WHEN IIS HISTORICAL FEEDBACK ENOUGH TO CHANGE A RECOMMENDATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Historical feedback can help a recommender decide whether to change its recommendations. However, more likes need not mean more sustained viewing, and the relation between these responses can change over time. We study when historical observations provide enough evidence for an improvement despite this uncertainty. Our method estimates changes in the feedback relation jointly and uses the same calibration set for every proposed recommendation. It checks both whether the estimated gain is large enough and whether the current feedback still points in the right direction. The decision accounts for repeated observations from a user and for choosing how much history to use. Under the stated assumptions, the rule bounds the chance of accepting a change that falls short of the required gain. A KuaiRand-Pure study shows that the benefit is regime dependent: shared calibration is useful in a calibration-limited matched setting, while fixed borrowing, separate-stage fitting or additional current labels can be preferable elsewhere. These findings connect the value of historical feedback to the information still needed before a recommendation changes.

1 INTRODUCTION

Recommendation systems often learn from feedback that is related to, but different from, the outcome they aim to improve. A short video can receive more likes without leading to more sustained viewing. The relation between these responses may change with the interface, audience or feedback process. Historical observations can still be useful, but their value depends on how well that change is understood. We study a practical decision: when does the available feedback justify replacing the current set of recommended items with another set?

Existing work addresses several parts of this problem. Exposure correction accounts for which items were observed, using inverse propensity weighting or doubly robust estimation (Schnabel et al., 2016; Wang et al., 2019). Label-shift methods use relations between observed responses and outcomes to adjust predictions (Lipton et al., 2018). Prediction-powered inference combines reliable outcome labels with auxiliary predictions (Angelopoulos et al., 2023; 2024). Our focus is the decision that remains after the observation support and weights have been fixed: whether historical feedback adds enough information to support an improvement under a changing feedback–outcome relation.

Two questions motivate this study. First, how much history should the decision use? More borrowing can reduce sampling uncertainty, but it also increases sensitivity to differences between historical and current feedback. Second, what must be established before a recommendation changes? A favorable feedback comparison supports an outcome improvement only when the current feedback relation has the required direction. Both checks use estimated quantities, and choosing a recommendation or a borrowing level from the data adds uncertainty. Calibrating one part accurately is therefore useful only if the complete decision has enough evidence to pass.

We estimate the changes in the feedback relation jointly and use the resulting confidence set for every proposed recommendation and borrowing level. This shared calibration preserves information about how the changes vary together. A recommendation is accepted only when the improvement score and the current feedback slope both remain positive after their uncertainty is accounted for. The rule also keeps repeated observations from one user together. In KuaiRand-Pure, the primary

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.

gain appears in a calibration-limited matched regime; transition, current-rich and purchase settings delineate when separate-stage fitting, fixed borrowing or additional current labels are preferable.

Our contributions are as follows:
• We formulate a recommendation change as two linked checks and give an error bound that covers selection over both recommendations and borrowing levels.
• We construct a shared calibration set for those checks, preserving precise combinations of stage differences while accounting for reused observations and dependence within users.
• We distinguish the value of borrowing, calibration precision and observation cost through complete decision comparisons, component analyses and controlled stress tests.

2 Related work

Recommendation from biased feedback. Inverse propensity and doubly robust methods address biased exposure in recommendation (Schnabel et al., 2016; Wang et al., 2019; Swaminathan & Joachims, 2015). Label-shift correction uses conditional feedback laws to adjust for changes between populations (Lipton et al., 2018). We study uncertainty in the feedback–outcome relation on fixed support. Historical and current relations may differ, and the relevant question is whether the remaining uncertainty permits a particular recommendation change.

Inference with auxiliary and proxy outcomes. PPI and PPI++ combine reliable labels and predictions to estimate an outcome while correcting prediction error (Angelopoulos et al., 2023; 2024). Classical surrogate-endpoint and surrogate-index work asks when short-run proxies can stand in for delayed outcomes (Prentice, 1989; Athey et al., 2019). We do not validate likes as a surrogate endpoint; instead, the feedback–outcome relation is allowed to change and its uncertainty is carried into a specific recommendation decision. Bias-aware inference and robust policy improvement provide related ways to trade additional information against misspecification (Armstrong & Kolesár, 2021; Kallus & Zhou, 2018; Petrik et al., 2016).

Safe decisions and historical borrowing. High-confidence policy improvement and Seldonian methods control undesirable decisions under statistical uncertainty (Thomas et al., 2015; 2019). Dynamic historical borrowing likewise adapts how strongly earlier data enter a current analysis (Viele et al., 2014; Rosenman et al., 2023). Our borrowing parameter is related to this idea, but here borrowed information enters a selected recommendation certificate through a separately calibrated change in the feedback–outcome relation.

Simultaneous and decision-based inference. Simultaneous contrasts protect several comparisons, risk-controlling selection protects a chosen action, and ratio inference handles uncertainty in a denominator (Scheffé, 1953; Angelopoulos et al., 2022; von Luxburg & Franz, 2009). Bounded-mean betting and concentration inequalities provide our calibration and sampling tools (Waudby-Smith & Ramdas, 2024; Maurer & Pontil, 2009). We use one inference rule for the improvement score and feedback slope, with explicit accounting for calibration, repeated users and selection.

3 Method

The method answers two questions before changing a recommendation: is the proposed improvement large enough, and does the feedback still point in the right direction? Figure 1 shows how shared calibration connects these checks. We first define the decision and available observations, then describe estimation, calibration and selection.

3.1 Recommendation and observation roles

Let $Y \in \{0, 1\}$ indicate sustained viewing and let $W \in [0, 1]$ be auxiliary feedback, such as a like. Write $p_i = \mathbb{E}(Y \mid i)$ for the outcome mean of item i in the target group. A *slate* is a set of s recommended items. Its additive value is $V(S; p) = s^{-1} \sum_{i \in S} p_i$. Relative to the reference slate S_0, define

$$d(S) = s^{-1}(\mathbf{1}_S - \mathbf{1}_{S_0}), \qquad \Delta(S) = d(S)^\top p, \qquad \mathbf{1}^\top d(S) = 0. \tag{1}$$

The vector $d(S)$ records which items are added or removed, and $\Delta(S)$ is their average outcome gain. The rule requires $\Delta(S) > \tau$ before replacing S_0. Keeping the reference contributes zero gain. This target is a mean over the specified group and item support.

Historical records F and current records D contain (Y, W, x). Calibration records C measure differences between stages and can also enter the labeled fit. Target records T reveal W but withhold Y from inference. Evaluation records E are reserved before fitting, and their outcomes are revealed after selection. Independent development data fix the representation $x \in \mathbb{R}^p$, reference, candidate family and tuning settings. Table 4 summarizes the notation.

For the current stage and target group, we assume

$$\begin{aligned} \mathbb{E}_D(W \mid Y, x) &= \theta_0 + a_D Y + x^\top \beta, \\ r_i = \mathbb{E}_T(W \mid i) &= \theta_0 + a_D p_i + x_i^\top \beta + b_i, \qquad |b_i| \le \eta_i. \end{aligned} \tag{2}$$

The slope a_D describes how feedback changes with the outcome after accounting for context. The term b_i allows a remaining difference between the current and target groups. The historical relation adds $\delta_0 + \delta_a Y$, so $\delta = (\delta_0, \delta_a)^\top$ captures changes in level and slope. Calibration of these historical–current changes is separate from the current–target allowance η_i.

Conditional on independent development and the specified design, user vectors are independent, the conditional mean relations hold, and weights are fixed. Responses within a user may be dependent. Each required parameter function must be estimable on the observed support; Appendix A.1 gives the corresponding checks. A frozen nonlinear representation is permitted when the mean is affine in that representation. Appendix A.2 also covers a larger retained subspace of context changes.

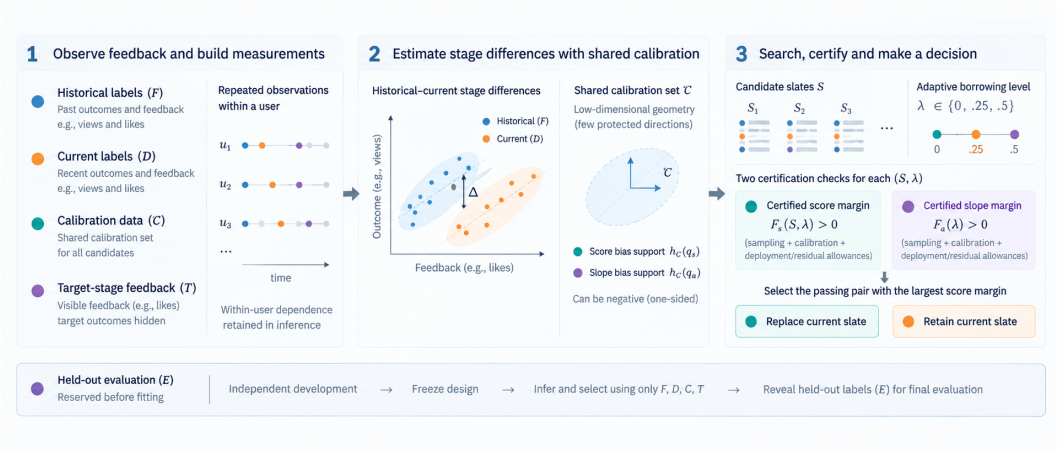


Figure 1: From observed feedback to a recommendation change. The same calibration set bounds the bias in the improvement score and current feedback slope. A candidate slate and borrowing level must pass both checks. Evaluation records are reserved before fitting, and their labels are revealed after selection.

3.2 Estimation with historical feedback

We vary how much the estimator borrows from history while keeping its observations fixed. Let W_L contain the labeled feedback, with mean $X\vartheta + H\delta$, where $\vartheta = (\theta_0, a_D, \beta^\top)^\top$. The columns of X represent the current relation; H marks historical changes. Stage aware fits separate stage levels and slopes. Its linear map B_A satisfies $B_A X = I$ and $B_A H = 0$ on the required functions. Pooled regression shares stage parameters and has map B_P, with $B_P X = I$. For borrowing level λ,

$$B_\lambda = (1-\lambda)B_A + \lambda B_P, \qquad \widehat{\vartheta}_\lambda = B_\lambda W_L, \qquad C_\lambda = B_\lambda H = \lambda B_P H. \tag{3}$$

At $\lambda = 0$, historical labels still estimate the shared context effects. Positive λ also borrows information about the historical level and slope. Every estimator uses the same permitted labels, including the calibration responses.

