-to-zero occupancy ties are allowed whenever the stated approximation and variance conditions hold. If all estimated signs clear the threshold, the formula reduces to ordinary Gaussian outward derivatives.

Proof. All probability bounds and remainders below are uniform over $P \in \mathcal{P}_n$. Let $b_{L,x}$ minimize $c_x q_{x,b}$ and let $b_{U,x}$ maximize it, using ℓ and u, respectively, when $c_x = 0$. For any (A_b) and (B_b),

$$\min_b A_b - \min_b B_b \le A_{b_L} - B_{b_L}, \qquad \max_b B_b - \max_b A_b \le B_{b_U} - A_{b_U},$$

where $b_L \in \arg\min_b B_b$ and $b_U \in \arg\max_b B_b$. Take $A_b = \widehat{c}_x \widehat{q}_{x,b}$ and $B_b = c_x q_{x,b}$. Writing $\delta c_x = \widehat{c}_x - c_x$ and $\delta q_{x,b} = \widehat{q}_{x,b} - q_{x,b}$ gives

$$A_b - B_b = q_{x,b}\delta c_x + c_x \delta q_{x,b} + \delta c_x \delta q_{x,b}.$$

Uniform tightness in (58) implies

$$r_n \sum_x \max_b |\delta c_x \delta q_{x,b}| = O_P(r_n^{-1}) = o_P(1).$$

Consequently,

$$r_n(\widehat{L}_a - L_a) \le H_L(Z_n) + o_P(1), \qquad r_n(U_a - \widehat{U}_a) \le H_U(Z_n) + o_P(1).$$

On $A_n = \{\|\widehat{c} - c\|_\infty < a_n\}$, every true label $b_{j,x}$ lies in $\widehat{B}_{j,x}$: excluding it would require an estimated coefficient of the opposite sign with magnitude exceeding a_n. Thus, for every bootstrap draw,

$$G_j^*(Z_n^*) \ge \widehat{H}_j(Z_n^*) \quad \text{on } A_n, \qquad q_{j,n}^* \ge q_{\widehat{H}_j,n}^*,$$

where $\widehat{H}_j$ uses the true labels and estimated coefficients, and $q_{\widehat{H}_j,n}^*$ is its conditional upper $(1-\alpha/2)$ quantile. The common approximation, coefficient consistency, and the scalar Gaussian variance lower bound imply

$$\sup_{P \in \mathcal{P}_n} \mathbb{P}_P\{H_j(Z_n) + o_P(1) > q_{\widehat{H}_j,n}^*\} \le \alpha/2 + o(1).$$

Since $\sup_P \mathbb{P}_P(A_n^c) \to 0$, positive-part critical values and a union bound over $j = L, U$ prove (62). □

The maximum over retained branches is part of a single pre-specified procedure; it does not choose a reported interval according to whether it certifies. For Q pre-specified contrasts, using level α/Q for each contrast yields simultaneous whole-set coverage by a union bound. Pointwise versions of the approximation and positive true-branch variance conditions give the corresponding pointwise statement.

Conservatism at ties. At $c_x = 0$, the lower endpoint derivative is $m_x z_{c,x} - h_x|z_{c,x}|$, whereas the retained-branch upper envelope is $m_x z_{c,x} + h_x|z_{c,x}|$. This deliberate domination ensures uniform outer coverage, not exact nominal coverage or shortest intervals. Its first-order slack can persist along ties; Table 14 quantifies the coverage–length trade-off.

Corollary 10 (Independent bounded sources at unequal sample sizes)**.** *Condition on the independent design/pilot sigma-field, and retain the fixed finite cell set, policies, audit support, and allocation of Proposition 9. For each of a fixed number K of independent inference sources, let E_{ki}, $i \le n_k$, be row-wise i.i.d. sampling units, independent across k. The source sizes are fixed after conditioning on the design. Keep $1 \le \Gamma < \infty$ fixed. Set*

$$m_n = \min_{k \le K} n_k \longrightarrow \infty, \qquad r_n = \sqrt{m_n}.$$

No lower bound on $n_k / \sum_h n_h$ is imposed. A unit may be an episode of length at most $\bar{H}$; a passive or independent confirmation record is the special case $\bar{H} = 1$. Episode boundaries are fixed by the sampling design, not by unobserved rewards or transitions. Suppose uniformly over $P \in \mathcal{P}_n$: rewards lie in $[0,1]$, $\gamma \le 1 - \epsilon$; every probability in a transition, response, or frozen-confirmation denominator within its source *is at least ϵ; and, for $\mathcal{T} = \{\tau_-, \tau_+\}$,*

$$\lim_{\delta \downarrow 0} \sup_n \sup_{P \in \mathcal{P}_n} \max_{x, \tau \in \mathcal{T}} \mathbb{P}_P(|R - e_{\tau,x,P}| \le \delta \mid O = 1, X = x) = 0. \tag{63}$$

Use the empirical initial-law (when unknown), transition, response, mean, expectile, occupancy, and outcome-independent confirmation plug-ins above. If activation uses successful-confirmation minima, assume on every planned cell that the planned attempt count $n_x^{\text{attempt}} \to \infty$, $n_{x,\min} = o(n_x^{\text{attempt}})$, and success probability is at least ϵ.

Let $\widehat{\psi}_{ki}$ be the source-k sampling-unit influence contribution and let $\bar{\widehat{\psi}}_k = n_k^{-1} \sum_i \widehat{\psi}_{ki}$. Draw $\xi_{ki} \overset{\text{iid}}{\sim} N(0,1)$ independently of all data and set

$$Z_n^* = \sum_{k=1}^K \sqrt{\frac{m_n}{n_k}} \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} \xi_{ki}(\widehat{\psi}_{ki} - \bar{\widehat{\psi}}_k), \qquad \widehat{\theta}^* = \widehat{\theta} + m_n^{-1/2} Z_n^*. \tag{64}$$

For the true-branch maps H_L, H_U in Proposition 9, assume

$$0 < \underline{\sigma}^2 \le \mathrm{Var}\{H_j(Z_{n,P})\}, \qquad j \in \{L, U\}, \tag{65}$$

uniformly over n, P, where $Z_{n,P}$ is the Gaussian influence-function reference constructed below. Then $\sqrt{m_n}(\widehat{\theta} - \theta_{n,P})$ and Z_n^* satisfy (58). Under (59), (62) holds. Audit support may leave positive identification width, and independent confirmation sources may be a vanishing fraction of the passive sample.

Proof. Bounded-unit concentration and $|\mathcal{X}| < \infty$ give, for a constant $c > 0$ depending only on the uniform bounds,

$$\sup_{P \in \mathcal{P}_n} \mathbb{P}_P(\widehat{\mathcal{C}}_{\mathrm{act}} \neq \mathcal{C}) \le \sum_{x \in \mathcal{C}} \exp\{-c n_x^{\mathrm{attempt}} / \bar{H}\} \longrightarrow 0.$$

Uniform no-atom and positivity give the regular expectile expansion from (51); its Jacobian obeys $\kappa_{\tau,x} \ge \min\{\tau, 1-\tau\}$. For occupancies, write

$$R_\pi = (I - \gamma P^\pi)^{-1}, \qquad dR_\pi[H] = \gamma R_\pi H^\pi R_\pi, \qquad \|R_\pi\|_\infty \le (1-\gamma)^{-1} \le \epsilon^{-1}.$$

These formulas and the regular empirical ratio expansions give

$$\sqrt{m_n}(\widehat{\theta} - \theta_{n,P}) = \sum_{k=1}^{K} \frac{\sqrt{m_n}}{n_k} \sum_{i=1}^{n_k} \psi_{k,P}(E_{ki}) + o_P(1),$$

where the true influence contributions are uniformly bounded and mean zero, and

$$\max_{k \le K} \frac{1}{n_k} \sum_i \|\widehat{\psi}_{ki} - \psi_{k,P}(E_{ki})\|^2 = o_P(1).$$

All sourcewise remainders remain negligible because $m_n/n_k \le 1$. The total third absolute moment of the independent normalized summands is at most

$$C \sum_{k=1}^{K} \frac{m_n^{3/2}}{n_k^2} \le \frac{CK}{\sqrt{m_n}} \longrightarrow 0.$$

For bounded Lipschitz test functions, Gaussian smoothing at bandwidth δ has error $O(\delta)$ and third derivatives $O(\delta^{-2})$. Lindeberg replacement therefore has total error $O(\delta + \delta^{-2} m_n^{-1/2}) = o(1)$ with $\delta = m_n^{-1/6}$. This proves the uniform Gaussian approximation, without inverting the covariance, for

$$\Sigma_{n,P} = \sum_{k=1}^{K} \frac{m_n}{n_k} \mathbb{E}_P[\psi_{k,P} \psi_{k,P}^\top].$$

Conditionally, the multiplier is exactly $N(0, \widehat{\Sigma}_n)$, where

$$\widehat{\Sigma}_n = \sum_{k=1}^{K} \frac{m_n}{n_k} \frac{1}{n_k} \sum_i (\widehat{\psi}_{ki} - \bar{\widehat{\psi}}_k)(\widehat{\psi}_{ki} - \bar{\widehat{\psi}}_k)^\top.$$

Influence-function consistency and the bounded-unit law of large numbers imply $\widehat{\Sigma}_n - \Sigma_{n,P} = o_P(1)$ uniformly. Hence both laws have the same moving Gaussian reference. The true-branch variance lower bound supplies scalar Gaussian quantile continuity; Proposition 9 proves coverage. Activation failure has vanishing probability and does not change the result. □

Nondegeneracy and discrete rewards. Vanishing audit/passive ratios remain subject to (65). For independent cellwise confirmation means based on n_x successes, their contribution at the $\sqrt{m_n}$ scale is $\sum_{x \in \mathcal{C}} (m_n/n_x) c_x^2 (1-\rho_x)^2 \sigma_{0x}^2$. When $m_n/N_{\mathrm{pas}} \to 0$, auditing only zero-contrast cells makes this contribution vanish; merely including a nonzero-contrast cell need not suffice if $m_n/n_x \to 0$. Fixed-scale projection below permits zero endpoint variances only under its stated primitive approximation and nondegeneracy conditions.

Condition (63) is substantive for multivalued discrete rewards: at $\Gamma > 1$, an expectile can coincide with an internal support atom, invalidating the ordinary score (51). Such cases require separate

primitive inference. Uniform separation from support atoms suffices for fixed finite support; individual continuity alone does not establish the uniform condition. At $\Gamma = 1$, expectiles are means, whose ordinary score $R - \mathbb{E}R$ remains valid at atoms; the no-atom assumption is unnecessary for this special primitive.

The common primitive rate suffices for screening if $\sqrt{m_n}a_n \to \infty$. When every occupancy primitive is estimated from a passive source of size N_{pas}, its sharper root-N_{pas} rate instead permits $\sqrt{N_{\mathrm{pas}}}a_n \to \infty$; we use $a_n = N_{\mathrm{pas}}^{-2/5}$. The latter remains valid when $m_n/N_{\mathrm{pas}} \to 0$. No requirement that all positive-gain cells be audited is used here: unaudited cells keep their passive endpoints.

The unequal-source result concerns separately sampled confirmations with positive *within-source* denominators. A vanishing Bernoulli inclusion probability inside a single two-phase source can make its inverse-inclusion score unbounded and is not covered automatically. A fixed-positive-inclusion two-phase design may instead be treated with its bounded unit score in (53); without-replacement audits require their design-respecting expansion and bootstrap. The corollary does not cover the selective-success program in Proposition 5, whose active basis may change, or one continuing trajectory and other dependent, adaptive, or growing-dimensional regimes. Those settings require their own primitive approximation; learned nuisances additionally require orthogonal scores and explicit rate conditions.

Projection in (47) covers the entire identified set whenever its joint primitive region covers the truth. For a single fixed contrast, (c, ℓ, u) may replace (D, ℓ, u); projection accommodates zero coefficients but can be conservative.