The two checks require an improvement score and a current slope estimate. Let $x_d = \sum_i d_i x_i$, $f_s(d) = (0, \tau, x_d^\top)^\top$, and $f_a = (0, 1, 0, \ldots, 0)^\top$. With target feedback means $\widehat{r}$, define

$$\widehat{g}_{d,\lambda} = d^\top \widehat{r} - f_s(d)^\top \widehat{\vartheta}_\lambda, \qquad q_s = -C_\lambda^\top f_s(d),$$
$$\widehat{a}_\lambda = f_a^\top \widehat{\vartheta}_\lambda, \qquad q_a = C_\lambda^\top f_a. \tag{4}$$

Their expectations are $a_D(\Delta - \tau) + d^\top b + q_s^\top \delta$ and $a_D + q_a^\top \delta$. Thus q_s and q_a specify which stage changes affect each check. The common intercept cancels from the slate contrast. Uncertainty in the fitted parameters remains part of both checks, with coefficients of reused responses combined before sampling error is bounded.

The score subtracts the required gain after expressing it in feedback units. This avoids dividing by an uncertain slope: a positive score establishes the desired gain only when the slope is also positive. For example, a larger like rate can reflect an improvement in viewing or a change in how viewing produces likes. The two checks separate these explanations under the measurement relation. A score margin is therefore evidence for a recommendation change, not an outcome gain reported in percentage points.

3.3 Shared calibration for the decision

Both checks depend on the same stage differences. We estimate these differences together rather than calibrating each candidate separately. The two outcome-specific gaps are $\kappa = (\delta_0, \delta_0 + \delta_a)^\top$. For $q = (q_0, q_1)^\top$, write $\bar{q} = (q_0 - q_1, q_1)^\top$, so $q^\top \delta = \bar{q}^\top \kappa$. Simultaneous intervals along fixed directions h_j give

$$\mathcal{C} = \{v \in [-1, 1]^2 : \ell_j \le h_j^\top v \le u_j,\ j \le J\}, \qquad h_{\mathcal{C}}(q) = \sup_{v \in \mathcal{C}} \bar{q}^\top v. \tag{5}$$

The set $\mathcal{C}$ contains the stage differences allowed by calibration. Its support $h_{\mathcal{C}}(q)$ bounds the bias that could make a check too optimistic. This bound is one-sided and can be negative. An absolute bound would also protect the opposite direction. Once $\kappa \in \mathcal{C}$, every query is protected, including one chosen after seeing the data.

Only an overly optimistic bias can cause a false acceptance, so the query determines which side needs protection. If every calibrated value lowers a score, its upper bias bound can be negative. Subtracting that bound then raises the decision margin. The coverage argument is unchanged because it uses the same signed upper bound.

Proposition 1 (Finite query geometry)**.** *For two changing channel parameters, normalize the nonzero full-borrowing queries to unit ℓ_1 norm. Their symmetric convex hull has at most four opposite vertex pairs. One generator per pair covers the finite candidate family and all nonnegative borrowing levels. Each query has a representation $\bar{q} = \sum_j t_j h_j$ with $\sum_j |t_j| \le \|\bar{q}\|_1$.*

The result follows from the four edges of the two-dimensional ℓ_1 diamond. It lets a small set of protected directions serve many slate comparisons. Increasing the context representation changes the queries, but the four-pair bound concerns the two stage differences. Appendix A.2 gives the proof and the extension to more changing parameters.

Calibration precision depends on the observed variation along each direction. A matched unit contains $(W_{F0}, W_{F1}, W_{D0}, W_{D1})$ from one user and supplies $Z = (W_{F0} - W_{D0}, W_{F1} - W_{D1})$. Ordinary calibration uses independent users in these four strata. Under the calibration design, $h^\top Z$ is bounded and has mean $h^\top \kappa$. For example,

$$\mathrm{Var}(Z_1 - Z_0) = \mathrm{Var}(Z_1) + \mathrm{Var}(Z_0) - 2\,\mathrm{Cov}(Z_1, Z_0). \tag{6}$$

Positive covariance can make a difference precise even when its coordinates vary. Matching preserves this observed dependence; grouping unrelated records does not create it.

We use bounded-mean betting intervals (Waudby-Smith & Ramdas, 2024) to retain that precision. Normalize a projection to $X_u \in [0, 1]$ with mean m. For a fixed fraction $c \in [0, 1]$,

$$K_n^+(m) = \prod_{u=1}^{n} (1 - c + cX_u/m) \tag{7}$$

has expectation one at the true mean for $m > 0$. A large product is evidence against a proposed mean. Applying the same construction to $1 - X_u$ bounds the other side. A fixed mixture over fractions gives an interval without choosing the fraction from test results. Appendix A.3 states the full construction, including boundary means. The total calibration error budget is α_C.

3.4 Sampling uncertainty and selection

A precise calibration set is only one part of the decision. The score and slope also have sampling error. We combine every occurrence of the same response into one coefficient and then group responses by user. For user contributions Z_u with range lengths $R_u \le \sum_j |w_{uj}|$, a Hoeffding radius is

$$e_H(\epsilon) = \sqrt{\tfrac{1}{2}\log(2/\epsilon)\sum_u R_u^2}. \tag{8}$$

This keeps dependence within users in the bound. Different response errors need not cancel when their coefficients have opposite signs. Independent development selects either this construction or the empirical Bernstein alternative detailed in Appendix A.6 (Maurer & Pontil, 2009).

For M candidate contrasts and L borrowing levels, score errors receive $\alpha_s/(ML)$ and slope errors receive α_a/L. The slope estimate is common to the slates at each borrowing level. Subtracting sampling, calibration and residual allowances gives

$$\begin{aligned} F_s(d,\lambda) &= \widehat{g}_{d,\lambda} - e_s(d,\lambda) - h_{\mathcal{C}}(q_s) - \eta_d - R_s, \\ F_a(\lambda) &= \widehat{a}_\lambda - e_a(\lambda) - h_{\mathcal{C}}(q_a) - R_a, \qquad\qquad \eta_d = \sum_i |d_i|\eta_i. \end{aligned} \tag{9}$$

The residual terms are zero under the retained mean model. When calibration responses also enter regression, Appendix A.7 gives the remaining residual influence after accounting for that reuse. The rule selects the passing pair with the largest score margin, breaking ties in a fixed order. It retains S_0 when no pair passes.

Theorem 1 (Protection after selection)**.** *Under the protected design, estimability, conditional mean relations and valid residual bounds, suppose the calibration, score and slope events have failure probabilities at most $\alpha_C, \alpha_s, \alpha_a$. Requiring the same (d,λ) to satisfy $F_s > 0$ and $F_a > 0$ gives*

$$\mathbb{P}\{\widehat{S} \neq S_0,\ \Delta(\widehat{S}) \le \tau\} \le \alpha_C + \alpha_s + \alpha_a.$$

The events may be dependent. The bound includes selection over both slates and borrowing levels.

On their intersection, $F_s \le a_D(\Delta - \tau)$ and $F_a \le a_D$. Positivity of both therefore implies $\Delta > \tau$. The guarantee controls the chance of accepting an insufficient improvement over the full procedure, rather than an error rate conditional on acceptance. Appendix A.8 gives the proof, implementation steps and protection for independent development.

3.5 When borrowing is useful

The decision margins identify which additional information can help. Before subtracting calibration, let $U_s = \widehat{g} - e_s - \eta_d - R_s$ and $U_a = \widehat{a} - e_a - R_a$. A pair passes when $U_s > h_{\mathcal{C}}(q_s)$ and $U_a > h_{\mathcal{C}}(q_a)$. If a non-calibration margin is nonpositive and its bias bound is nonnegative, shrinking that bound to zero still leaves the pair blocked.

For the same candidate, comparison with the standalone Stage aware rule gives

$$\begin{aligned} F_s^{\mathrm{Shared}}(d,\lambda) - F_s^{\mathrm{Stage}}(d) = (\widehat{g}_{d,\lambda} - \widehat{g}_{d,0}) + (e_s^{\mathrm{Stage}}(d) - e_s(d,\lambda)) \\ - h_{\mathcal{C}}(q_s) - (R_s(d,\lambda) - R_s(d,0)). \end{aligned} \tag{10}$$

Borrowing changes the score center, saves or spends sampling precision, and introduces a calibration term. The corresponding slope comparison must also pass. This identity explains the margin for one candidate; final recommendation value additionally depends on which candidate is selected.

Even a positive current slope can leave a decision unresolved. For $0 < a_0 < a_1 < 1$ and $0 < c \le a_0$ with $c/a_0 > \tau \ge c/a_1 > 0$, Theorem 2 constructs two models with gains $c/a_0, c/a_1$ and identical

historical and target-feedback laws. Current labels distinguish the models; extra history and target feedback do not. A change-of-measure argument (Kaufmann et al., 2016) gives the required count in the balanced-label experiment for power $1 - \beta$ and error at most α:

$$N_D \geq \frac{\mathrm{kl}(1-\beta, \alpha)}{\mathrm{kl}((1+a_0)/2, (1+a_1)/2)}. \tag{11}$$

4 Experiments and results

4.1 Experimental setup

We evaluate complete decision rules on KuaiRand-Pure logged user–video interactions (Gao et al., 2022). The outcome Y is `long_view`, and feedback W is `is_like`. Our empirical target is fixed-support slate replacement under an additive mean-value objective: development fixes 16 videos, a three-video reference slate, 120 alternatives and the context representation. Each condition contains 200 paired analyses, evaluation outcomes are reserved before fitting, and the minimum gain is $\tau = .01$. The observation budgets defining the calibration, transition and current-rich settings are fixed before the development search; development selects the sampling construction, borrowing grid and error allocation within those settings.

The three main settings vary the available information. The calibration window combines many historical labels and target feedback records with few current labels. The transition window increases calibration and reduces target feedback. The current-rich window provides more current labels. All calibration labels also enter the labeled fit. Appendix B gives the observation budgets and protocol. Table 1 lists the comparisons.

Independent development selects the sampling bound, borrowing grid and error allocation. Shared and its calibration variants, together with Fixed and Five, use the selected allocation $(\alpha_C, \alpha_s, \alpha_a) = (.005, .025, .020)$; Stage aware uses $(.02, .03)$ for score and slope. Shared selects among $\{0, .25, .5\}$, allowing modest borrowing without paying for an unnecessarily aggressive level. Fixed uses only $\lambda = .5$. Zero retains Shared's allocation while permitting only zero borrowing; Five uses five levels. The primary comparisons are Shared minus Stage aware, Ellipsoid, Zero and Fixed in the calibration window. Their paired bootstrap intervals adjust for four comparisons. Other intervals are pointwise. Gain includes zero for reference retention; Appendix B.5 defines both empirical evaluation summaries.

Table 1: Comparison methods and their information use. Stage aware is a complete alternative; Zero isolates borrowing under Shared's error allocation. Shared and its calibration variants, together with Fixed and Five, use $(.005, .025, .020)$; Stage aware uses $(.02, .03)$ for score and slope. Calibration methods use identical responses and the same calibration error budget. Purchase methods receive the same additional resource budget.

Method	Information use
Shared	Shared calibration; selects a slate and borrowing level.
Stage aware	All labels, separate stage parameters, no borrowing search.
Ellipsoid	Covariance-shaped calibration region using pilot directions.
Box / Projected	Coordinate intervals / one interval for each distinct query.
Zero / Fixed / Five	Zero borrowing / one borrowing level / five borrowing levels.
Current	Uses the calibration budget for more current labels.
Direct / PPI	Target labels; PPI also uses a frozen outcome prediction.

4.2 When does borrowing improve the decision?

We first compare the value of the final recommendation, including decisions to retain the reference. In the calibration window, Shared reaches 16.22 percentage points, compared with 13.48 for Stage aware (Figure 2). The paired gain is 2.74 [1.40, 4.08] points. Shared accepts in 94.00% of replicates and Stage aware in 79.00%. There are 37 replicates where only Shared accepts and 7 where only Stage aware accepts. The improvement therefore concerns the complete decision rule.

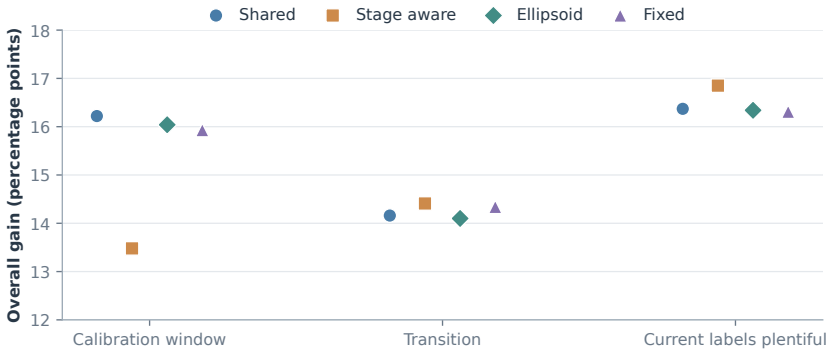


Figure 2: Overall gain under three information settings. Points average 200 paired replicates. Shared and Ellipsoid remain close, but the calibration-window decisions are no longer identical. The Shared–Stage aware contrast in Table 2 uses the adjusted interval. Gain includes retention of the reference.

Five reaches 15.58 points. Fixed remains competitive at 15.93 because it avoids the borrowing-level search. Shared nevertheless reaches 16.22, a paired +0.29 [0.05, 0.54], showing that adaptive borrowing can recover the selection cost when evidence varies across replicates. Shared selects positive borrowing in 52.0% of replicates. Stage aware remains slightly stronger than Shared in the transition setting (14.41 vs. 14.16). When current labels are plentiful, Stage aware remains stronger than all borrowing variants (16.85 vs. 16.37 for Shared), locating where adaptive borrowing helps and where more current information is preferable.

Table 2: Paired differences in overall gain in the calibration window, in percentage points. Intervals use four-comparison Bonferroni tail allocation. Counts list both accepting, Shared only, comparator only, and neither.

Comparison	Difference	Adjusted interval	Both	Shared	Other	Neither
Shared − Stage aware	+2.74	[1.40, 4.08]	151	37	7	5
Shared − Ellipsoid	+0.18	[-0.03, 0.39]	181	7	4	8
Shared − Zero	+6.85	[5.32, 8.39]	112	76	4	8
Shared − Fixed	+0.29	[0.05, 0.54]	182	6	4	8

The Zero comparison answers a narrower question than the Stage aware comparison. Zero retains Shared's selection accounting while excluding positive borrowing, whereas Stage aware spends no error budget on that search. Shared improves on Zero by 6.85 [5.32, 8.39] points. The smaller gain over Stage aware reflects a stronger complete procedure, not a change in the observations. Separating these controls identifies both the contribution of borrowing within Shared and the value of adopting the full method.

Ordinary calibration accepts in 181 of 200 replicates; 26 accepted decisions use positive borrowing. Thus positive borrowing occurs in 13.0% of all replicates (14.4% of accepted decisions), versus 52.0% with matched calibration. Its overall gain remains below Stage aware (15.92 vs. 16.44). Matching therefore changes how often historical information can be used safely rather than creating borrowing from nothing.

4.3 What does shared calibration contribute?

Holding the observations fixed, Shared reaches 16.22 points versus 10.18 for Box. Projected remains close in complete value at 16.06 but uses 121 protected directions. Ellipsoid remains close at 16.04, but no longer makes identical decisions: Shared uses two protected directions rather than twelve and has a paired difference of +0.18 [-0.03, 0.39]. This separates calibration economy from a claim that every tighter representation must improve final value. One-sided support gives 16.22 versus 16.16 for two-sided support, while empirical Bernstein or Hoeffding calibration gives 9.37.

Score precision remains the broadest first-stage filter, but it no longer dominates the full candidate analysis. After the sampling checks are cleared, calibration blocks 29.2% of the remaining candidates at $\lambda = .25$ and 35.8% at $\lambda = .5$. Thus sampling determines which candidates are competitive, while calibration materially determines which competitive candidates can be certified. Appendix C.2 reports both unconditional and conditional views.

Table 3 explains one candidate from the first replicate where Shared accepts and Box does not. Its point estimates and sampling radii are identical. Both score margins remain positive, while only the slope check changes from negative under Box to positive under Shared. The calibration set changes whether the feedback direction is established for that candidate. Aggregate results in Tables 2 and 5 include all replicates.

Table 3: A fixed candidate under Shared and Box calibration. Entries are multiplied by 100 and have feedback units. Margin equals Center minus Sampling, Calibration and Other. Shared changes the slope check from negative to positive while leaving the fitted estimates and sampling radii unchanged.

Method	Test	Center	Sampling	Calibration	Other	Margin
Shared	Score	2.395	1.820	0.031	0.200	0.344
Shared	Slope	13.685	7.537	4.249	0.000	1.899
Box	Score	2.395	1.820	0.072	0.200	0.303
Box	Slope	13.685	7.537	7.874	0.000	-1.726

4.4 Which observations are worth acquiring?

The purchase comparison asks which additional observations improve the decision under a fixed resource budget. Each method starts with the same history, current labels and target feedback. The resource model charges four labeled responses per matched unit, failed eligibility screens and half a label unit for matching. Current spends this budget on current labels. Direct and PPI buy representative target outcomes; PPI also corrects the errors of a predictor fitted on independent development data (Angelopoulos et al., 2023; 2024). Appendix D gives the costs and finite-sample inference procedures.

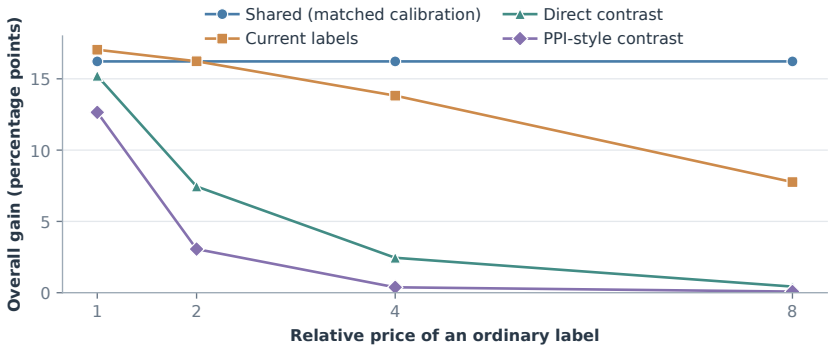


Figure 3: Overall gain at a fixed additional resource budget. Increasing the relative label price reduces the number of available labels. Screening and matching costs are included for Shared. Direct and PPI require target outcomes at decision time.

At relative price one, Current reaches 17.04 points and Direct reaches 15.21, compared with Shared's 16.22 (Figure 3). At price four, Current and Direct reach 13.82 and 2.45. Their decline reflects fewer labels, with the outcome definition and reference unchanged. PPI reaches 12.65 at price one. Its residual correction protects inference, but a weak proxy need not improve precision. The comparison uses direct inference for slate contrasts; Appendix D.2 also retains the more conservative coordinate variants on the same purchased records.

When target outcomes are unavailable at decision time, the eligible options are current labels and calibrated feedback. Earlier mature outcomes can inform the current relation, while transfer to a later target window still requires its own allowance. Shared is therefore not preferred when informative current labels are cheap; its value emerges when those labels are expensive or unavailable. The price axis describes an analysis budget rather than platform prices.

4.5 Further checks

The following checks are controlled diagnostics layered on the logged KuaiRand-Pure study; they are not additional naturally occurring data regimes. We examine whether calibration benefits persist when other information is limited. With repeated target users, Shared reaches 3.10 points; with limited fitting and target information, it reaches 0.42 points. In the controlled representation study, Shared's gains are 18.05, 9.75 and 3.05 points as the nonzero coefficient count increases from 8 to 32 and 64; Stage aware remains at 13.39, 12.19 and 4.00. Each representation uses 100 replicates; each larger-candidate design uses 40. The stage discrepancy remains two-dimensional. A small calibration family therefore coexists with a growing estimation burden. Larger candidate designs cover up to 5,000 alternatives; acceptance remains high but decreases slightly for the largest families. Appendix C reports their joint changes in support, slate size and family size.

Controlled zero-gain and exact-threshold checks retain the reference throughout, giving a 95% binomial upper endpoint of 1.83% for zero errors in each 200-replicate condition. When the actual deployment residual is five times the declared allowance, false acceptance occurs in 3.0% of Shared replicates and 7.0% of Stage aware replicates. These checks separate uncertainty under the stated model from the consequences of violating its allowance.

5 Discussion

Calibration precision and decision value. Shared improves over Box and remains close to Ellipsoid while using two rather than twelve protected directions. The fixed-candidate and bottleneck analyses show that tighter calibration matters only when it changes a passing status or selected slate; fewer directions reduce calibration complexity without implying higher value in every replicate.

Choosing how much history to use. Fixed avoids a selection cost and remains competitive at 15.93 points, while Shared reaches 16.22. Positive borrowing is selected in 52.0% of matched-calibration replicates and 13.0% with ordinary calibration. Fixed cannot retreat from $\lambda = .5$ when matching weakens or residual protection grows, whereas Shared can fall back to smaller borrowing levels.