A fixed-scale region that permits zero coordinate variances. Choose deterministic scales $0 < \underline{s} \le s_j \le \overline{s} < \infty$ before inference, and let $q^*_{1-\alpha}$ be the conditional upper $(1-\alpha)$ quantile of $M^*_n = \max_j |Z^*_{n,j}|/s_j$. Set

$$\Theta_{1-\alpha} = \left\{ \vartheta : |\vartheta_j - \widehat{\theta}_j| \le \frac{s_j q^*_{1-\alpha}}{\sqrt{m_n}} \text{ for every } j \right\} \cap \Theta_{\mathrm{structural}}, \tag{66}$$

where $\Theta_{\mathrm{structural}}$ imposes endpoint ordering, audited equality, and the occupancy constraints. The scales are fixed units, not estimated coordinate standard errors. Deterministic coordinates may be fixed at their known values. In addition to (58), it is enough that

$$\max_j \Sigma_{n,P,jj}/s_j^2 \ge \underline{\sigma}_0^2 > 0$$

uniformly; the other variances may vanish. This condition follows from (65) when the true-branch coefficient norms and the scales are uniformly bounded. Then $\liminf_n \inf_P \mathbb{P}_P(\theta_{n,P} \in \Theta_{1-\alpha}) \ge 1 - \alpha$.

Justification. The scaled covariance matrices range over a compact bounded subset of the positive semidefinite matrices with at least one diagonal bounded below. For each such matrix, $\max_j |Z_j|/s_j$ has no atom: at any positive value an atom would require a nondegenerate Gaussian coordinate to equal a fixed number, and at zero at least one nondegenerate coordinate must equal zero. Gaussian weak continuity and compactness imply a uniform continuity modulus for these distribution functions. Indeed, a contrary sequence of shrinking intervals with nonvanishing probability has, by tightness and covariance compactness, a subsequence converging to an atom of a limiting Gaussian maximum. This is impossible. The common-reference approximation transfers quantiles; intersection with structural constraints preserves coverage because the true parameter satisfies them. If all primitives are known without sampling error by design, their exact singleton is used instead. No claim for a common-rate sequence with every variance tending to zero is needed.

The same construction on the independent design split certifies support selection. Project its joint region through (38) after imposing $u_x \ge \ell_x$ and the occupancy constraints. Projection occurs before the absolute value, so the score interval remains valid at $c_x(a) = 0$. If simultaneous marginal ranges are $c_x \in [\underline{c}_x, \overline{c}_x]$ and $\Delta_x \in [\underline{\Delta}_x, \overline{\Delta}_x]$, a conservative closed form is

$$\underline{g}_x = \mathrm{dist}\{0, [\underline{c}_x, \overline{c}_x]\}\, \underline{\Delta}_x, \qquad \overline{g}_x = \max\{|\underline{c}_x|, |\overline{c}_x|\}\, \overline{\Delta}_x. \tag{67}$$

For exactly K equal-cost selections, strict separation $\min_{x \in S^{\mathrm{LCB}}} \underline{g}_x > \max_{y \notin S^{\mathrm{LCB}}} \overline{g}_y$ certifies exact support recovery on the same coverage event. Design-stage selection confidence and the independent

endpoint-inference claim can be combined using levels $\alpha_s + \alpha_I \leq \alpha$. These are fixed finite primitive results; growing cell collections require additional conditions. The sign-envelope and projection are reported as separately predeclared procedures, rather than selecting the narrower or certifying interval after observing the data.

For the regular sign-separated case, joint CLTs, diverging cell counts, and regular expectile derivatives give

$$\liminf \mathbb{P}\{\mathcal{I}_{\Gamma,a}(\mathcal{C}) \subseteq \widehat{\mathcal{I}}^{\mathrm{out}}_{a,1-\alpha}\} \geq 1-\alpha. \tag{68}$$

D Function Approximation and Continuous State Spaces

The identification theory is stated for finite $\mathcal{X}$ because audit support is an operational object. In large spaces, we form a pre-specified partition $\{\mathcal{X}_b\}_{b=1}^{B}$ using outcome-free (S, A, S') on an independent design split, with final membership a function of (S, A). If membership itself uses S', the cell variable must instead be redefined as (S, A, S') and weighted by its target transition occupancy. Conditional expectile regressions estimate $e_{\tau_\pm}(F_1 \mid X = x)$, and a fitted transition model or marginalized density-ratio method estimates the target occupancies. For a contrast a, the score of a candidate audit region is

$$\widehat{g}_b(a) = \int_{\mathcal{X}_b} \left| \sum_j a_j \, d\widehat{D}^{\pi_j} / d\widehat{\lambda}(x) \right| (1 - \widehat{\rho}(x)) \{\widehat{e}_{\tau_+}(x) - \widehat{e}_{\tau_-}(x)\} \, d\widehat{\lambda}(x), \tag{69}$$

for any dominating reference measure $\widehat{\lambda}$. Sharpness holds relative to the maintained conditional model or partition, not automatically under arbitrary within-bin heterogeneity. Our experiments freeze the partition before applying missingness.

Equation (69) scores resolution of the conditional reward means throughout the region; it is not generally the gain from learning one pooled missing-reward mean. Under a maintained bin-constant reward model, the bin score instead is $|\sum_j a_j D^{\pi_j}(\mathcal{X}_b)|\Delta_b$, with the absolute value outside the occupancy integral. Arbitrary pooling does not identify the conditional function. We claim neither nonparametric inference rates nor exact audit gains for heterogeneous aggregate bins; approximate sharing is treated separately in Section E.5.

E Empirical and Experimental Details

E.1 Real randomized follow-up studies

Both applications are horizon-one treatment comparisons. They test recovery from outcome nonresponse and its stability conditions; they do not test Markov occupancy estimation, sequential support targeting, or shared-cell cancellation. Indeed, the deterministic treat-all and treat-none policies occupy disjoint action cells, so the shared-law and marginal contrast sets coincide. The controlled MDP tests sequential targeting. KuaiRec separately tests shared-law cancellation and audit ranking with real outcomes under an imposed mask and constructed policy hierarchy.

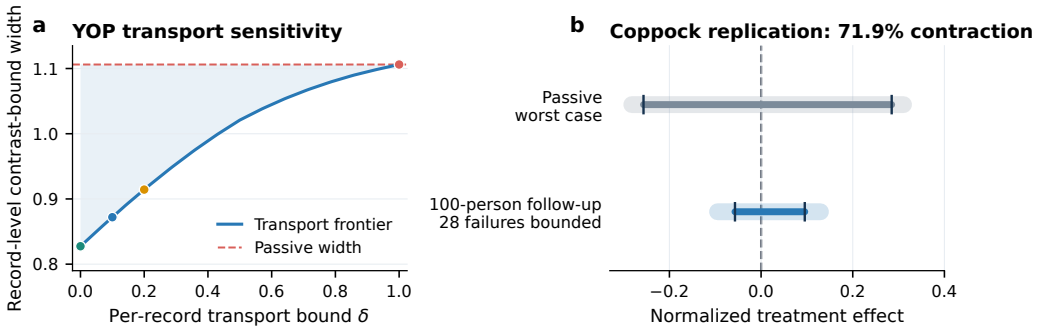


Figure 3: **Recovery can contract ambiguity, but its identifying conditions remain visible.** (a) In YOP, 362 later record-linked outcomes reduce finite-sample Manski width by 25.2% only under exact temporal transport; recordwise drift allowances restore the passive width at $\delta = 1$. (b) In an exact replication of Coppock et al. (2017), 100 randomized follow-up invitations, with all 28 failures bounded, reduce plug-in worst-case ATE-bound width by 71.9% under the original outcome-stability condition. Pale segments are descriptive 95% outer factor-bootstrap regions. Both applications have $H = 1$ and test outcome recovery, not sequential occupancy targeting.

E.1.1 YOP CASH-GRANT EXPERIMENT: MAPPING AND AUDIT BOUNDARY

We use the CC0 replication archive for Fiala et al. (2025), Dataverse DOI 10.7910/DVN/HU0BV1, file version 1, whose downloaded archive has MD5 `e282a1bb5cc2b7c9d96c2eb0f2cb2fc1` (Rose, 2026). Table 3 records the complete horizon-one mapping. The contrast is the original 2008 randomized grant assignment: $A = 1$ for grant treatment versus $A = 0$ for control. Higher food security is better; the primary bounded outcome is

$$R = 1 - \frac{\texttt{proport} + \texttt{reduceno}}{14} \in [0, 1], \tag{70}$$

and the binary diagnostic is one when both items are zero.

Table 3: YOP mapping to the one-step audit notation. A tilde marks the later-wave measurement.

Symbol	Archive field	Role
A	`assigned`	randomized cash-grant assignment arm
O	phase-one consent and food items	routine outcome observed
E	final phase-one status	final-status noncontact eligibility
I_{raw}	`selected2`	archived phase-two selection flag
I	`selected2` with $E = 1$	design-compatible archived selection
Z	phase-two consent and valid food items with $I = 1$	successful food-outcome ascertainment
R	phase-one `proport`, `reduceno`	routine reward when $O = 1$
$\widetilde{R}$	`second_short`/`second_long` items	later reward for a recovered record

Invitation frame. The primary target is the 2,598 tracking-sheet records with a valid assignment (1,276 control and 1,322 treated), rather than a frame obtained by deleting later status codes. It contains 1,162 routine food outcomes and 1,436 unresolved outcomes. In particular, all 46 refusals and 40 death codes remain in the target and contribute $[0, 1]$ unless recovered; this prevents post-assignment deletion, while making clear that a food-security interpretation for deceased participants requires a composite or hypothetical-outcome convention. The baseline roster has 2,677 records, so we report that alternative merge only as a frame sensitivity, not as the headline estimand.

Of the 1,436 records without a routine food outcome, 1,350 are final-status noncontacts (682 control and 668 treated); the other 46 refusals and 40 deaths remain support-bounded. The paper reports phase-two randomization stratified by assignment and district. The public archive has no frozen pre-draw roster, so the 28 assignment-by-district invitation rates are reconstructed from this final-status noncontact frame and used only in secondary design-weighted projections.

The round-two selection file contains 603 IDs. Five conflict with final phase-one status (three later appear complete, one refused, and one deceased), and the archive lacks the frozen pre-draw frame. All target records and conflict flags are preserved, but phase-two recovery is counted only when an archived selection is compatible with final noncontact status. This gives 598 design-compatible selections and 362 successful food-outcome recoveries. The primary recordwise bound does not extrapolate by the invitation randomization. No unflagged record is excluded. The release rule is therefore the archive's 2,598 assigned tracking records, with every refusal and death retained and bounded.

Direct recovery without extrapolation. The primary analysis does not impute from invitees to noninvitees, model audit success, or impose a finite Γ. For arm a, let $\mathcal{O}_a$ index routine phase-one outcomes, $\mathcal{J}_a$ successful record-linked recoveries, and $\mathcal{M}_a$ everyone still unresolved. If each recovered later outcome obeys $|R_i - \widetilde{R}_i| \le \delta$, define

$$L_a(\delta) = \frac{1}{N_a}\left\{\sum_{i\in\mathcal{O}_a} R_i + \sum_{i\in\mathcal{J}_a} \max(0, \widetilde{R}_i - \delta)\right\},$$
$$U_a(\delta) = \frac{1}{N_a}\left\{\sum_{i\in\mathcal{O}_a} R_i + \sum_{i\in\mathcal{J}_a} \min(1, \widetilde{R}_i + \delta) + |\mathcal{M}_a|\right\}. \tag{71}$$

The record-level interval for the realized treated-minus-control arm-mean contrast is exactly $[L_1(\delta) - U_0(\delta), U_1(\delta) - L_0(\delta)]$ under this transport restriction. The displayed object is the realized arm-mean contrast in the fixed published target; a design-based ITT interpretation additionally requires target-frame membership not to condition on post-assignment outcomes. The arithmetic is not a literal finite-population SATE set over both potential outcomes. Every noninvitee, refusal, failure, and other unresolved record retains its full range. Randomization governs who was offered recovery, but the validity of (71) does not extrapolate their outcomes to anyone else.