Scope and information choice. The guarantee targets additive fixed-support slate value; position effects or cross-item interactions require a different value functional. The finite-family union bound pays a $\log(ML)$ cost, so the larger-family checks study inference scaling rather than production ranking. Matched calibration can favor users with sufficient cross-stage coverage, and the 0.8 screening success probability is a declared resource assumption. Operationally, weak sampling margins point to more current labels, while a remaining calibration obstacle favors shared calibration; these diagnostics are fixed before Test evaluation.

6 Conclusion

Historical feedback supports a recommendation change when its added information outweighs uncertainty in the changing feedback–outcome relation. Shared calibration carries this uncertainty through the score and slope while protecting the selected recommendation. Its benefit is regime dependent: matched calibration can make borrowing more certifiable, while separate-stage fitting, fixed borrowing or additional current labels can be preferable elsewhere. The present guarantee concerns additive fixed-support slate replacement rather than full production ranking.

AI USE STATEMENT

Generative AI tools were used for writing and language editing, figure preparation, and assistance with parts of the software implementation and experimental analysis, including code drafting, debugging, refactoring, and consistency checking. All AI-assisted text, figures, analyses, and code were reviewed and checked by the authors. The authors take responsibility for the final manuscript, implementation, reported results, and conclusions.

ETHICS STATEMENT

This study uses the publicly released KuaiRand-Pure dataset and does not collect new user data or conduct interventions on users. Results are reported in aggregate, and the study does not attempt to identify individual users. Practical deployment would require appropriate privacy and data-access safeguards.

REPRODUCIBILITY STATEMENT

The appendices document the theoretical results, experimental protocol, observation roles, statistical procedures, and additional analyses. The supplementary material provides anonymized code and stored outputs for reproducing the reported analyses and figures. The empirical study uses the publicly released KuaiRand-Pure dataset.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2022. URL https://arxiv.org/abs/2110.01052v5. Version 5.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *arXiv preprint arXiv:2301.09633*, 2023. URL https://arxiv.org/abs/2301.09633v4. Version 4.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2024. URL https://arxiv.org/abs/2311.01453v2. Version 2.

Timothy B. Armstrong and Michal Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021. doi: 10.3982/QE1609. URL https://arxiv.org/abs/1808.07387.

Susan Athey, Raj Chetty, Guido W. Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical Report 26463, National Bureau of Economic Research, 2019.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://academic.oup.com/biomet/article/26/4/404/291538.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 3953–3957, 2022. doi: 10.1145/3511808.3557624. URL https://kuairand.com/.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1805.08593.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016. URL https://arxiv.org/abs/1407.4443.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Marek Petrik, Yinlam Chow, and Mohammad Ghavamzadeh. Safe policy improvement by minimizing robust baseline regret. *arXiv preprint arXiv:1607.03842*, 2016. URL https://arxiv.org/abs/1607.03842.

Ross L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989. doi: 10.1002/sim.4780080407.

Evan T. R. Rosenman, Guillaume Basse, Art B. Owen, and Mike Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 2023. doi: 10.1111/biom.13827.

Henry Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1–2): 87–110, 1953. doi: 10.1093/biomet/40.1-2.87.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the International Conference on Machine Learning*, volume 48, pp. 1670–1679, 2016. URL https://proceedings.mlr.press/v48/schnabel16.html.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004, 2019. doi: 10.1126/science.aag3311.

Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G. Ibrahim, Nelson Kinnersley, Stacy Lindborg, Sandrine Micallef, Satrajit Roychoudhury, and Laura Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2014. doi: 10.1002/pst.1589.

Ulrike von Luxburg and Volker H. Franz. A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations, and bootstrap. *Statistica Sinica*, 19(3):1095–1117, 2009. URL https://arxiv.org/abs/0711.0198.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *Proceedings of the International Conference on Machine Learning*, volume 97, pp. 6638–6647, 2019. URL https://proceedings.mlr.press/v97/wang19n.html.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009. URL https://arxiv.org/abs/2010.09686.

A THEORY AND NUMERICAL INFERENCE

A.1 LINEAR FUNCTIONS, BIAS, AND ESTIMABILITY

Table 4: The main objects. Each observation has one identity even when it enters several computations.

Symbol	Meaning
F, D, C, T, E	Historical labels, current labels, calibration, target feedback, evaluation
d, Δ, τ	Slate contrast, target gain, minimum required gain
δ, κ	Stage parameter difference and its class-gap coordinates
$\lambda, \widehat{g}, \widehat{a}$	Borrowing level, improvement score, current slope estimate
$h_{\mathcal{C}}(q)$	Upper bound on bias in query direction q
e_s, e_a	Sampling radii for the score and current slope

The labeled observations may be individual responses or means within a fixed design cell. In the latter case, the coefficient of a cell mean is divided by its count before a response-level uncertainty calculation. Let $Z = [X, H]$ be the unrestricted stage design. Weighted least squares uses the observed cell counts, not one equal weight per cell. The first block of its coefficient map is B_A. The pooled design uses X alone and gives B_P.

For the score or slope function f, the required identities are

$$f^\top B_A X = f^\top, \qquad f^\top B_A H = 0, \qquad f^\top B_P X = f^\top.$$

They are checked on the relevant functions rather than requiring every nuisance coefficient to be identifiable. If a required identity fails, that candidate receives an unsupported status. If only an unrelated nuisance direction is unidentified, its presence does not force rejection. The numerical tolerance for these identities is 10^{-7}.

By linearity and the identities above,

$$\mathbb{E}(\widehat{\vartheta}_\lambda) = \vartheta + C_\lambda \delta.$$

Using $\mathbf{1}^\top d = 0$ in the target mean relation gives

$$\mathbb{E}(\widehat{g}_{d,\lambda}) = a_D(d^\top p - \tau) + d^\top b - f_s(d)^\top C_\lambda \delta,$$
$$\mathbb{E}(\widehat{a}_\lambda) = a_D + f_a^\top C_\lambda \delta.$$

This establishes the signs of q_s and q_a in Equation (4). The level term disappears from $d^\top r$, but can still affect the biased parameter fit through a nonorthogonal design. The score has feedback units. A positive score margin is not itself a gain measured in outcome percentage points.

A.2 PROOF OF THE FINITE QUERY RESULT

In this proof, q denotes a query in the κ coordinates, written as $\bar{q}$ in Section 3.3. Let $\mathcal{Q}$ contain the nonzero queries at $\lambda = 1$, including the slope query. Here $\lambda = 1$ is used only to normalize the query directions in the geometric argument; the empirical borrowing grid need not include $\lambda = 1$. Divide each by its ℓ_1 norm and form the symmetric set $\mathcal{V} = \{\pm q/\|q\|_1 : q \in \mathcal{Q}\}$. Every point lies on the boundary of the diamond $\{v : \|v\|_1 \le 1\}$. A finite collection on one edge has at most two extreme points on that edge. Across four edges the convex hull therefore has at most eight vertices. Central symmetry pairs them into at most four opposite pairs.

Choose representatives $h_1, \ldots, h_J$ from those pairs. Every normalized query has a convex combination of their signed versions. Multiplying by the query norm gives $q = \sum_j t_j h_j$, with $\sum_j |t_j| \le \|q\|_1$. A zero query has all coefficients zero. Since $C_\lambda = \lambda B_P H$, nonnegative borrowing scales the same query and creates no new calibration event. The directions are computed from the protected design, before calibration responses are opened. This proves Proposition 1.

Suppose intervals cover $h_j^\top \kappa$ simultaneously. Their intersection with the physical box contains κ; hence every upper support query bounds $q^\top \kappa$. The particular representation coefficients are not needed for the exact support calculation. They explain why the protected generator family is small.

A larger space of stage changes. A basis U fixed on independent development data permits $\beta_F - \beta_D = U\gamma$, with discrepancy $\delta^+ = (\delta_0, \delta_a, \gamma^\top)^\top$ of dimension $k = r+2$. The basis retains specified context changes; any omitted effect needs a separate residual allowance. In this higher-dimensional space, the normalized query hull can have many vertices. A fixed invertible matrix G_C defines the region $v = \widehat{\delta}^+ + G_C^{-1} z$, $|z_j| \le e_j$. Its support is

$$\sup_v q^\top v = q^\top \widehat{\delta}^+ + \sum_j |(G_C^{-\top} q)_j| e_j. \tag{12}$$

A physical-domain intersection can tighten this expression. Basis choice changes the widths, not the basic coverage argument. The four-pair result applies to the two stage parameters; the formula above applies to a larger retained shift space.

Frozen nonlinear representations. Let $x = \phi(x_{\mathrm{raw}})$, with ϕ selected on independent development data. Conditional on that development sample, ϕ is fixed. The same linear-function analysis applies if the conditional mean is affine in this representation. This permits nonlinear functions of raw context, while preserving a finite-dimensional coefficient model. It does not require the raw covariates themselves to enter linearly. Residual protection below handles a specified omitted component; a small empirical prediction loss alone is not used as a uniform residual bound.

A.3 The betting calibration interval

For independent calibration units Z_u with common mean κ, choose a nonzero generator h and write $b_h = \|h\|_1$. The normalized observation

$$X_u = \frac{h^\top Z_u + b_h}{2b_h} \in [0,1]$$

has mean $m = (h^\top \kappa + b_h)/(2b_h)$. The construction uses fixed fractions

$$\mathcal{F} = \{.001, .003, .01, .03, .10, .30, .60, .90\}.$$

For a candidate $m \in (0,1)$, define two nonnegative mixtures

$$K_n^+(m) = \frac{1}{|\mathcal{F}|} \sum_{c \in \mathcal{F}} \prod_{u=1}^{n} \left(1 - c + c\frac{X_u}{m}\right),$$

$$K_n^-(m) = \frac{1}{|\mathcal{F}|} \sum_{c \in \mathcal{F}} \prod_{u=1}^{n} \left(1 - c + c\frac{1 - X_u}{1 - m}\right).$$

At the true mean, every factor has conditional expectation one and is nonnegative. Each product is a mean-one martingale and each mixture remains mean one. Markov's inequality gives $\mathbb{P}\{K_n^\pm(m) \ge 2/\epsilon\} \le \epsilon/2$. Retaining means for which both mixtures are below their threshold therefore covers m with probability at least $1 - \epsilon$. This is a fixed-fraction implementation of betting-based bounded-mean inference (Waudby-Smith & Ramdas, 2024).

The positive mixture is nonincreasing in m and the negative mixture is nondecreasing. The retained set is thus an interval. Bisection returns outward endpoints; boundary means zero and one are handled by limits. A histogram of the observed support gives exactly the same products as expanding the individual observations. In the binary stage-gap experiment, Z has nine possible values in $\{-1, 0, 1\}^2$. The mixture is evaluated in log space to avoid overflow.