Passive Manski bounds are $[-0.549, 0.557]$, width 1.106. At $\delta = 0$, the 362 recoveries give $[-0.4187, 0.4088]$, width 0.8275: a 25.2% contraction. Widths at $\delta = 0.1$ and 0.2 are 0.872 and 0.914; $\delta = 1$ yields 1.106, exactly the passive width. A secondary design-weighted analysis over (Γ, Γ_A) is included as sensitivity analysis. Its separately projected and intersected restrictions are valid outer intervals but need not equal the sharp joint density-ratio program in Proposition 5; they are not the headline.

Table 4: YOP finite-sample width under recordwise temporal tolerance. Contraction is relative to passive Manski bounds; no row extrapolates from recovered to unresolved records.

Restriction	Width	Contraction
Passive; no recovered outcome transported	1.106140	0%
$\delta = 0$	0.827485	25.19%
$\delta = 0.1$	0.872014	21.17%
$\delta = 0.2$	0.914356	17.34%
$\delta = 1$	1.106140	0%

Time drift is not audit failure. The $\delta = 0$ row equates outcomes across waves. Among the 259 of 362 design-compatible recoveries with both dates, phase-two collection follows the recorded phase-one attempt by a median 55 days. All 53 same-phone diagnostic pairs have dated observations and a median gap of 41 days; normalized food security rises by 0.158 (standard error 0.050; 95% t interval $[0.058, 0.258]$) and within-person correlation is 0.231. Exact transport is therefore not credible by default. The δ frontier states the required per-person bound transparently; values 0.1 and 0.2 are sensitivity points, not confidence limits. The paired mean change neither bounds every individual's drift nor identifies arm-specific drift. Thus YOP shows how much direct recovery can buy under an explicit time-scale tolerance, not an unconditional point effect.

As an explicitly conditional falsification exercise, impose $\delta = 0$, $\Gamma_A = 1$, and the reconstructed design-compatible cohort. The plug-in binary respondent and audit laws require a maximum armwise routine sensitivity of at least 1.825; a 5,000-draw group-cluster bootstrap gives a descriptive one-sided 95% percentile lower endpoint of 1.361. Thus $\Gamma < 1.825$ fails the plug-in diagnostic, whereas

$\Gamma < 1.361$ lies below its descriptive bootstrap lower endpoint, under those extra conditions. Because temporal stability and ignorable audit success are strong here, this diagnostic neither calibrates the primary YOP analysis nor transports a finite Γ to unaudited cells; it demonstrates the falsification logic in Section 3.

E.1.2 COPPOCK ET AL.: EXACT REPLICATION AND THE OUTCOME-STABILITY BOUNDARY

We use the authors' CC0 archive, Dataverse DOI 10.7910/DVN/AQB4MP, and reproduce the Polarized-versus-Moderate comparison in Coppock et al. (2017). Of 1,980 Wave-1 participants, 1,444 provide the routine Wave-2 outcome and 536 do not. Within each arm, exactly 50 nonrespondents are randomly invited to a higher-incentive follow-up; 39 Moderate and 33 Polarized invitees respond. Routine Wave-2 collection paid \$1, whereas selected nonrespondents were offered \$4. The published design therefore requires outcome stability: the incentive and delay must not change the outcome that would have been recorded initially. The horizon-one mapping is $A = \texttt{Z1}$, $O = \texttt{R1}$, $I = \texttt{Attempt}$, $Z = \texttt{R2}$, and $R = \texttt{L_dif_w2} \in [0, 6]$ (divided by six in normalized plots).

Sixteen source rows contain a nonmissing outcome despite $O = Z = I = 0$. The authors' scripts exclude them from every response stratum. We retain the source value in a separate column, set a discrepancy flag, and mask it in the exact replication. Every estimate and variance in published Table 3 is reconstructed; the maximum discrepancy from its four-decimal entries is 4.54×10^{-5}.

Table 5: Coppock plug-in ATE bound intervals on the published 0–6 scale. The passive row uses bounded outcomes; the randomized-follow-up row additionally requires randomized invitation and outcome stability across the Wave-2 collection attempts.

Restriction	Lower	Upper	Width
Passive worst case	-1.539	1.710	3.249
Random audit; failures in $[0, 6]$	-0.342	0.572	0.914
Passive $\Gamma = 4$	-0.249	0.497	0.746
Audit-success $\Gamma_A = 2$ projection	0.104	0.215	0.111

Random assignment supplies the treatment-effect interpretation; the displayed numbers are estimated ATE bounds, not exact finite-population SATE sets. Conditional on outcome stability, randomized follow-up is $3.56\times$ narrower than the passive worst-case set, a 71.8802% contraction, while still crossing zero. Because the \$1/\$4 incentive change can itself alter responses, same-Wave-2 collection reduces but does not eliminate transport risk. The randomized-follow-up interval is slightly wider than the passive $\Gamma = 4$ row because the latter imposes a strong finite response-odds restriction; invitation randomization does not guarantee domination of every model-based interval. The $\Gamma_A = 2$ row is the follow-up-success projection; its descriptive 95% outer factor-bootstrap region is $[-0.093, 0.408]$. Intersecting it with a passive Γ projection is a valid outer bound, not generally the sharp joint LP, and we do not label it otherwise. A 5,000-draw design-aware factor bootstrap gives descriptive outer endpoint regions and is not substituted for the paper's fixed-tabular inference theorem.

E.2 CONTROLLED MARKOV ENVIRONMENT

The controlled continuing MDP has 20 states, four actions, and 80 audit cells. With seed 20260807, sticky-Dirichlet transition rows favor the current and action-indexed states; Bernoulli reward logits contain state, action, and interaction effects. We use $\pi(a \mid s) \propto \exp\{2.5r(s, a)\}$, behavior $0.55\pi + 0.45\,\mathrm{Unif}(\mathcal{A})$, a uniform initial law, and $\gamma = 0.95$. Finite linear systems give exact discounted occupancies and normalized value.

Routine response follows

$$\mathbb{P}(O = 1 \mid X = x, R = r) = \mathrm{logit}^{-1}\{\alpha_x - \log(4)r\}, \tag{72}$$

with α_x targeting rates $\{0.38, 0.45, 0.52, 0.59, 0.66\}$; the missing-to-observed density ratio reaches $\Gamma = 4$ between binary outcomes, so truth lies on the boundary. Population curves average 500 random support orders. Conditional inference uses 1,000 repetitions at 5, 10, 25, 50, 100, 250, and 500 independent audited rewards per selected cell, with either top-20 or full-80 support and Agresti–Coull variance stabilization at small counts. Figure 4 reports containment of the complete supported population set; bars are 95% binomial Monte Carlo intervals.

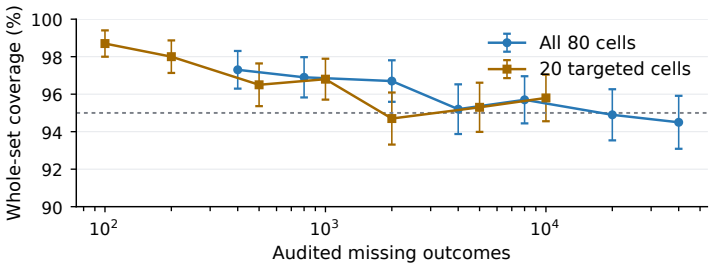


Figure 4: **Conditional whole-set coverage for Figure 2.** With the population fixed, audit rewards alone are resampled. Bars are pointwise 95% normal-approximation binomial Monte Carlo intervals over 1,000 repetitions; the dashed line is 95%. The complete sign-envelope procedure is evaluated separately in Tables 7–9.

For the static–sequential ablation, write $d_S^j(s) = \sum_a d^j(s,a)$ and $\pi_j(a \mid s) = d^j(s,a)/d_S^j(s)$. The static ranking score is

$$\frac{d_S^\pi(s) + d_S^\mu(s)}{2} \, |\pi(a \mid s) - \mu(a \mid s)| \, \Delta_{sa},$$

which omits policy-induced state visitation. The exact score $|d^\pi(s,a) - d^\mu(s,a)|\Delta_{sa}$ includes it; at budgets 20, 40, and 60 the static residual width is larger by 0.8%, 4.3%, and 46.9%.

Table 6: Controlled-MDP population results at $\Gamma = 4$. Set rows report identification widths with no audit or 20 equally costly activated cells; random width averages 500 orders.

Method	$\lvert\widehat{V} - V\rvert$	Truth in set	Width
Complete case	0.1321	–	–
MAR-IPW	0.1331	–	–
Oracle MNAR-IPW	0.0000	–	–
Unaudited Γ set	–	yes	0.2611
Random support (20)	–	yes	0.1962
Directional audit support (20)	–	yes	**0.1531**

Complete sign-envelope pipeline and decision-boundary controls. This is the primary inference experiment, implementing Algorithm 1 and Proposition 9. The target and reference policies, transition kernel, and known initial law are fixed across four reward laws. Besides the original positive contrast $\delta_0 = 0.0144709$, we replace its Bernoulli mean vector r^0 by $r_x^\delta = 1/2 + (\delta/\delta_0)(r_x^0 - 1/2)$ for $\delta \in \{-0.003, 0, 0.003\}$. Since $\sum_x c_x = 0$, the resulting contrast is $\sum_x c_x r_x^\delta = \delta$. All means lie in $[0,1]$; response intercepts are recomputed to preserve the response rates and $\Gamma = 4$.

Each of 200 independent repetitions uses 30,000 design records and 30,000 inference records. Only (X, S', O, OR) is passed to the estimator. The design split estimates comparison scores and chooses no audit, top 20, or all 80 cells. A separate pilot obtains 20 missing rewards per selected cell and estimates its variance. With k selected cells and frozen weights $w_x = |\widehat{c}_x|(1-\widehat{\rho}_x)\widehat{\sigma}_{0x}$, we allocate $n_x = 10 + (B - 10k)w_x / \sum_z w_z$, rounded by largest remainders. This is a minimum-plus-Neyman-weighted rule, not the exact minimum-constrained variance optimum. Confirmation budgets are $B = 2{,}000$ and $8{,}000$; pilot costs add 400 and 1,600 outcomes. These count successfully acquired missing rewards, not attempted contacts. Pilot outcomes are excluded from the final endpoints, and the design is frozen before independent confirmation.

All occupancies and reward primitives are re-estimated. The 999 Gaussian influence draws retain joint passive-sample covariance and independent cellwise confirmation variation. We use the sign-envelope rule with $a_n = 30{,}000^{-2/5} = 0.01619$ and separate 97.5% outward quantiles. Every coefficient is screened in every replication at this working point; thus it tests the full envelope, not the transition between screened and singleton branches. The predeclared comparison projects a 95% primitive rectangle with deterministic scales $s_c = 1/80$ and $s_q = 1$. Its computed outer projection drops

Table 7: Complete prescribed inference in the 80-cell MDP. Both methods use the fixed design detailed above (200 repetitions); coverage concerns the entire sharp set. MCSEs are in parentheses, and confirmations exclude pilot outcomes.

Support	Method	Coverage (MCSE)	CI width	Certify > 0	Confirmations
None	Sign envelope	100.0 (0.0)	0.0304	0.0	0
None	Primitive projection	100.0 (0.0)	0.2118	0.0	0
Learned top 20	Sign envelope	99.0 (0.7)	0.0154	100.0	2,000
Learned top 20	Primitive projection	100.0 (0.0)	0.2045	0.0	2,000
All 80	Sign envelope	93.0 (1.8)	0.0015	100.0	8,000
All 80	Primitive projection	100.0 (0.0)	0.2495	0.0	8,000

Table 8: Joint sample-size diagnostic of the unchanged 95% sign-envelope procedure, with all 80 cells audited. Each passive split has n records; B is the independent confirmation budget. Pilot/minimum counts are 10/5, 20/10, 40/20, 80/40 per cell. Coverage and false-sign rates are percentages; brackets give pointwise 95% Wilson Monte Carlo intervals. False sign certifies a positive (negative) contrast when truth is nonpositive (nonnegative).

Reward law	n	B	Reps	Cover [95% MC interval]	Width	False sign
Positive	15,000	4,000	500	96.2 [94.1, 97.6]	0.0021	0.0
Positive	30,000	8,000	1,000	95.2 [93.7, 96.4]	0.0015	0.0
Positive	60,000	16,000	500	95.6 [93.4, 97.1]	0.0010	0.0
Positive	120,000	32,000	500	95.4 [93.2, 96.9]	0.0007	0.0
Zero	15,000	4,000	500	94.2 [91.8, 95.9]	0.0022	5.8
Zero	30,000	8,000	1,000	95.6 [94.1, 96.7]	0.0016	4.4
Zero	60,000	16,000	500	95.2 [93.0, 96.8]	0.0011	4.8
Zero	120,000	32,000	500	95.0 [92.7, 96.6]	0.0008	5.0

coupled occupancy constraints, which can only widen the interval. Neither procedure is selected or tuned according to the observed certificates.

The all-80 coverage estimate in Table 7 has MCSE 1.8 percentage points and alone does not establish systematic undercoverage. Fixed-scale projection covers every set but never certifies at this working point. Both procedures are retained without outcome-based selection; the following study uses independent random streams.

Joint sample-size and Monte Carlo precision check. The new design is fixed before running any full diagnostic and leaves the confidence procedure, screening exponent, and allocation rule unchanged. We use $n \in \{15{,}000, 30{,}000, 60{,}000, 120{,}000\}$ records in each independent passive split. The all-80 confirmation budgets are respectively $4{,}000, 8{,}000, 16{,}000, 32{,}000$; pilot counts and per-cell confirmation minima also scale proportionally. The original positive and zero reward laws each receive $500, 1{,}000, 500, 500$ independent repetitions, with 999 Gaussian influence draws throughout. Thus there are 5,000 data repetitions, with no-audit, learned-top-20, and all-80 results archived for each. Top-20 uses one quarter of the all-80 budget; those two supports are not compared at equal confirmation cost. All samples consist of independent conditionally ascertained rewards and passive tuples; this study does not model a finite audit pool, recovery failures, or a continuing trajectory.

Table 8 reports all-80 results. Every coverage interval for Monte Carlo uncertainty contains 95%; the positive contrast's mean confidence width decreases from 0.00207 to 0.000725 across the grid. At $n = 30{,}000$, positive-effect coverage is 95.2% (MCSE 0.68 percentage points); null false-sign frequency is 4.4% with Wilson interval $[3.3, 5.9]\%$. Under the null and full support, false-sign certification is exactly noncoverage of the singleton zero. Increasing repetitions reduces Monte Carlo error; increasing passive and audit samples examines the statistical procedure. Empirical coverage need not approach 95% monotonically. These intervals are pointwise Monte Carlo intervals, not simultaneous or policy-value confidence bands. Full-support endpoints are smooth, so this check addresses sampling calibration; the local-kink experiment below addresses nonregularity.

Table 9: Decision-boundary checks with the same fixed policies and transition kernel. Only the reward law changes, giving true contrasts $-0.003, 0, +0.003$. Entries are percentages over 200 replications. False sign denotes either a positive certificate for a nonpositive contrast or a negative certificate for a nonnegative contrast.

Truth	Support	Method	Cover	Certify > 0	Certify < 0	False sign
+0.000	Learned top 20	Sign envelope	100.0	0.0	0.0	0.0
+0.000	Learned top 20	Primitive projection	100.0	0.0	0.0	0.0
+0.000	All 80	Sign envelope	95.5	2.0	2.5	4.5
+0.000	All 80	Primitive projection	100.0	0.0	0.0	0.0
-0.003	Learned top 20	Sign envelope	99.0	0.0	0.0	0.0
-0.003	Learned top 20	Primitive projection	100.0	0.0	0.0	0.0
-0.003	All 80	Sign envelope	95.0	0.0	100.0	0.0
-0.003	All 80	Primitive projection	100.0	0.0	0.0	0.0
+0.003	Learned top 20	Sign envelope	100.0	0.0	0.0	0.0
+0.003	Learned top 20	Primitive projection	100.0	0.0	0.0	0.0
+0.003	All 80	Sign envelope	95.5	100.0	0.0	0.0
+0.003	All 80	Primitive projection	100.0	0.0	0.0	0.0

At zero total effect and full support, positive and negative false certificates occur in 2% and 2.5% of repetitions. The two nonzero controls are correctly ordered with full support; top-20 support remains inconclusive in all three boundary controls. The public CSV retains all 24 combinations of reward law, support, and procedure, including no-audit results and primitive-region coverage. All retained source sizes must grow for Corollary 10; this finite experiment is not evidence for vanishing-probability inverse-weighted scores in one source.

Pointwise single-policy comparator. This comparator regenerates every stage with an ordinary bootstrap-SE expansion. Each of 200 repetitions uses independent design and inference splits of 30,000 i.i.d. discounted behavior-policy tuples. We estimate all transition rows, d^π, response rates, and respondent expectiles; audit 100 missing rewards per active cell; and hold independently selected support fixed across 499 bootstrap draws. Table 10 reports whole-set, not scalar-truth, coverage.

Table 10: End-to-end fixed-tabular inference over 200 complete repetitions. Parentheses give Monte Carlo standard errors for coverage.

Support	Set coverage	Mean true width	Mean confidence width
None	94.0% (1.7%)	0.2611	0.2738
Top 20	94.5% (1.6%)	0.1536	0.1663
All 80	94.5% (1.6%)	0	0.0125

These are i.i.d. occupancy tuples, not one trajectory. All 499 draws were retained; because percentile intervals undercovered at the boundary, we use the two one-sided 97.5% bootstrap-SE expansions in (56).

Pointwise centered-endpoint comparison. A second experiment targets the policy contrast. In each of 200 repetitions, 30,000 design tuples select support by $|\widehat{d}_x^\pi - \widehat{d}_x^\mu|\widehat{\Delta}_x$ without population scores or true means. A fresh 30,000-tuple split, 100 audits per active cell, and 499 bootstrap draws re-estimate all primitives and form a joint max-statistic endpoint band. Table 11 reports whole-advantage-set coverage under independently selected support.

All population contrasts are nonzero (minimum 1.51×10^{-5}); estimated signs nevertheless mismatch in 133 of 200 repetitions—one cell in 75 runs, two in 46, and three in 12. Each draw resamples tuples and, independently, binomial audit successes, then refits occupancies, signs, and endpoints with support frozen. The 95th percentile of the maximum endpoint-standardized deviation is the joint critical value. This is a fixed-DGP pointwise diagnostic; Proposition 9 handles exact and local zeros, while (47) certifies support scores conservatively.

Table 11: Pointwise centered-endpoint contrast inference. Parentheses are Monte Carlo standard errors; "Cert." is the fraction of confidence sets entirely above zero.

Support	Set coverage	Cert.	True width	95% set width
None	95.5% (1.5%)	0.0%	0.0270	0.0288
Top 1	95.0% (1.5%)	16.0%	0.0260	0.0278
Top 20	96.5% (1.3%)	100.0%	0.0128	0.0145
All 80	96.0% (1.4%)	100.0%	0	0.0016

Table 12: Equal-support population audit design on the unchanged 30-case MDP grid (ten seeds, three policy mixtures). Entries are relative remaining-width regret $(W_{\mathrm{rule}} - W_\star)/W_\star$, in percent: median [minimum, maximum]. Every rule uses the same shared reward model and activates exactly K of 80 cells at unit activation cost. Random entries use the exact expected width of uniformly random support. These calculations evaluate identification width, not learned support or finite-sample inference.

Audit rule	$K = 20$	$K = 40$
Comparison-specific (oracle)	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Local-action disagreement	0.8 [0.0, 2.8]	5.8 [2.6, 8.3]
Occupancy difference only	7.1 [2.9, 10.7]	10.4 [1.6, 28.3]
Marginal-targeted	30.4 [18.1, 39.5]	93.2 [53.1, 123.9]
Ambiguity only	44.8 [37.2, 57.4]	96.8 [72.6, 134.0]
Uniform random (expected)	64.2 [55.5, 79.5]	151.0 [136.8, 205.9]

Estimated top-20 support overlaps 18.825 oracle cells on average (Jaccard 0.891) and loses 0.42% of removable width. Although the population top cell certifies, its estimate selects that cell in 47% of repetitions and certifies in only 16.0%. A local-action top-20 design has coverage 96.0% (MCSE 1.4%) and width 0.0146.

Same-model support selection across the full MDP grid. To separate support selection from shared-law tightening, we retain all ten seeds and three behavior-policy mixtures of the existing population sweep. Every rule uses the same reward intervals and activates exactly $K = 20$ or 40 of the 80 cells. The local-action proxy multiplies average discounted state occupancy by within-state action disagreement and reward width; it omits the change in state visitation between policies. Uniform-random support is integrated exactly, giving expected remaining width $(1 - K/80)W(\varnothing)$. Table 12 reports relative regret against the additive cardinality oracle. This calculation assumes population quantities and equal activation costs; it does not test learned rankings or recruitment attempts. Main-text percentage reductions use the comparator width as denominator; the table's regret instead uses the oracle width.

Held-out primitive projection. This comparator projects a studentized rectangular primitive region through (47); it is distinct from the fixed-scale alternative above. A calibration replay (IDs 0–199) found only 84.5–91.5% coverage for a nominal-95% high-dimensional box. We therefore fixed a 99%-nominal rule and evaluated it once on 200 disjoint held-out repetitions (IDs 10000–10199; 1,999 draws each). Primitive coverage was 98.5–100% and every projected set covered, but projection was much wider and never certified. These are fixed-DGP diagnostics, not finite-sample exact or uniform guarantees.

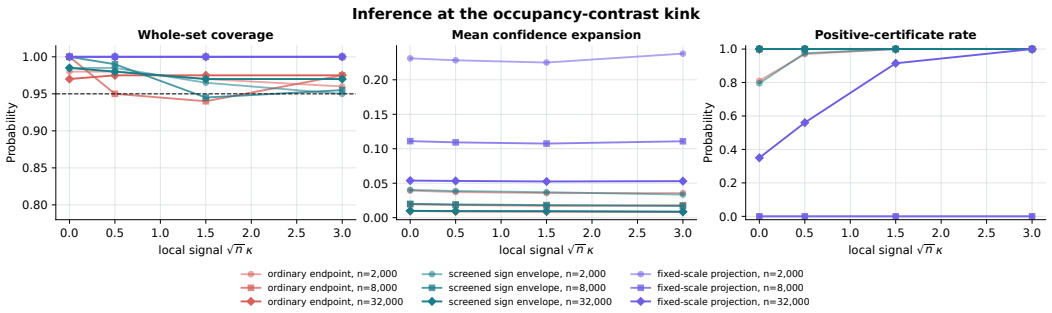


Figure 5: **Inference at the occupancy-contrast kink.** Lines vary the local signal $\sqrt{n}\,\kappa$ and marker shapes denote the size of each of the three independent samples. The screened sign-envelope band has 94.5–100% whole-set coverage and requires 14.1–18.6% of fixed-scale projection's confidence expansion in this diagnostic.

Table 13: Held-out paired end-to-end comparison over 200 independent replications (IDs 10000–10199). Primitive regions use 1,999 bootstrap draws; the pointwise comparator retains its original first 499 draws. "Primitive cover" is joint coverage of the full contrast-sufficient primitive vector (c, ℓ, u), available because this is a simulation; "set cover" is containment of the entire true identified set, and "Cert." is the fraction of reported contrast sets entirely above zero. The 99%-nominal box was fixed after a separate calibration replay and is a deliberately conservative diagnostic, not a claim of finite-sample exactness. Parentheses give Monte Carlo standard errors.