For J generators, use $\epsilon = \alpha_C / J$ and transform each interval back to feedback units. A union bound protects all slabs. No extra calibration allocation is charged for M candidate slates or L positive rescalings. The score and slope sampling events still protect their actual families separately. Calibration, labeled fitting, and score events can use the same responses; their joint protection uses a union bound rather than an independence claim.

Ordinary calibration. An ordinary unit in the computation is a block of four independently observed users, one from each stage and class stratum at the standardization context. Grouping them into an iid block for the calibration projection does not create covariance between them. The regression calculation treats them as four independent users. With matched measurements, the four responses belong to one user and are grouped as one vector in regression uncertainty. Both designs cost four labeled responses per calibration block, before screening and matching costs.

A.4 One-sided support and numerical certification

For $\mathcal{C} = \{v : Av \le b,\ |v_i| \le 1\}$ and any nonnegative dual vector z, the following computable upper bound holds:

$$\sup_{v \in \mathcal{C}} q^\top v \le z^\top b + \|q - A^\top z\|_1. \tag{13}$$

Indeed, $q^\top v = z^\top Av + (q - A^\top z)^\top v$, and the two terms are bounded by $z^\top b$ and the residual norm. This remains valid with a nonzero dual residual. In two dimensions we enumerate nonsingular pairs of constraint normals, obtain dual candidates, clip negligible negative coordinates, and include the resulting residual term. The smallest valid upper bound is used. A feasible primal maximizer alone would be a lower bound on this support function and is not used as a safety certificate.

The implementation intersects slabs with $[-1, 1]^2$ and checks vertices. Empty intersections return an inconsistent-set status. A missing finite support upper bound is unresolved. Neither status is interpreted as evidence that no true improvement exists. A guard of order 10^{-8} is added to the support bound. The final score and slope tolerance is also 10^{-8}. The checks compare the numerical upper support with independent linear-program solutions and verify physical-domain constraints.

An absolute support replaces $h_{\mathcal{C}}(q)$ by $\max\{h_{\mathcal{C}}(q), h_{\mathcal{C}}(-q)\}$. It is useful as a conservative comparison but is not needed in Theorem 1. The relevant optimism direction is already specified by q_s or q_a. In particular, a negative upper support is legitimate: it says every admissible stage difference biases the estimate downward in that direction.

A.5 A covariance-aware ellipsoid

An independent pilot estimates the covariance of the two class gaps. A ridge of $.003I$ makes its square root A invertible. Twelve equally spaced normal directions t_j on the half-circle define normalized directions proportional to $A^{-\top} t_j$. Each receives probability $\alpha_C/12$. Their betting intervals define a covered polytope, intersected with the physical box.

We compute an enclosing ellipse of that polytope using a minimum-volume iteration and then explicitly inflate it until every enumerated vertex satisfies the quadratic constraint. If the ellipse has center c and shape matrix S, its support is

$$q^\top c + \sqrt{q^\top S q}.$$

Because the ellipse contains the covered polytope, it contains the true discrepancy on the same calibration event. The construction avoids replacing all directional radii by their maximum. The resulting comparison is with this explicit covariance-shaped finite-sample region. Enclosing a polytope remains an approximation, and the mean value comparison is not a minimax comparison over all ellipsoidal procedures.

A.6 User-level sampling bounds

For a fixed linear functional, combine every occurrence of response W_j into coefficient w_j. For user u, let $Z_u = \sum_{j \in u} w_j (W_j - m_j)$, with a response-independent centering m_j. The sum has the same centered error as the uncentered statistic. Its range length is at most $R_u = \sum_{j \in u} |w_j|$. Applying the bounded-difference moment-generating-function inequality to independent users and optimizing the exponential parameter yields

$$\mathbb{P}\left\{ \left| \sum_u (Z_u - \mathbb{E} Z_u) \right| > t \right\} \le 2 \exp\left(-\frac{2t^2}{\sum_u R_u^2} \right).$$

Equation (8) follows. Different response distributions and means are permitted; independence is required between the user vectors.

For clarity, the empirical Bernstein comparison uses a *common* interval $[A, B]$ containing every centered Z_u, with $R = B - A$, n users, and sample variance s_Z^2. Applying Maurer–Pontil Theorem 11 to $(Z_u - A)/R$ and its complement gives the two-sided bound

$$e_{EB}(\epsilon) = \sqrt{2 n s_Z^2 \log(4/\epsilon)} + \frac{7nR \log(4/\epsilon)}{3(n-1)}. \tag{14}$$

The cited theorem already covers independent variables with different distributions (Maurer & Pontil, 2009). Its proof uses $n^{-1}\sum_u \mathrm{Var}(X_u) \le \mathbb{E}s_X^2$, concentration of sample standard deviation, and a bounded-variable tail inequality. Thus no iid extension is assumed here. Multiplication by nR yields Equation (14); applying both tails spends $\epsilon/2$ on each.

The common-range term can be expensive when a few users carry large weights. In contrast, Hoeffding keeps the separate R_u. We therefore select the construction using independent development responses, not by taking the smaller of two nominal 95% intervals after observing the test result. Both methods are then reported on the same test arrays. A data-dependent intersection would require a joint error allocation.

Repeated responses. If one user supplies several binary responses, the safe range is the sum of absolute coefficients. For instance, coefficients $(1, -1)$ and responses $(B, 1-B)$ produce error proportional to $2B-1$. The coefficient sum is zero, but the random contribution is not. With k perfectly repeated binary target events per user, each group contains only one distinct response. Effective target information is n_T/k, not n_T independent events.

A.7 Residual influence when calibration is reused

Suppose labeled means satisfy $\mathbb{E}W_L = X\vartheta + H\delta + \rho$. Write the standardized calibration mean as $\widetilde{\kappa} = \kappa + K\rho$, where K has nonzero columns only for the calibration responses. The calibration intervals cover $\widetilde{\kappa}$, the mean actually observed. If a score has labeled influence $w_s^\top$, then

$$\mathbb{E}\widehat{g} = a_D(\Delta - \tau) + d^\top b + \bar{q}_s^\top \widetilde{\kappa} + (w_s^\top - \bar{q}_s^\top K)\rho.$$

Consequently, under $|\rho_j| \le v_j$, valid residual penalties are

$$R_s = \sum_j |(w_s - K^\top \bar{q}_s)_j| v_j,$$
$$R_a = \sum_j |(w_a - K^\top \bar{q}_a)_j| v_j.$$

This is the influence left after the residual already entering the calibrated discrepancy has been removed. Charging $|w|^\top v$ and a separate full calibration-residual penalty would count some effects twice. In the four-response design, $K = [I, -I]$ on the calibration block. The implementation applies the same borrowing factor to both the bias query and this correction.

The protected residual experiment specifies a bounded smooth conditional-mean effect and uses its valid envelope. The omitted-protection experiment uses the same type of effect with zero protection and is marked as misspecified. The residual allowance is an input with a stated scope; it is not inferred from final evaluation labels.

A.8 Proof of the selected-decision guarantee

Let $\mathcal{E}_C$ be the simultaneous calibration event. Let $\mathcal{E}_s$ cover every score's sampling deviation and $\mathcal{E}_a$ every borrowing level's slope deviation. On their intersection, the bias support and residual bounds imply

$$F_s(d, \lambda) \le a_D(\Delta(d) - \tau), \qquad F_a(\lambda) \le a_D$$

for every allowed pair. If the selected pair has both margins positive, then $a_D > 0$ and $\Delta(d) > \tau$. Thus an accepted gain at or below the threshold can occur only outside the intersection. A union bound gives failure probability at most $\alpha_C + \alpha_s + \alpha_a$. No independence among the three events is needed. Testing a score for one pair and a slope for another would not support the implication, which is why the algorithm requires the same pair.

The main Shared design uses $M = 120$, $L = 3$, and $(\alpha_C, \alpha_s, \alpha_a) = (.005, .025, .020)$. Its calibration variants use the same three-level borrowing family and allocation. Fixed uses the same total allocation with $L = 1$, Five uses it with $L = 5$, and Zero retains $L = 3$ while masking the positive-borrowing rows. Stage aware uses $L = 1$, no calibration bias, and $(\alpha_s, \alpha_a) = (.02, .03)$. Calibration is shared across queries, while score and slope budgets are $\alpha_s/(ML)$ and α_a/L for the three-level Shared family; each method uses the allocation appropriate to its registered borrowing

family. Stage aware has $L = 1$, no calibration bias, and uses $(.02, .03)$ for score and slope. Zero retains $L = 3$ and the Shared allocation, but masks the positive-borrowing rows. It is a mechanism ablation, not a substitute for Stage aware.

Algorithm 1. A protected recommendation change
Input: stage labels, calibration, target feedback, fixed candidates, reference, borrowing levels, and error budget.
1. Check support and estimability of each required score and slope function.
2. Fit Stage aware and Pooled maps using all permitted labels.
3. Build the shared calibration set from the fixed generator directions.
4. For each candidate and borrowing level, merge response coefficients by identity and user.
5. Compute its score and slope sampling radii and signed bias support bounds.
6. Subtract the deployment and remaining model-residual allowances.
7. Keep only pairs whose score and slope margins both exceed the numerical tolerance.
8. Select the passing pair with largest score margin; use a fixed tie order.
9. Otherwise retain the reference; report unsupported, inconsistent or unresolved inputs separately.
10. Reveal the reserved evaluation labels only after this decision is fixed.

Corollary 1 (Independent development)**.** *Suppose development data select the representation, retained shift basis, finite candidate and borrowing families, and error allocation. If the protected-design conditions and Theorem 1 hold conditional on almost every development realization, the same unconditional error bound holds after averaging over development.*

Proof. Conditional on development, apply Theorem 1. The conditional error probability is at most the fixed total α. Taking expectation preserves that inequality. The statement requires the conditional assumptions; statistical independence of a split does not by itself establish a correct measurement model or a future transport allowance. □

A.9 Decision margins, power, and information requirements

The non-calibration margins U_s, U_a in the main text are observable diagnostics for a fixed pair. If either is nonpositive, any nonnegative calibration penalty on that side leaves the pair blocked. With a signed support, its actual value must be used: a sufficiently negative bias bound can improve a margin. The generic diagnostic is therefore to compare U_s with $h_{\mathcal{C}}(q_s)$ and U_a with $h_{\mathcal{C}}(q_a)$, rather than to assume every bias bound is nonnegative.

Let $b_s = q_s^\top \delta$ and $b_a = q_a^\top \delta$ in a zero-residual model. Suppose a calibration precision event guarantees

$$h_{\mathcal{C}}(q_s) - b_s \le \rho_s, \qquad h_{\mathcal{C}}(q_a) - b_a \le \rho_a$$

with failure probability at most β_C. Suppose lower-tail sampling deviations are bounded by t_s, t_a with failure probabilities β_s, β_a. A sufficient power condition for a fixed true gain Δ is

$$a_D(\Delta - \tau) > 2\eta_d + e_s + t_s + \rho_s, \qquad a_D > e_a + t_a + \rho_a. \tag{15}$$

The full pair then passes with probability at least $1 - \beta_C - \beta_s - \beta_a$. Sampling radii and the design must satisfy the stated inequalities on the events being used. In a random design, failure of that design condition needs its own probability term. The error bound for an incorrect acceptance and the probability of passing a valid action are different guarantees.

The slope condition explains why a small threshold does not eliminate calibration needs. The score contribution of slope uncertainty is multiplied by τ, while confirmation of a positive slope still needs sufficient information. An accurate score alone therefore does not give the power of the full rule.

For comparison, a bounded projection with variance σ_h^2 and range $2\|h\|_1$ has empirical Bernstein uncertainty of order

$$\sqrt{\frac{\sigma_h^2 \log(J/\alpha_C)}{n_C}} + \frac{\|h\|_1 \log(J/\alpha_C)}{n_C}.$$

After non-calibration errors leave room ρ, the associated sufficient sample order contains both σ_h^2/ρ^2 and $\|h\|_1/\rho$. The faster term dominates only when relevant variance is sufficiently small. Positive residual variance restores the square-root term. More precise calibration is valuable only if its reduction reaches the remaining margin in both parts of Equation (15).

A.10 POSITIVE-SIGN CALIBRATION LOWER BOUND

Theorem 2 (Magnitude calibration). *Let $0 < a_0 < a_1 < 1$, $0 < c \le a_0$, and $c/a_0 > \tau \ge c/a_1 > 0$. There exist two compatible models with the same historical and target-feedback distributions and positive current slopes a_0, a_1. For a fixed recommendation pair, their target gains are $c/a_0, c/a_1$. If each current calibration user supplies one balanced binary class label and its feedback, an acceptance rule with probabilities at least $1 - \beta$ and at most α in the respective models requires Equation* (11)*, for $\alpha + \beta < 1$.*

Proof. Take two items, reference item 1 and proposed item 2, and set $r_1 = (1-c)/2$, $r_2 = (1+c)/2$. In world j, set $\theta_j = (1 - a_j)/2$, zero context effects and zero deployment error. Then

$$p_1^{(j)} = \frac{1}{2} - \frac{c}{2a_j}, \qquad p_2^{(j)} = \frac{1}{2} + \frac{c}{2a_j},$$

which are physical probabilities because $c \le a_0$. Both worlds have the same target-feedback means, but gain c/a_j. Give them an identical historical response law, as the partial-sharing model permits. Each calibration response has $Y \sim \text{Bernoulli}(1/2)$ and $W \mid Y = y \sim \text{Bernoulli}(\theta_j + a_j y)$. Its KL divergence between worlds equals

$$\tfrac{1}{2}\,\text{kl}((1 - a_0)/2, (1 - a_1)/2) + \tfrac{1}{2}\,\text{kl}((1 + a_0)/2, (1 + a_1)/2) = \text{kl}((1 + a_0)/2, (1 + a_1)/2).$$

The product experiment has N_D times this divergence; historical and target-feedback observations add zero. Mapping all observations and any shared randomization to the binary acceptance event cannot increase KL. Its divergence is at least $\text{kl}(1 - \beta, \alpha)$, yielding the result. □

The testing argument follows standard change-of-measure reasoning (Kaufmann et al., 2016). It establishes a necessary calibration count in this observation model, not an implementation-specific slope gate. Both signs are already positive. When $\tau = 0$, this magnitude construction no longer separates acceptable and unacceptable worlds. Direct target labels, known equal stage slopes, or different side information also change the experiment.

For interior slopes separated by h, the single-observation KL is of order h^2. A direct binomial calibration test achieves the corresponding inverse-square order in this symmetric two-world experiment. The matching inverse-square order concerns this local two-world experiment.

B EXPERIMENTAL PROTOCOL

B.1 STUDY IDENTITY AND OBSERVATION DESIGN

The empirical study uses the released KuaiRand-Pure user–video interaction records (Gao et al., 2022). We use `long_view` as the sustained-view outcome Y and `is_like` as auxiliary feedback W. Recorded user and video identities define the observation units. Development data fix the representation, item support, reference slate and candidate family before the decision methods access the analysis records.

The item support contains $m = 16$ videos. Items 1–3 form the reference, and 120 distinct alternative size-three slates are fixed before inference. All methods receive this same family. Item identities remain fixed across repetitions. Item-level viewing summaries are obtained from the corresponding evaluation records; evaluation outcomes do not determine the reference or the candidate family.

Historical F and current D records supply (Y, W, x). Calibration C supplies the stage comparisons and can also enter the labeled fit. Target T exposes W to the decision procedures while withholding Y. Evaluation E is reserved before fitting and selection. Its outcome fields remain masked until every method has returned a decision. Observation roles control access to existing records, and a record retains one identity when reused in several calculations.

Stage assignment. Eligible KuaiRand-Pure interactions are ordered by recorded interaction time before any model fitting or method comparison. Earlier interactions form the historical stage F, later interactions form the current stage used for D and calibration C, and a still later block supplies target feedback T and held-out evaluation E. The boundaries are shared by all methods and do not

move when an observation budget changes. Users may appear in more than one stage; matched calibration additionally requires the same user to supply the required historical/current stage–class records.

The context representation is frozen using development data. The same bounded feature transformation is applied to the historical, current and target contexts. Within each stage, the analysis preserves the observed context and outcome-class composition of the eligible logged records; it does not rebalance the stages to force a common distribution. Labeled counts determine the regression weights. Target prevalence is not inferred from the class proportions of the labeled sampling design.

The inference design includes current columns $(1, Y, x)$ and historical level and slope indicator columns. Pooled and Stage aware regression use the same labeled records and context representation. User identities, rather than the number of uses of a record, determine the groups in the sampling bound. The representation and residual perturbations used in the controlled diagnostics are specified separately below.

B.2 Stage estimation and calibration records

The current-stage parameters (θ_0, a_D, β) and the historical-to-current differences (δ_0, δ_a) are estimated using the permitted labeled and calibration records. The fitted quantities enter the score and slope functions in Equation (4). The measurement relation is the model used for inference, not a rule for replacing observed viewing or like labels. Target and evaluation outcomes do not enter regression, calibration or candidate selection.

A matched calibration unit consists of four eligible records for one user: the two outcome classes in each of the historical and current stages. Write

$$Z_y = W_{Fy} - W_{Dy}, \qquad y \in \{0, 1\}, \qquad \kappa_y = \delta_0 + y\delta_a.$$

The observed stage gaps form one user-level vector (Z_0, Z_1). Their covariance is retained when calibrating a projection. An unrelated observation cannot replace a missing stage–class cell and still be treated as a matched response from that user. Ordinary calibration instead uses separately observed users in the four strata, keeping the distinction between matched and independent observations.

All four labeled responses in a calibration block also enter the labeled fit. A matched block contributes one dependent user vector to the sampling bound; an ordinary block contains four independent user vectors. Reused responses are charged once, and their coefficients are merged before uncertainty is calculated. A separate 256-block pilot fixes the ellipse orientation. Main calibration outcomes do not choose which directions receive protection.

The target relation in Equation (2) retains the deployment allowance $\eta_i = .001$. This is a specified inference allowance rather than an observed item-level residual. The protected-residual and deployment-error checks vary the discrepancy separately to examine the consequence of satisfying or violating that allowance. No final target or evaluation outcome is used to reduce it.

B.3 Sample roles and conditions

Historical fitting has $N_F = 24,576$ labeled responses and current fitting has $N_D = 192$. Calibration contributes $n_C = 448$ blocks; target feedback has $n_T = 16,384$ responses per item. Evaluation has $n_E = 2,048$ held-out records per item, reserved before fitting. The primary comparison changes no data across methods. The reported sample sizes are obtained by subsampling the corresponding frozen stage-specific pools; changing an observation budget changes only how many eligible records are used, not the stage boundaries.

The transition configuration uses $n_C = 512$ and $n_T = 8,192$. The current-rich configuration instead sets $N_D = 4,096$ with the other calibration-window observation budgets. Additional observation conditions use ordinary calibration or $N_F = n_T = 4,096$ for sampling limitation. These settings change the available records, while methods within a setting receive matched inputs.

Controlled channel checks examine a stage difference of $(.06, .10)$ and a weak-channel condition with $a_D = .05$ and $\delta_a = .005$. These and the near-threshold, zero-gain, residual and deployment-error constructions below are controlled diagnostics layered on the logged study, not naturally occurring KuaiRand-Pure regimes.

In the controlled near-threshold check, reference means are $.49$ and other means range from $.48$ to $.508$. In the zero-gain setting all means equal $.49$. At the boundary, all nonreference item means are $.50$ and reference means $.49$, so no candidate exceeds $\tau = .01$. The outcome mean checks use unrounded values with a 10^{-12} comparison tolerance.

The repeated-user setting repeats each target response four times within the same user. The protected-residual setting adds $.012\cos(\pi x_1)$ in the historical stage and minus one half of that effect in the current labeled stage. Calibration means are changed consistently. Its envelope is paid through the residual influence correction. The omitted version uses the same residual family but sets the protection to zero and is explicitly marked as a misspecified analysis.

B.4 Development, choices, and freezing

Development uses separate logged records and contains a finite initial search, a local refinement, and a final inference-choice comparison. The observation budgets defining the calibration, transition and current-rich settings are fixed before this search. The initial search has 18 configurations with eight replicates each; the refinement has 12 configurations with 16 replicates each. Within the frozen information settings, the inference-choice comparison uses 12 replicates per setting, five probability allocations and two sampling bounds. These are paired candidate analyses, not extra independent datasets.

The five candidate allocations $(\alpha_C, \alpha_s, \alpha_a)$ are

$$(.005, .035, .01), \quad (.01, .03, .01), \quad (.01, .02, .02),$$
$$(.005, .025, .020), \quad (.015, .02, .015).$$

Development selects $(.005, .025, .020)$ for Shared and its calibration variants, together with Fixed and Five, allocating more error budget to the score while retaining separate protection for calibration and the slope. Stage aware retains its separately selected $(.02, .03)$ score/slope allocation. The methods use a fixed tie order. After these settings and the 13 analysis configurations are fixed, the records for each repetition are assigned their observation roles, including the reserved evaluation partition, before the decision methods are run.