Support	Method	Primitive cover	Set cover	Mean width	Cert.
None	Endpoint sup-t (95%)	–	0.965 (0.013)	0.028832	0.010
	Primitive box (95%)	0.930 (0.018)	1.000 (0.000)	0.065695	0.000
	Primitive box (99%)	1.000 (0.000)	1.000 (0.000)	0.070542	0.000
Top 1	Endpoint sup-t (95%)	–	0.965 (0.013)	0.027791	0.205
	Primitive box (95%)	0.930 (0.018)	1.000 (0.000)	0.064634	0.000
	Primitive box (99%)	1.000 (0.000)	1.000 (0.000)	0.069471	0.000
Top 20	Endpoint sup-t (95%)	–	0.970 (0.012)	0.014523	1.000
	Primitive box (95%)	0.930 (0.018)	1.000 (0.000)	0.051884	0.000
	Primitive box (99%)	0.990 (0.007)	1.000 (0.000)	0.056822	0.000
All 80	Endpoint sup-t (95%)	–	0.960 (0.014)	0.001577	1.000
	Primitive box (95%)	0.920 (0.019)	1.000 (0.000)	0.038778	0.000
	Primitive box (99%)	0.985 (0.009)	1.000 (0.000)	0.043436	0.000

Local-to-zero stress test. The kink stress test uses four Bernoulli cells at $\Gamma = 3$, reference occupancy $(0.22, 0.28, 0.25, 0.25)$, contrast $(\kappa, -\kappa, 0.12, -0.12)$, respondent means $(0.80, 0.25, 0.70, 0.30)$, and response rate 0.65. "Exact zero" concerns only the first pair; the total contrast remains positive. For $n \in \{2{,}000, 8{,}000, 32{,}000\}$, three independent size-n samples estimate both occupancies and the observed law. Across $\kappa \in \{0, 0.5/\sqrt{n}, 1.5/\sqrt{n}, 3/\sqrt{n}, 0.06\}$, we compare centered endpoint sup-t, the screened band with $a_n = n^{-0.4}$, and fixed-scale primitive projection ($s_c = 1/4$, $s_q = 1$). All procedures use the same nonparametric primitive bootstrap with 200 repetitions and 399 draws per condition; coverage refers to the full sharp set. This diagnostic retains endpoint noise in screened cells; the complete MDP pipeline above uses Gaussian influence multipliers.

Sign-envelope whole-set coverage is 94.5–100% (MCSE at most 1.6 points) and its expansion is 14.1–18.6% of fixed-scale projection's. At exact zero, projection certifies in 0%, 0%, and 35% as n grows, versus 79.5%, 100%, and 100% for the sign-envelope band. The theorem, not this finite-sample diagnostic, supplies the asymptotic claim.

Independent-episode end-to-end stress test. The preceding i.i.d.-tuple experiment does not reproduce serial dependence. We therefore keep the fixed policies, reward law, boundary MNAR mechanism, and initial law unchanged and generate 400 independent behavior-policy episodes of

Table 14: Local-to-zero inference for the entire sharp contrast set; n is the size of each of three independent samples. "Exact zero" refers to the first coefficient pair, not the total contrast. Parentheses after coverage are Monte Carlo standard errors.

n	κ	Method	Coverage	Width expansion	Certify > 0
2,000	Exact zero	Endpoint sup-t	98.0 (1.0)	0.0393	81.0
2,000	Exact zero	Screened sign envelope	98.5 (0.9)	0.0404	79.5
2,000	Exact zero	Fixed-scale projection	100.0 (0.0)	0.2311	0.0
2,000	$1.5/\sqrt{n}$	Endpoint sup-t	97.0 (1.2)	0.0357	100.0
2,000	$1.5/\sqrt{n}$	Screened sign envelope	96.5 (1.3)	0.0369	100.0
2,000	$1.5/\sqrt{n}$	Fixed-scale projection	100.0 (0.0)	0.2250	0.0
2,000	Fixed 0.06	Endpoint sup-t	95.5 (1.5)	0.0376	100.0
2,000	Fixed 0.06	Screened sign envelope	95.0 (1.5)	0.0358	100.0
2,000	Fixed 0.06	Fixed-scale projection	100.0 (0.0)	0.2390	0.0
8,000	Exact zero	Endpoint sup-t	100.0 (0.0)	0.0194	100.0
8,000	Exact zero	Screened sign envelope	100.0 (0.0)	0.0201	100.0
8,000	Exact zero	Fixed-scale projection	100.0 (0.0)	0.1111	0.0
8,000	$1.5/\sqrt{n}$	Endpoint sup-t	94.0 (1.7)	0.0173	100.0
8,000	$1.5/\sqrt{n}$	Screened sign envelope	94.5 (1.6)	0.0183	100.0
8,000	$1.5/\sqrt{n}$	Fixed-scale projection	100.0 (0.0)	0.1075	0.0
8,000	Fixed 0.06	Endpoint sup-t	99.0 (0.7)	0.0181	100.0
8,000	Fixed 0.06	Screened sign envelope	96.5 (1.3)	0.0170	100.0
8,000	Fixed 0.06	Fixed-scale projection	100.0 (0.0)	0.1180	0.5
32,000	Exact zero	Endpoint sup-t	97.0 (1.2)	0.0097	100.0
32,000	Exact zero	Screened sign envelope	98.5 (0.9)	0.0100	100.0
32,000	Exact zero	Fixed-scale projection	100.0 (0.0)	0.0538	35.0
32,000	$1.5/\sqrt{n}$	Endpoint sup-t	97.5 (1.1)	0.0089	100.0
32,000	$1.5/\sqrt{n}$	Screened sign envelope	97.0 (1.2)	0.0094	100.0
32,000	$1.5/\sqrt{n}$	Fixed-scale projection	100.0 (0.0)	0.0525	91.5
32,000	Fixed 0.06	Endpoint sup-t	96.0 (1.4)	0.0089	100.0
32,000	Fixed 0.06	Screened sign envelope	95.0 (1.5)	0.0083	100.0
32,000	Fixed 0.06	Fixed-scale projection	100.0 (0.0)	0.0586	100.0

length 75 per split under two transition kernels. The uniform-refresh kernel has absolute spectral gap 0.909, whereas the self-loop kernel has gap 0.022. Changing the transition kernel also changes discounted occupancies and hence the numerical contrast estimand slightly; this is a comparison of two otherwise matched DGPs, not a causal isolation of serial dependence.

Each of 200 replications uses 30,000 transitions in an independent design split and 30,000 fresh transitions in the inference split. Design episodes select the top 20 cells. Each selected cell then supplies exactly 100 successfully ascertained missing-reward outcomes under the idealized outcome-independent audit-success model—these are successful outcomes, not 100 invitation attempts. Mean-one exponential multipliers reweight whole inference episodes, while each draw independently bootstraps the successful audit samples; transitions, both occupancies, signs, and endpoints are refit. The sampling unit is therefore an independent fixed-length episode, not one continuing Markov trajectory.

Table 15 reports the pointwise results. Top-20 coverage is 96.0% (1.4-point MCSE) in the fast DGP and 95.5% (1.5 points) in the slow DGP, with positive certification in all 200 replications. These fixed-DGP diagnostics concern many independent fixed-length episodes; they are not uniform over vanishing spectral gaps or local-to-zero occupancy contrasts and do not validate a block bootstrap for one continuing trajectory, adaptive logging, or outcome-dependent audit refusal.

Seed and overlap stress test. We also cross ten MDP seeds with behavior target-mixtures 0.25, 0.55, and 0.80 (30 exact populations). Every true advantage is positive, spanning 0.0047–0.0258. Marginal subtraction is $11.5\times$–$51.5\times$ wider than the shared-law interval (median $20.4\times$); zero, one, or at most 2 population support cells certify in every run. Twenty and forty cells remove on average

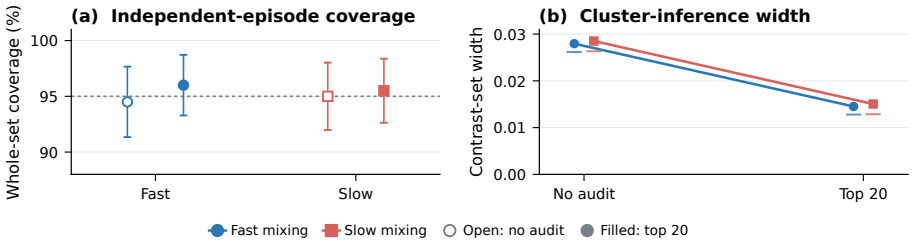


Figure 6: **Independent episodes preserve pointwise calibration under two fixed transition DGPs.** Each of 200 replications uses disjoint design and inference samples of 400 length-75 episodes; whole episodes are reweighted in every bootstrap draw. (a) Open and filled markers denote no audit and frozen top-20 support; Monte Carlo error bars use 1.96 standard errors. Top-20 coverage is 96.0% under fast mixing and 95.5% under slow mixing. (b) Points are mean simultaneous confidence widths and short ticks are the corresponding population sharp widths. This is not evidence for a block bootstrap on one continuing trajectory.

Table 15: Pointwise end-to-end inference from 400 independent behavior-policy episodes of length 75 per split, not one continuing trajectory. Transition kernels change occupancies, so sharp widths are reported separately. Whole episodes receive common multiplier weights. Top-20 rows use 100 successfully ascertained missing outcomes per selected cell under outcome-independent audit success (not 100 attempts). Parentheses are Monte Carlo standard errors over 200 replications. ESS is the median finite-episode effective count for $Z_t = \mu(S_t, A_t)$ out of 30,000; it is a signal diagnostic, not estimator-wide ESS.

Mixing	Audit	Gap	ESS	Set cover	Cert.	Sharp width	Excess	95% width
Fast	None	0.909	29437	0.945 (0.016)	0.020 (0.010)	0.02617	0.00179	0.02795
Fast	Top 20	0.909	29437	0.960 (0.014)	1.000 (0.000)	0.01279	0.00173	0.01452
Slow	None	0.022	1381	0.950 (0.015)	0.010 (0.007)	0.02632	0.00220	0.02852
Slow	Top 20	0.022	1381	0.955 (0.015)	1.000 (0.000)	0.01286	0.00218	0.01505

54.5% and 80.7% of sharp width. The original seed/mix row is reproduced exactly, while all width identities and no-direct-reward-leakage checks pass in every run.

Interpreting improvements across designs. Population optimality does not imply a large gain over every heuristic. At $K = 20$, the local-action gain is below 1% in 20 of the 30 grid cases. The original full-support pipeline has 93.0% coverage in 200 replications (MCSE 1.8 percentage points), while the fresh 1,000-run scan gives 95.2% for the positive contrast and 95.6% under the null, with 4.4% wrong-sign certification. These are distinct prescribed experiments. Their Monte Carlo intervals quantify simulation uncertainty and do not establish exact finite-sample calibration. The population audit objective minimizes remaining identification width; maximizing the chance of a particular sign certificate also depends on the outcomes recovered by the audit.

Equal reward-budget comparison. We compare learned top-20 and all-80 support at confirmation budgets $B \in \{2{,}000, 8{,}000\}$ and pilot budgets $B/5$ under the original positive and zero contrasts. Each of the four conditions has 200 paired replications. The supports share the same independent 30,000-record design and inference samples. Supports share passive Gaussian draws; their pilot, confirmation, and audit multiplier streams are independent. Every selected cell receives at least 10 confirmations; the remaining budget uses the original minimum-plus-proportional allocation. The pilot total is also matched, so a smaller support receives more pilot observations per cell. The DGP, estimator, $N_{\text{pas}}^{-2/5}$ screen, and 999 Gaussian draws are unchanged. Seed 20260925083 and all settings were recorded before execution.

All-80 produces narrower confidence intervals in every condition because it removes the identification floor. Its positive-effect coverage is $96.0\%/95.5\%$ at the two budgets. Under the null, coverage is $94.0\%/92.5\%$ and wrong-sign certification is $6.0\%/7.5\%$. The Wilson intervals in Table 16

Table 16: Matched reward budgets with the unchanged 95% sign envelope. Both supports receive confirmation budget B and pilot budget $B/5$, with 10 minimum confirmations per selected cell. Each condition has 200 replications, 30,000 records per independent passive split, and 999 Gaussian influence draws. Coverage is whole-set coverage with a 95% Wilson Monte Carlo interval. ID and CI give mean identification and confidence widths. Activation costs and failed attempts are not modeled.

Law	B	Support	Cover [MC interval]	ID width	CI width	Certify > 0	False sign
Positive	2,000	Top 20	99.5 [97.2, 99.9]	0.0127	0.0155	100.0	0.0
Positive	2,000	All 80	96.0 [92.3, 98.0]	0.0000	0.0022	100.0	0.0
Positive	8,000	Top 20	100.0 [98.1, 100.0]	0.0128	0.0153	100.0	0.0
Positive	8,000	All 80	95.5 [91.7, 97.6]	0.0000	0.0015	100.0	0.0
Zero	2,000	Top 20	100.0 [98.1, 100.0]	0.0140	0.0168	0.0	0.0
Zero	2,000	All 80	94.0 [89.8, 96.5]	0.0000	0.0025	3.5	6.0
Zero	8,000	Top 20	99.0 [96.4, 99.7]	0.0140	0.0165	0.0	0.0
Zero	8,000	All 80	92.5 [88.0, 95.4]	0.0000	0.0015	2.0	7.5

retain Monte Carlo uncertainty; they do not establish exact finite-sample calibration. The top-20 intervals cover their full identified sets in 99.0–100% of runs. This comparison matches observed reward counts, not total deployment cost: activation fees and unsuccessful attempts are absent. It therefore supports full auditing when all cells are affordable, without changing the top-K result for a fixed activation limit. Every condition and replication is retained in `results/equal_budget/`, including the pre-run design record, source hashes, validation, and a deterministic replay.

E.3 KuaiRec construction and data provenance

KuaiRec 2.0 contains 4,676,570 interactions from 1,411 users with 3,327 videos in its nearly complete small matrix (Gao et al., 2022). We retain finite watch ratios and define $R = \min\{\max(\texttt{watch_ratio}, 0), 2\}/2 \in [0, 1]$. With seed 20260807, 282 users (20%) form a design split. Their video-level mean rewards are used to fix the target policy. The other 1,129 users and 3,741,883 interactions form a disjoint evaluation population and supply exact finite-population truth.

For evaluation user u, the row-level diagnostic policy obeys $\pi_\beta(a \mid u) \propto \exp\{\beta \widehat{m}_a\}$ over the videos available to that user, where $\widehat{m}_a$ is the design-user score. We study $\beta \in \{0, 1, 2, 3, 4\}$ and pre-specify $\beta = 3$ for the single-value experiment. Fifteen outcome-free audit cells cross account-tenure tertile with design-score quintile. Within cell $x = (U, Q)$, routine response is

$$\mathbb{P}(O = 1 \mid X = x, R = r) = \text{logit}^{-1}\{\alpha_x - \log(4)r\},$$

where α_x targets rate $0.36 + 0.08\{(2Q + U) \bmod 5\}$ using zero-based indices. The realized overall response rate is 51.97%. This mask has true $\Gamma = 4$ and is fixed once; it is a transparent semi-synthetic intervention, not a claim about Kuaishou's production mechanism. The complete evaluation matrix supplies exact finite-population values. Response rates and respondent laws are empirical under this one realized mask, so substituting them into the population formulas gives plug-in sensitivity intervals, not exact finite-population identified sets. The generating odds restriction does not ensure coverage for every realized empirical mask.

Within an activated cell, audits sample initially hidden evaluation interactions without replacement. We compare complete case, stratified MAR-IPW, generating-probability MNAR-IPW, and plug-in sensitivity intervals under random, occupancy-only, ambiguity-only, and directional support. Support curves replace selected reward-mean intervals by full-feedback means and average 200 random permutations; these are oracle mean replacements. Conditional on the fixed evaluation population, passive mask, and support, finite-audit coverage uses 500 independent audit samples at within-cell fractions $q \in \{0.0005, 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05\}$. Thus the reported random-support curves average 200 permutations, whereas finite-audit coverage averages 500 randomized audits; neither regenerates the MNAR mask or includes passive-estimation uncertainty. Full-support auditing targets the exact scalar full-feedback value.

For the constructed $\pi_4 - \pi_0$ shared-law comparison, β changes only the induced occupancy over the pre-outcome cells; both policies use the $\beta = 0$ conditional interaction kernel inside each cell. The cell reward law is therefore fixed while occupancy varies, exactly matching Theorem 4. The raw row-level softmax contrast changes the within-cell video mix and is retained only as a coarsening diagnostic. We repeat the construction with 15, 30, 60, and 120 cells; the diagnostic gap falls monotonically from

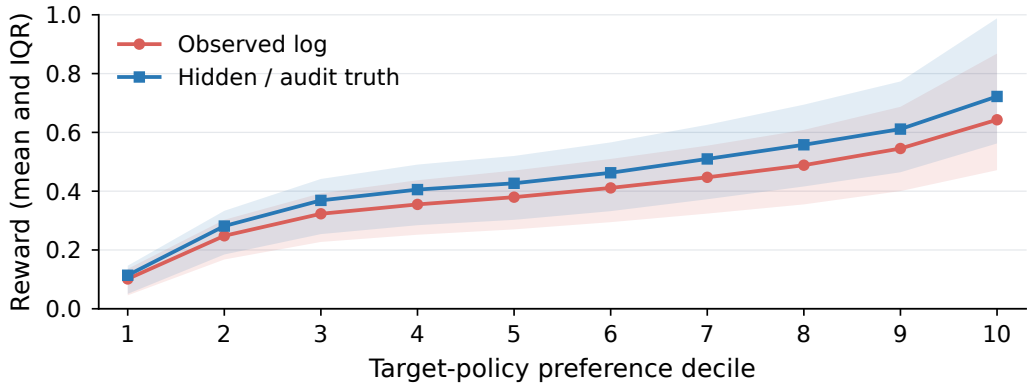


Figure 7: **The controlled outcome-dependent mask creates a visible passive-data gap.** Target-weighted real watch reward among routinely observed (red) and hidden-but-auditable (blue) KuaiRec interactions across target-preference deciles. Curves are weighted means; ribbons span weighted quartiles.

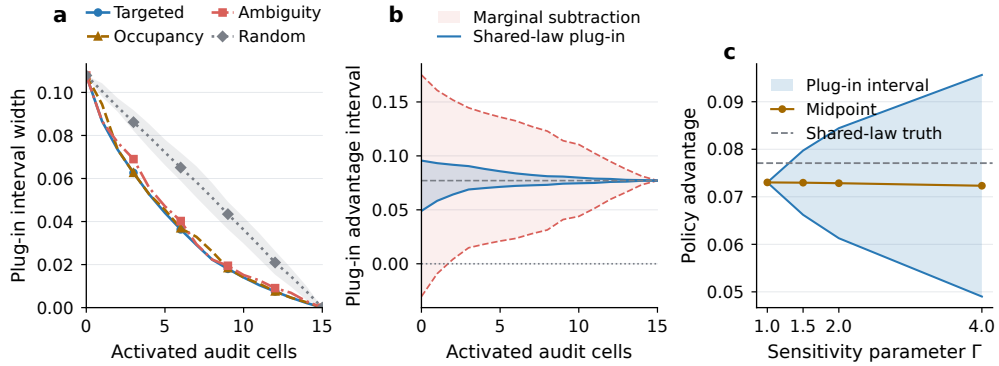


Figure 8: **Shared-law comparison of KuaiRec plug-in bounds.** Targeted support contracts the single-policy plug-in interval (left); marginal subtraction crosses zero while the $4.41\times$ narrower shared-law plug-in interval is positive (center); sensitivity intervals widen with Γ (right). The left-panel random band summarizes support orders; center and right shading show plug-in intervals. Positive plug-in endpoints are not sampling-valid certificates.

0.0127 to 0.0020 and the raw contrast lies in every reported plug-in interval; this containment is a diagnostic, not a sampling guarantee.

Table 17: KuaiRec plug-in results at $\Gamma = 4$ under one fixed mask. The complete matrix gives exact values; support rows replace the reward-mean intervals in 5/15 cells by full-feedback means. Random width averages 200 permutations.

Method	$\lvert\widehat{V} - V\rvert$	Truth in set	Width
Complete case	0.0247	–	–
Stratified MAR-IPW	0.0308	–	–
Oracle MNAR-IPW	0.0001	–	–
Unaudited Γ set	–	yes	0.1079
Random support (5)	–	yes	0.0721
Directional audit support (5)	–	yes	**0.0440**

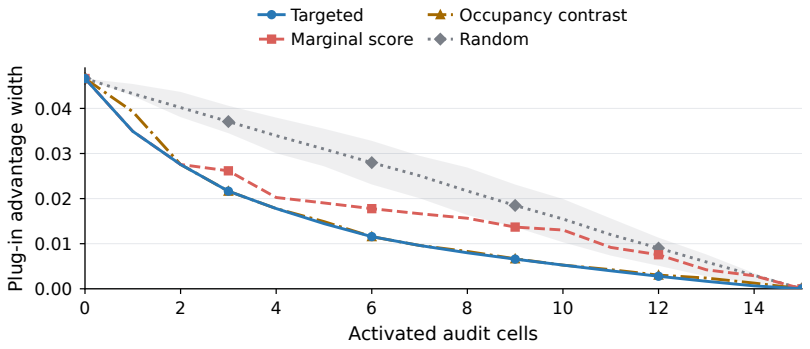


Figure 9: **Contrast-directed support in KuaiRec.** For the constructed shared-policy comparison, ranking cells by occupancy disagreement times ambiguity contracts the plug-in contrast interval; ribbons summarize random support orders.

Table 18: Contrast-level bounds and oracle mean-replacement baselines at the displayed audit budgets.

Data	estimand	method	k	width	Truth in set	$L > 0$
Bound baselines						
Exact MDP	contrast	Complete case	–	0.0000	0.000	1.000
Exact MDP	contrast	Manski shared law	–	0.0507	1.000	0.000
Exact MDP	contrast	Gamma=4 marginal subtraction	–	0.5193	1.000	0.000
Exact MDP	contrast	Gamma=4 shared-law sharp	–	0.0270	1.000	0.000
Exact MDP	contrast	Full-feedback oracle	–	0.0000	1.000	1.000
KuaiRec	contrast	Complete case	–	0.0000	0.000	1.000
KuaiRec	contrast	Manski shared-law plug-in	–	0.2212	1.000	0.000
KuaiRec	contrast	Gamma=4 marginal subtraction	–	0.2054	1.000	0.000
KuaiRec	contrast	Gamma=4 shared-law plug-in	–	0.0466	1.000	1.000
KuaiRec	contrast	Full-feedback oracle	–	0.0000	1.000	1.000
Oracle mean-replacement baselines						
Exact MDP	contrast	Comparison-specific	20	0.0127	–	–
Exact MDP	contrast	Local-action disagreement	20	0.0128	–	–
Exact MDP	contrast	Marginal targeted	20	0.0154	–	–
Exact MDP	contrast	Occupancy only	20	0.0137	–	–
Exact MDP	contrast	Ambiguity only	20	0.0177	–	–
Exact MDP	contrast	Random (mean)	20	0.0202	–	–
KuaiRec	contrast	Comparison-specific	5	0.0145	–	–
KuaiRec	contrast	Marginal targeted	5	0.0190	–	–
KuaiRec	contrast	Occupancy only	5	0.0149	–	–
KuaiRec	contrast	Ambiguity only	5	0.0299	–	–
KuaiRec	contrast	Random (mean)	5	0.0310	–	–

MDP bounds use exact population primitives. KuaiRec bounds are plug-in intervals from one realized mask and exclude passive-mask uncertainty; truths use full feedback. Selected cells are replaced by oracle means in the support rows. The full-feedback oracle is an infeasible scale reference.