The main borrowing grid is $\{0, .25, .5\}$. The five-level ablation uses $\{0, .25, .5, .75, 1\}$, while Fixed uses only $.5$. Zero uses the three-level probability accounting and masks positive levels. Changing the method's family changes its sampling allocation; this is part of the comparison, not an unreported gain in information.

The supplementary representation, size, and deployment-error checks were specified after the main study, before opening their own analysis outcomes. The direct-contrast purchase analysis was added after inspecting the coordinate-interval purchase output. It reuses the exact same purchased labels and is identified as a same-array comparator audit, not a new independent primary result. The feature map, candidate family, borrowing grid and primary comparisons remain fixed once the analysis outcomes are opened.

B.5 Statistics and evaluation

Within a replicate, all methods receive the same permitted records and use the same reference and candidate family. Evaluation records are reserved before fitting and selection. Their sustained-view labels are masked from all development, calibration, inference and slate-selection procedures, and are revealed only after every method has fixed its decision. These are held-out fields of existing KuaiRand-Pure records.

For empirical reporting, let p^{eval} denote the evaluation-side item-rate vector underlying the recorded Mean gain column, and let $\bar{Y}_E$ denote the item means in the replicate's held-out evaluation records. The corresponding summaries are $d^\top p^{\mathrm{eval}}$ and $\widehat{\Delta}_E = d^\top \bar{Y}_E$, both zero on retention. Neither evaluation summary enters the decision rule. These empirical rates are distinct from the population mean p in the theoretical guarantee. In the controlled boundary and residual checks, the specified reference condition supplies the target contrast for the error diagnostic.

Acceptance and error-rate summaries use two-sided Clopper–Pearson intervals (Clopper & Pearson, 1934). The controlled error event is acceptance with $\Delta \leq \tau$. Under independent repetitions of

the specified diagnostic, zero errors among 200 repetitions give a 95% upper endpoint of 1.83%. No-acceptance conditions have an undefined conditional error fraction; the unconditional fraction remains defined. An empirical gain estimate below τ is not by itself an observed violation of the population guarantee.

For each primary method contrast, form the paired gain difference within every replicate. Resample these 200 pairs 10,000 times and use percentile tails $.025/4$ and $1 - .025/4$ for the four central primary comparisons. These are Bonferroni-adjusted bootstrap approximations, not exact simultaneous confidence sets. Other reported bootstrap intervals use pointwise 95% tails. The same independent resampling seed is attached to each named contrast. Single-method intervals use the same resampling indices within a condition. Methods with identical value records therefore have identical marginal intervals, as well as an exactly zero paired difference.

The paired resampling unit is one complete analysis repetition, retaining the decisions and evaluation values of all compared methods together. Items, candidate slates, borrowing levels and optimizer calls are not separate replications. The 200 repetitions describe variability under the frozen logged-data sampling protocol; repeated use of a record does not increase the number of distinct users. User dependence within each analysis remains in the sampling bounds. Candidate-wise precision and full-decision value have different denominators and are reported separately.

C Complete results and mechanism checks

C.1 Central comparison and all primary conditions

Table 5: Calibration window, 200 paired replicates. Acceptance includes exact 95% binomial intervals. Mean gain and held-out Evaluation are the empirical summaries defined in Appendix B.5, in percentage points and including retention. Positive λ counts replicates accepting a positive borrowing level. Directions count protected calibration projections, not sampling events.

Method	Acceptance (%)	Mean gain	Evaluation	Positive λ (%)	Directions
Shared	94.0 [89.8, 96.9]	16.22	16.16	52.0	2
Stage aware	79.0 [72.7, 84.4]	13.48	13.34	0.0	0
Ellipsoid	92.5 [87.9, 95.7]	16.04	15.98	49.5	12
Box	59.0 [51.8, 65.9]	10.18	10.02	4.5	2
Projected	92.5 [87.9, 95.7]	16.06	16.00	49.0	121
Zero	58.0 [50.8, 64.9]	9.37	9.28	0.0	0
Fixed	93.0 [88.5, 96.1]	15.93	15.88	93.0	2
Five	91.0 [86.1, 94.6]	15.58	15.50	54.0	2

The four primary paired contrasts are reported with the main results in Table 2.

Table 6: All 13 observation and diagnostic configurations, overall gain in percentage points. Each entry averages 200 paired replicates. Proj.: Projected; Ellip.: Ellipsoid; Stage: independent Stage aware. The asterisk marks deliberately omitted residual protection, not a valid-method superiority comparison. Zero gain and boundary remain separate conditions.

Setting	Shared	Box	Proj.	Ellip.	Zero	Stage	Fixed	Five
Calibration window	16.22	10.18	16.06	16.04	9.37	13.48	15.93	15.58
Transition	14.16	13.08	14.02	14.10	12.64	14.41	14.34	13.88
Current labels plentiful	16.37	16.20	16.30	16.34	16.18	16.85	16.31	16.26
Ordinary calibration	15.92	15.70	15.84	15.88	15.62	16.44	10.85	15.36
Sampling limited	0.42	0.34	0.46	0.47	0.35	0.77	0.53	0.18
Larger stage difference	14.98	9.75	13.62	14.82	9.55	12.54	15.10	14.48
Weak channel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Near threshold	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zero gain	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
At the threshold	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Repeated target users	3.10	2.05	3.02	3.05	2.00	4.56	4.05	2.55
Protected residual	1.22	1.00	1.18	1.20	0.98	2.47	0.20	0.85
Omitted protection*	11.20	11.05	11.18	11.19	11.00	13.98	5.60	9.45

Shared selects 21 distinct accepted slates in the calibration window. Its most frequently selected slate occurs in 92 of 200 replicates; the target vector and candidate family permit multiple competitive alternatives. Ellipsoid remains close in mean gain but differs from Shared in the accept/retain outcome in 11 replicates; Projected is also comparable rather than identical in value. The held-out evaluation summary remains separate from the item-level empirical gain summary. Neither supplies labels to the decision rule. Under ordinary calibration, Shared accepts in 181 of 200 replicates, and 26 accepted decisions use positive borrowing.

C.2 Borrowing and bottlenecks

For every registered candidate and positive borrowing level, we retain its score center, sampling radius, signed calibration penalty, residual penalty and final margin. The decomposition below uses the same candidate at zero borrowing as its reference. It averages across all candidates, so its mean can be negative while a useful subset improves. The standalone Stage aware comparison also changes probability accounting and is reported separately in Table 2.

Table 7: Borrowing decomposition for the same candidates in the calibration window, multiplied by 100. Center change plus radius saved minus calibration penalty equals the margin change because the residual allowance is zero and the deployment term is unchanged. Entries describe the improvement score in feedback units; outcome gains are reported separately.

λ	Center change	Radius saved	Calibration	Margin change	Slope penalty
0.25	-0.002	+0.011	0.007	+0.002	1.910
0.5	-0.005	+0.015	0.023	-0.013	3.820

Table 8: Candidate–replicate percentages with sequential categories: score sampling blocked, slope sampling blocked among the remainder, calibration blocked among the remainder, and both checks passing. The 120 candidate slates per replicate do not constitute 120 independent datasets.

λ	Score sampling	Slope sampling	Calibration	Passing
0	72.30	7.10	0.00	20.60
0.25	68.40	1.80	8.70	21.10
0.5	69.10	0.70	10.80	19.40

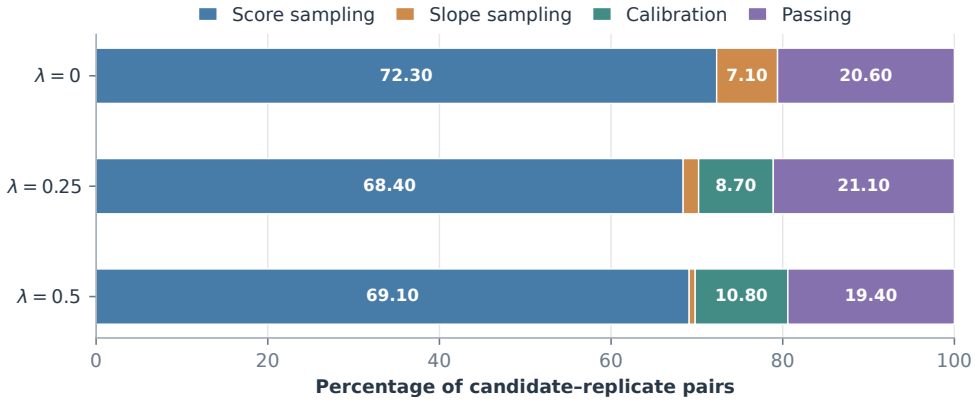


Figure 4: The candidate-level decomposition. Many slate alternatives have insufficient score precision even when another candidate in the same replicate passes. Consequently, the affected candidate fraction and the fraction of complete decisions changed by calibration answer different questions.

Calibration among candidates that clear both sampling checks. The sequential percentages in Table 8 use all candidate–replicate pairs as their denominator. Let b_{score}, b_{slope} and b_{cal} denote the sequential blocked percentages. Among candidates that clear both sampling checks, the calibration-blocked fraction is

$$100\,\frac{b_{\text{cal}}}{100 - b_{\text{score}} - b_{\text{slope}}}\,\%.$$

Using the displayed, rounded entries gives approximately 29.2% at $\lambda = .25$ and 35.8% at $\lambda = .5$. Thus calibration blocks roughly one third of the candidates that survive sampling at the positive borrowing levels, while the unconditional percentages retain all candidate–replicate pairs as their denominator.

C.3 Calibration and sampling constructions

The empirical Bernstein calibration radius has a sample-variance component and a finite-range remainder. Both use the same protected direction count. Betting inverts a valid bounded-mean test without this particular additive remainder. Its improvement in this experiment is therefore part of the interval implementation, while Shared versus Box with betting fixed isolates geometry. The negligible one-sided support gain shows that the theoretical reduction need not create a large value difference in every configuration.

Table 9: Same-array inference sensitivity, calibration window. The primary construction uses betting calibration and Hoeffding sampling. The two-sided support variant changes only the support function; the remaining rows change one concentration construction. Gain reports mean overall gain. Error is the percentage of all repetitions accepting a contrast at or below the evaluation threshold; the population guarantee is stated separately in Theorem 1.