E.4 Sensitivity and stress tests

The main panels report $\Gamma \in \{1, 1.5, 2, 4\}$; Figure 10 extends the KuaiRec grid through $8, 16, 32, \Gamma_{\text{cert}}$, and 64. Across every saved nondegenerate value $\Gamma > 1$, the same five cells appear in the same order (`U3|Q5`, `U1|Q5`, `U2|Q5`, `U1|Q1`, `U2|Q1`), giving unit Jaccard overlap and zero support regret relative to $\Gamma = 4$; at $\Gamma = 1$ all ambiguity widths vanish, and only ranks nine and ten swap from $\Gamma = 32$. Values below the generating 4 are deliberate negative controls: the generating response law is incompatible with the maintained class; empirical truth containment is reported separately. We also contrast incomplete and full audit support across seven sample fractions. All reported intervals, support permutations, and audit draws are written to CSV before plotting.

In the 80-cell controlled MDP, we additionally cross generating sensitivities $\Gamma_0 \in \{1, 2, 4, 8\}$ with analyst values $\Gamma \in \{1, 1.5, 2, 4, 8\}$. Every $\Gamma \geq \Gamma_0$ covers all 80 generating cell means, whereas every $\Gamma < \Gamma_0$ covers none in this construction. Aggregate containment is less diagnostic: in six below-diagonal settings the contrast interval still contains and positively certifies the true advantage

Table 19: KuaiRec partition refinement. The raw row-level contrast is 0.089770. Counts are finite-population min/median/max within cells, not independent sample sizes; partition-specific comparison confidence widths were not run.

Cells	Hierarchical truth	Gap	Plug-in width	Rows/cell	Distinct users/cell
15	0.077083	0.012688	0.046628	247,628/249,019/251,770	374/376/379
30	0.083394	0.006376	0.044705	123,683/124,522/126,069	374/376/379
60	0.086044	0.003727	0.044236	61,765/62,264.5/63,197	374/376/379
120	0.087813	0.001957	0.044127	30,838/31,136.5/31,811	374/376/379

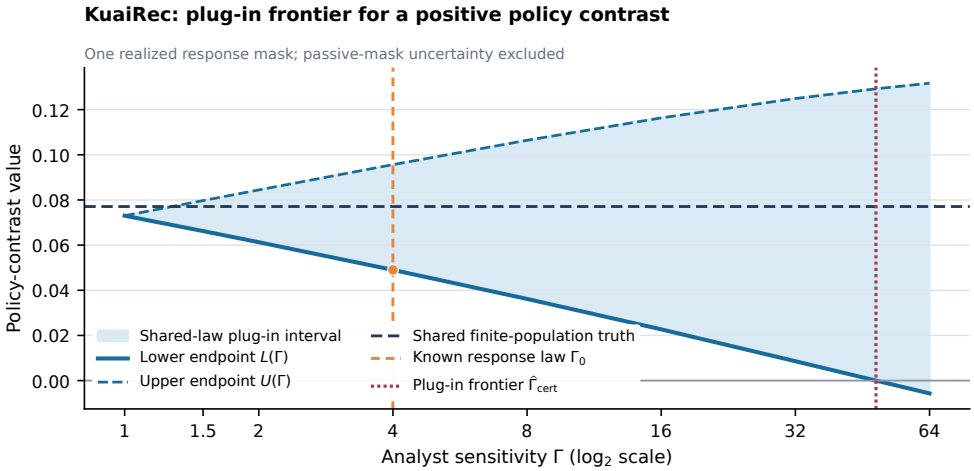


Figure 10: **KuaiRec's plug-in interval remains positive beyond the generating sensitivity.** The orange line marks the known generating $\Gamma_0 = 4$; the red line is the empirical positive-endpoint frontier $\widehat{\Gamma}_{\text{cert}} = 48.53$. Neither positive plug-in bounds nor aggregate containment validate the sensitivity model or provide a confidence guarantee.

although all local means are excluded. Calibration must therefore assess audited-law compatibility, not whether one aggregate happens to be contained.

For KuaiRec, the plug-in analogue of (80) gives $\widehat{\Gamma}_{\text{cert}} \in [48.528404745157, 48.528404745615]$. This is the positive-endpoint frontier of a fixed-mask plug-in interval, not a sampling-valid certificate or an estimate of response sensitivity. The known generating response law first satisfies the envelope at $\Gamma = 4$.

E.5 Approximate sharing within audit cells

The common-law comparison can be relaxed without silently changing its estimand. Let μ_{jx} be policy j's conditional reward mean in cell x, let μ_x be an anchor mean admissible under the shared-law model, and suppose the analyst supplies bounds $|\mu_{jx} - \mu_x| \le \epsilon_{jx}$. For a contrast a, write $T_a^{\text{sh}} = \sum_x \{\sum_j a_j d^{\pi_j}(x)\}\mu_x$. Then

$$\begin{aligned} T_a - T_a^{\text{sh}} &= \sum_{j,x} a_j d^{\pi_j}(x)(\mu_{jx} - \mu_x), \\ |T_a - T_a^{\text{sh}}| &\le \sum_{j,x} |a_j| d^{\pi_j}(x) |\mu_{jx} - \mu_x| \\ &\le \sum_{j,x} |a_j| d^{\pi_j}(x) \epsilon_{jx} =: E_a. \end{aligned} \tag{73}$$

Thus

$$T_a^{\text{sh}} \in [L_a, U_a] \quad \Longrightarrow \quad T_a \in [L_a - E_a, U_a + E_a].$$

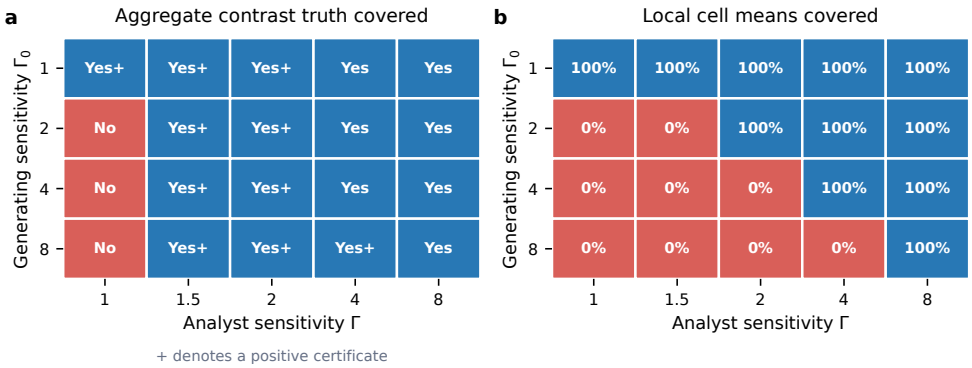


Figure 11: **Aggregate coverage can conceal local sensitivity misspecification in the 80-cell controlled MDP.** Rows generate at Γ_0 and columns analyze at Γ. Left: aggregate contrast containment; a plus denotes positive certification. Right: fraction of generating cell means contained.

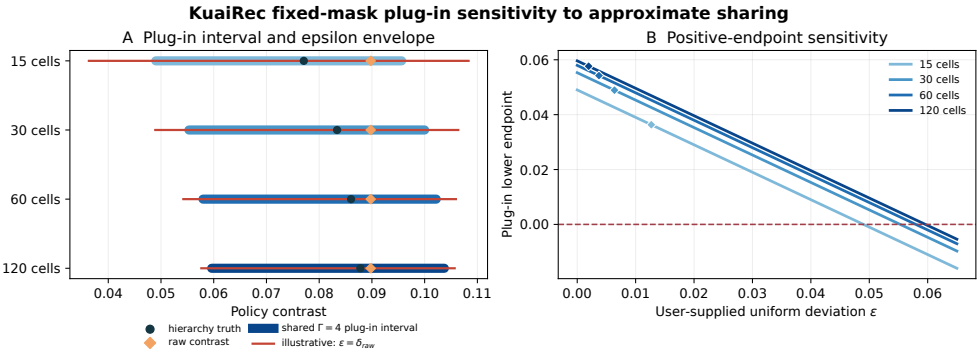


Figure 12: **Approximate sharing separates a guarantee from an ex-post diagnostic.** Thick blue segments are the fixed-mask shared $\Gamma = 4$ plug-in intervals. Thin red segments illustrate the outer envelope when the supplied ϵ is set equal to the realized aggregate gap; diamonds do not assert that this choice upper-bounds every cellwise deviation.

The result is a valid guarantee without identical within-cell laws; sharpness of the enlarged set would additionally require the bounded deviations to be jointly attainable.

For KuaiRec, the reference kernel is the anchor, so only the raw target mixture deviates. A uniform bound $|\mu_{1x} - \mu_{0x}| \leq \epsilon$ gives $E_a \leq \epsilon$ because normalized target occupancy sums to one. The observed raw-versus-hierarchy gap is a realized finite-population diagnostic and only a *lower bound* on any uniform cellwise ϵ capable of explaining it; it is neither an estimator nor an upper confidence bound for ϵ. Base endpoints are empirical plug-in bounds; applying (73) to them does not itself yield a confidence set. Refinement reduces that gap from 0.01269 in 15 cells to 0.00196 in 120 cells. Figure 12 illustrates setting the envelope radius to the realized gap for each partition; this is not a validated cellwise bound, although the lower endpoint remains positive under the illustration.

F When Reward and Next State Are Jointly Missing

The additive zonotope relies on observing S' when R is absent. Here we record the exact boundary of that result under either of two explicit observation schemes. In a conditionally sampled-tuple scheme, the design distribution of $X = (S, A)$ is ancillary, (X, O) are observed, and the packet $Y = (R, S')$ is observed only if $O = 1$. Alternatively, in an absorbing-dropout trajectory scheme, the known behavior policy generates actions and tuples are observed through the first $O = 0$ event, with Markov packet response and no later re-entry. The latter is related to parametric MNAR OPE for monotone trajectories (Wang et al., 2025). Call a cell target-relevant if it has positive normalized

occupancy under the fixed target policy for at least one admissible packet kernel, and take the union over those kernels. Assume every cell in this union has positive design probability in the tuple scheme or positive pre-dropout risk-set probability in the trajectory scheme. In either scheme, let $\rho_x = \mathbb{P}(O = 1 \mid X = x)$ and $F_{1x} = \mathcal{L}(Y \mid O = 1, X = x)$. Under full response positivity and the pairwise selection-odds bound over packet values, the sharp full packet-kernel set is

$$\mathcal{K}_x = \{\rho_x F_{1x} + (1 - \rho_x) w_x F_{1x} : w_x \in \mathcal{W}_\Gamma(F_{1x})\}. \tag{74}$$

For each w_x, set $q_x(y) = \rho_x / \{\rho_x + (1 - \rho_x) w_x(y)\}$. Then

$$\mathbb{P}_w(O = 1, Y \in dy \mid X = x) = \rho_x F_{1x}(dy), \qquad \mathbb{P}_w(O = 0 \mid X = x) = 1 - \rho_x.$$

In the absorbing-dropout scheme, a prefix whose first dropout is at time k has probability measure

$$\nu(ds_0) \prod_{t<k} \mu(da_t \mid s_t) \rho_{x_t} F_{1x_t}(dy_t)\, \mu(da_k \mid s_k)(1 - \rho_{x_k}),$$

which contains no w_x. Thus induction over prefix length gives the same observed trajectory law for every completion. Positive, outcome-independent ascertainment of missing packets makes $\mathcal{K}_x$ a singleton.

For bounded $f(Y)$, define $\mathfrak{m}_x^-(f)$ and $\mathfrak{m}_x^+(f)$ as $\rho_x \mathbb{E}_{F_{1x}} f$ plus $(1 - \rho_x)$ times the τ_- or τ_+ expectile of $f(Y)$ under F_{1x}; on audited cells both equal the full audited-kernel expectation. For a fixed stationary policy, let

$$(\mathcal{T}_{\mathcal{C}}^{\pm} v)(s) = \sum_a \pi(a \mid s) \mathfrak{m}_{(s,a),\mathcal{C}}^{\pm} \{R + \gamma v(S')\}. \tag{75}$$

Theorem 11. *Suppose the target-cell coverage and response-positivity conditions above hold, state, action, and respondent packet supports are finite, rewards are bounded, and packet kernels vary rectangularly across (s, a). The operators in (75) are monotone γ-contractions with unique fixed points $v_{\mathcal{C}}^-$ and $v_{\mathcal{C}}^+$. The sharp normalized value set is*

$$\left[(1 - \gamma)\nu^\top v_{\mathcal{C}}^-, (1 - \gamma)\nu^\top v_{\mathcal{C}}^+\right]. \tag{76}$$

Activating compatible audit cells weakly raises the lower endpoint and lowers the upper endpoint.

Proof. For every $K = (K_x)_x \in \prod_x \mathcal{K}_x$, define

$$(\mathcal{T}_K v)(s) = \sum_a \pi(a \mid s) \mathbb{E}_{K_{sa}} \{R + \gamma v(S')\}.$$

Proposition 3, applied to the scalar continuation reward, and rectangularity give

$$\mathcal{T}_{\mathcal{C}}^- v = \operatorname*{rect\,inf}_K \mathcal{T}_K v, \qquad \mathcal{T}_{\mathcal{C}}^+ v = \operatorname*{rect\,sup}_K \mathcal{T}_K v. \tag{T11.1}$$

The standard rectangular robust-policy-evaluation theorem, applied to these compact state–action-rectangular joint reward–transition kernels (Iyengar, 2005; Nilim & El Ghaoui, 2005), gives monotone γ-contractions, unique fixed points, and attaining kernels K^-, K^+ such that

$$v_{\mathcal{C}}^- \le v_K \le v_{\mathcal{C}}^+, \qquad v_{K^-} = v_{\mathcal{C}}^-, \quad v_{K^+} = v_{\mathcal{C}}^+. \tag{T11.2}$$

It remains to show that the identified scalar set contains every value between these attained endpoints. The map

$$J(K) = (1 - \gamma)\nu^\top (I - \gamma P_K^\pi)^{-1} r_K^\pi$$

is continuous in K. Since the finite-support product kernel set is compact and, by the convexity established in the proof of Proposition 3, each $\mathcal{K}_x$ is convex, the product is connected. Therefore

$$J\left(\prod_x \mathcal{K}_x\right) = [J(K^-), J(K^+)],$$

which proves sharpness of (76). If $\mathcal{C} \subseteq \mathcal{C}'$ and the additional audit laws are compatible,

$$\mathcal{T}_{\mathcal{C}}^- \le \mathcal{T}_{\mathcal{C}'}^- \le \mathcal{T}_{\mathcal{C}'}^+ \le \mathcal{T}_{\mathcal{C}}^+,$$

and monotone fixed-point iteration proves endpoint nesting. □

This recursion is a missingness-derived specialization of rectangular robust MDP evaluation and expectile Bellman operators (Iyengar, 2005; Nilim & El Ghaoui, 2005; Clavier et al., 2024; Bennett et al., 2024), not a new generic robust Bellman principle. Transition completion changes future occupancy, so audit gains can depend on both revealed kernels and other activated cells. A two-stage time-augmented example makes the boundary exact. Ignoring a common discount factor, let the value be pq, where the first packet controls reach probability $p \in [\ell, u]$ and the second controls terminal success $q \in [\ell, u]$. Then

$$\begin{aligned} W(\varnothing) &= u^2 - \ell^2, \\ W(\{p\}; p_0) &= p_0(u - \ell), \\ G_p(\varnothing; p_0) &= u^2 - \ell^2 - p_0(u - \ell). \end{aligned} \tag{77}$$

The gain from auditing p therefore depends on the revealed p_0 and cannot be a fixed additive cell score. Moreover, if the audited truths are $p_0 = q_0 = u$, the marginal gain from auditing p first is $\ell(u - \ell)$, while its gain after q is known is $u(u - \ell)$. This violates diminishing returns, so the audit-gain set function need not be submodular. Thus Theorem 11 deliberately does not inherit the top-K or knapsack claims of the reward-only model.

G Assumption Scope and Limitations

Sequential scope and observed transitions. Markov policies are fixed independently of missing rewards and response. Their values are the occupancy image of the rectangular reward-completion set; the audit geometry relies on always-observed next states. Feedback-adaptive policies or response dependence beyond the logged state require history augmentation. Theorem 11 covers rectangular joint dropout with nonadditive audit values; linked reappearance or later re-entry needs an explicit observation-process model.

Rectangular sensitivity and positivity. Cross-cell completion restrictions add modeling assumptions and may tighten bounds. Dependence of response on observed S' is allowed if the marginalized $q_x(r) = \mathbb{P}(O = 1 \mid R = r, X = x)$ satisfies Assumption 1; conditioning the envelope on S' gives a finer model. Full-law response positivity excludes reward values unique to nonrespondents and cannot be checked from routine respondents. Singular missing mass requires support bounds or an external audit, changing Proposition 3.

Audit validity. Identification requires positive, outcome-independent ascertainment given the cell and an unchanged outcome/time target. Selective success is covered by Proposition 5's second odds envelope. Drift or measurement changes require a measurement model; invitation followed by refusal requires a separate invitation indicator and a conditional audit-response model.

Finite-audit feasibility screen. Suppose each audit attempt in cell x has conditional success probability at least $\underline{\eta}_x$ given the full attempt history, and let c_{att} be its cost. Then $c_{\mathrm{att}}/\underline{\eta}_x$ is a conservative upper bound on expected cost per successful audit. Before treating a cell as operationally auditable, we require (i) at least $n_{1,\min}$ passive respondents, (ii) a missing pool allowing at least

$$A_x^\star = \min\{A : \Pr[\mathrm{Bin}(A, \underline{\eta}_x) \geq m_{\min}] \geq 1 - \alpha/|\mathcal{X}|\},$$

and (iii) $A_x^\star c_{\mathrm{att}}$ available audit budget. Cells that fail this screen retain their sensitivity ambiguity; they are not declared identified. We use $n_{1,\min} = m_{\min} = 30$ only as a transparent planning rule, not as a confidence-interval guarantee. For a target worst-case half-width h for a bounded $[0, 1]$ mean, the familiar conservative planning alternative is $m_{\min} \geq z^2_{1-\alpha/(2|\mathcal{X}|)}/(4h^2)$.

Joint budget and confirmation floors. Individual feasibility is insufficient for joint affordability. Let z_x indicate activation, n_x count attempts, m_x be the planned successful count, N_x the missing pool, and $\sum_x \alpha_x \leq \alpha_A$. After screening passive counts, enforce

$$\begin{aligned} &\sum_x (c_x^{\mathrm{on}} z_x + c_x^{\mathrm{attempt}} n_x) \leq B, \quad z_x \in \{0, 1\}, \quad n_x, m_x \in \mathbb{Z}_{\geq 0}, \\ &n_{x,\min} z_x \leq m_x \leq n_x \leq N_x z_x, \quad \Pr\{\mathrm{Bin}(n_x, \underline{\eta}_x) \geq m_x\} \geq 1 - \alpha_x \quad (z_x = 1). \end{aligned} \tag{78}$$

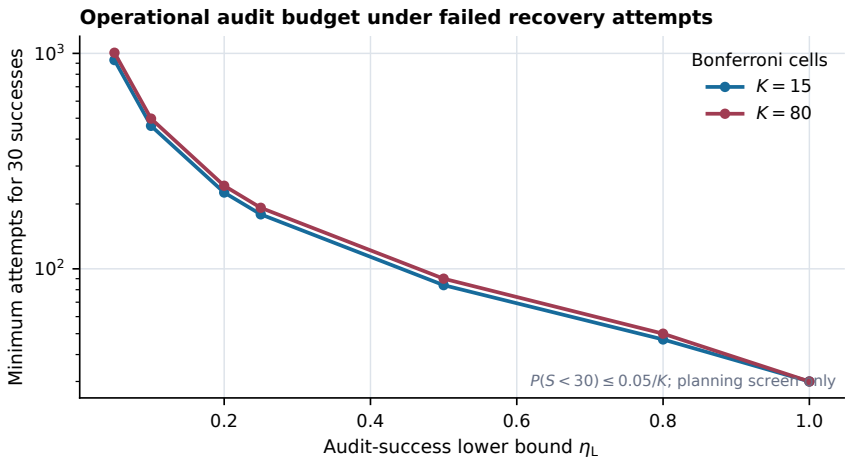


Figure 13: **Failed recovery attempts can dominate audit cost.** Minimum attempts required to obtain 30 successful audits with familywise failure probability at most 0.05 under a conservative success bound η_L. This is an operational screen, not a confidence-interval guarantee.

Thus the activation decision reserves its confirmation floor before any remaining allocation. With passive width W_0, a conditional planning surrogate minimizes $W_0 - \sum_x z_x \underline{g}_x^A + 2z_{1-\alpha_I/2}\{\sum_x z_x b_x^2/m_x\}^{1/2}$, where $b_x = |c_x|(1-\rho_x)\bar{\sigma}_{0x}$ and inactive terms are zero. For bounded rewards, $\bar{\sigma}_{0x} = 1/2$ is a pilot-free bound. This surrogate omits passive sampling uncertainty and is not an exact finite-sample confidence-width optimum. Under selective success, its variance term is only a proxy unless replaced by verified residual-endpoint scores. Pilot lower bounds and attempt-risk budgets must also account for their own failure probabilities; the final interval still uses the realized support in (46).

Sensitivity calibration. Γ is not learned from passive data. A valid randomized pilot can reject values that are too small on its audited support. If $w_x = dF_{0x}/dF_{1x}$, compatibility requires $\Gamma \ge \Gamma_{\min,x} := \operatorname{ess\,sup} w_x / \operatorname{ess\,inf} w_x$. For binary rewards, write $\theta_{ox} = \mathbb{P}(R = 1 \mid O = o, X = x)$. The displayed ratio formula assumes $\theta_{0x}, \theta_{1x} \in (0,1)$:

$$\Gamma_{\min,x} = \max\left\{ \frac{\theta_{0x}(1-\theta_{1x})}{\theta_{1x}(1-\theta_{0x})}, \frac{\theta_{1x}(1-\theta_{0x})}{\theta_{0x}(1-\theta_{1x})} \right\}. \tag{79}$$

A common degenerate Bernoulli law has $\Gamma_{\min,x} = 1$; if the two Bernoulli laws have different supports, response positivity rules out every finite Γ, so $\Gamma_{\min,x} = \infty$. A simultaneous region for $(\theta_{0x}, \theta_{1x})$ yields a one-sided lower bound by minimizing (79). A pilot can falsify a small Γ, but cannot validate a finite upper envelope or transport it without homogeneity. Continuous rewards also need binning or density-ratio regularity, so we report value and ranking over a pre-specified grid.

For a desired ordering, the decision-facing robustness summary is the *largest*, not the smallest, sensitivity value that preserves the certificate,

$$\Gamma_{\text{cert}} = \sup\{\Gamma \ge 1 : L_a(\Gamma) > 0\}. \tag{80}$$

We use this definition only when $L_a(1) > 0$; otherwise there is no baseline certificate and we report Γ_{cert} as *none*. Nested ambiguity sets make $L_a(\Gamma)$ nonincreasing. Thus Γ_{cert} measures tolerated sensitivity; it does not establish the law's Γ.

With observed audits, the operational frontier instead takes the supremum over the compatible domain $\{\Gamma : \mathfrak{P}_\Gamma(\mathcal{C}) \neq \varnothing\}$, reporting *none* if it never meets the positive-endpoint range. The reported fixed-mask KuaiRec frontier remains a descriptive plug-in calculation. A sampling-valid robustness statement instead uses outer-band lower endpoints on a predeclared sensitivity grid with simultaneous coverage; searching pointwise bands does not supply it.