Construction	Accept (%)	Gain	Error (%)
Betting calibration + Hoeffding score	94.0	16.22	0.0
Two-sided support	93.5	16.16	0.0
Empirical Bernstein calibration	58.0	9.37	0.0
Hoeffding calibration	58.0	9.37	0.0
Empirical Bernstein score / slope	82.0	14.12	0.0

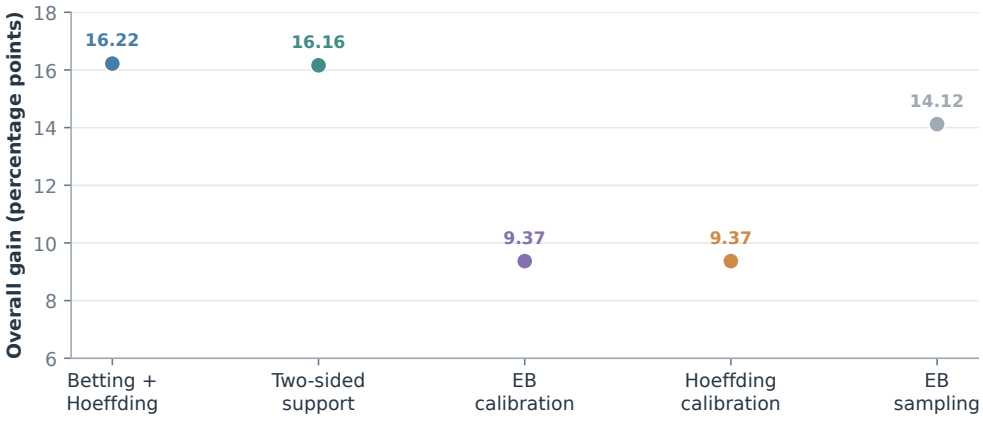


Figure 5: Value depends on the interval construction as well as calibration geometry. Points show mean overall gain for the same-array sensitivity comparisons. These test comparisons did not select a new default rule.

For the two calibration generators in the worked example, Table 10 separates the empirical Bernstein radius into its variance and range terms. With $L_h = 2\|h\|_1$, its radius is

$$e_h = \sqrt{2\widehat{v}_h \log(4J/\alpha_C)/n_C} + \frac{7L_h \log(4J/\alpha_C)}{3(n_C - 1)}.$$

The extra factor two inside the logarithm is retained in this sensitivity implementation. Both methods use the same user vectors, protected generators, and total calibration budget. Betting bounds are asymmetric endpoints, not a radius split into these two terms.

Table 10: Calibration components for the worked example in Table 3. Centers, radius components, and interval endpoints are multiplied by 100. The directions have unit ℓ_1 norm. The pair explains the fixed-candidate result; aggregate comparisons use all replicates.

Generator	h_0	h_1	Center	EB variance	EB range	Betting low	Betting high
1	0.651	-0.349	0.07	3.12	7.70	-3.58	3.69
2	0.381	-0.619	-1.25	3.12	7.70	-5.11	2.15

C.4 Representation, candidate family, and deployment error

Table 11: Controlled supplementary designs layered on the logged study. Value columns are percentage points. Candidate rows list items, slate size / number of alternatives. Accept and Error refer to Shared. These are diagnostics rather than additional naturally occurring KuaiRand-Pure regimes; the last two rows violate the declared deployment bound.

Design	Replicates	Shared	Stage	Fixed	Accept (%)	Error (%)
Cosine 8	100	18.05	13.39	17.80	100.0	0.0
Cosine 32	100	9.75	12.19	9.60	65.0	0.0
Cosine 64	100	3.05	4.00	3.00	20.0	0.0
32 items, 5 / 65	40	15.95	14.18	15.80	100.0	0.0
32 items, 5 / 250	40	16.48	14.32	16.22	100.0	0.0
40 items, 10 / 1,000	40	15.72	12.80	15.55	97.5	0.0
40 items, 10 / 5,000	40	15.96	12.23	15.72	95.0	0.0
2η at boundary	200	0.00	0.00	0.00	0.0	0.0
5η at boundary	200	0.03	0.07	0.04	3.0	3.0

The controlled representation study uses 100 replicates for each frozen cosine basis. All coefficients are nonzero and have total norm .025. Each dimension is a separate correctly specified design, rather than a low-dimensional truncation silently omitting coefficients. This tests growing effective signal dimension and estimation burden, not a claim of arbitrary nonlinear channel validity.

The candidate study uses 40 replicates per design and preserves the sampling probability cost for every registered pair. At 32 items the slate size is five; at 40 items it is ten. Families contain 65, 250, 1,000 or 5,000 response-independent alternatives. These are finite candidate menus, not exhaustive catalog enumeration. Acceptance remains high but decreases slightly for the largest families. The larger support, slate size and family are therefore reported together. The value changes therefore reflect the combined design rather than M alone.

The deployment-error stress keeps the declared $\eta = .001$ but directs the actual residual against the reference and toward alternative items. At two and five times that magnitude, the mean model no longer lies within the declared allowance. The target landscape remains at the practical boundary, so every strict acceptance is a false acceptance. These are violations of the guarantee's premise, not additional valid competing methods.

D Information purchase and reproducibility

D.1 Resource accounting

The common information consists of historical labels, $N_D = 192$ current labels, target feedback, and independent development information. Additional purchases are compared at the same replicate-specific budget. A matched calibration purchase has 448 successful units and four labeled responses per unit. Screening costs follow a negative-binomial resource model with a declared success-probability assumption of .8; this value is not estimated from a KuaiRand-Pure field. The model

charges one label unit per failed screen and .5 per successful match. Thus

$$B = 4n_C + N_{\text{screen}} + .5n_C.$$

This is a declared relative resource model. The study does not infer actual monetary prices or feedback delay from a dataset field name.

For ordinary-label price $c \in \{1, 2, 4, 8\}$, Current spends $\lfloor B/c \rfloor$ labels in the current conditional design and buys no matched calibration. Its Stage aware regression uses all common historical labels, common current labels and newly purchased labels. Direct and PPI buy $\lfloor \lfloor B/c \rfloor / m \rfloor$ representative target labels per item. Labels are not obtained through free outcome-class screening. Unspent rounding slack is retained in the cost record. Prediction development uses 512 representative observations per item and is a common sunk input, separate from the additional purchase budget.

Table 12: Same-budget purchases. The first three value columns use direction-level inference; the two coordinate variants retain their original results. All values include retention and are in percentage points. Label counts are averages after integer rounding, not independent sample sizes for uncertainty intervals. Shared purchases 448 matched blocks and has mean gain 16.22 points at all four prices.

Price	Current	Direct	PPI	Direct coord.	PPI coord.	Current labels	Target labels
1	17.04	15.21	12.65	1.38	0.00	2128.6	2120.2
2	16.23	7.45	3.06	0.19	0.00	1064.1	1056.2
4	13.82	2.45	0.38	0.02	0.00	531.8	526.6
8	7.76	0.43	0.08	0.00	0.00	265.6	256.0

D.2 Direct and proxy-assisted target inference

Direct contrast uses $\widehat{\Delta} = d^\top \bar{Y}_L$ and a two-sided Hoeffding radius

$$e_D(d) = \sqrt{\tfrac{1}{2} \log(2M/.05) \sum_i d_i^2 / n_i}.$$

Its acceptance margin is $\widehat{\Delta} - e_D(d) - \tau$. A coordinate version instead computes exact binomial intervals for each p_i and uses their rectangular support. Both are valid for the same representative target labels. The direct-contrast version is included because coordinate protection can be unnecessarily conservative for a fixed decision family.

For the PPI-style comparator, independent development learns bounded item-specific predictions $f_i(W)$ from the two observed feedback classes. The target estimator is

$$\widehat{p}_i^{\text{PPI}} = \frac{1}{n_T} \sum_{T_i} f_i(W) + \frac{1}{n_i} \sum_{L_i} \{Y - f_i(W)\}. \tag{16}$$

Labeled and unlabeled observations follow the same target joint law, including the target-stage residual. The predictor is fixed before these samples. Therefore Equation (16) has expectation p_i without a correct prediction model, following the residual-rectification principle (Angelopoulos et al., 2023).

For finite-sample selection, let s_i be the full possible range length of $Y - f_i(W)$, determined from all four binary pairs, and $t_i = |f_i(1) - f_i(0)|$. The PPI contrast radius is

$$e_P(d) = \sqrt{\tfrac{1}{2} \log(2M/.05) \sum_i d_i^2 (s_i^2/n_i + t_i^2/n_T)}.$$

The two samples are independent and all candidate errors are protected. A component-interval version uses betting intervals for prediction and residual means and takes their coordinate support. Both implementations are reported in the purchase table. They are finite-sample wrappers around residual correction, not reproductions of the asymptotic efficiency guarantees of PPI++ (Angelopoulos et al., 2024). With a weak like–view relation, extra predictions may reduce variance little while the finite-range term still increases; direct labels remain a substantive competitor.

The contrast-based Direct and PPI computations are an additional audit of the exact same purchased arrays. They were not used to select the primary Shared configuration. Figure 3 uses these stronger direction-based comparators, while the table preserves the original coordinate versions.

D.3 Availability and the role of the transport allowance

The timely-outcome setting allows direct target labels at their stated prices. In the delayed-outcome setting these purchases are unavailable at the decision time; they are not assigned zero gain as an algorithmic result. Current labeled measurements may come from an earlier mature cohort. The measurement method then still needs a D–T assumption or an independently supported allowance for the relevant population change.

An audit of mature earlier cohorts can constrain their own discrepancy. Applying that bound to a later cohort adds a persistence requirement. Neither a small fitted residual nor a failure to reject a stage difference establishes future transport. The explicit $1\times, 2\times, 5\times$ deployment-error study shows how the decision behaves as that requirement is violated. At 5η, false acceptance occurs in 3.0% of Shared replicates and 7.0% of Stage aware replicates. No final target outcome is used to reduce the declared η.

D.4 Implementation checks and scientific scope

The analysis stores KuaiRand-Pure record identities, observation-role assignments, response counts, matched user vectors, target counts, frozen designs, per-replicate decisions and held-out evaluation values. Counts summarize observed binary feedback without creating additional independent users. Labels used in more than one calculation are charged once and their influence is merged.

Implementation checks cover bias signs, required linear identities, support bounds, ellipse containment, and compressed versus explicit user-level calculations. The evaluator reveals the reserved labels after selection. Reported gains, acceptance rates, paired contrasts, intervals and figures are summaries of the frozen decision and evaluation records.

The empirical layers have distinct roles. The main KuaiRand-Pure study compares complete rules under matched and ordinary observation designs for additive fixed-support slate replacement. The purchase study changes which new observations are bought. Representation and larger-family studies examine estimation burden and finite-family selection cost. Boundary, residual and misspecification constructions are controlled diagnostics of the stated error event and its assumptions, not additional natural KuaiRand-Pure regimes. A geometric advantage at one query, a changed selected slate, and a better total-cost decision are recorded as separate findings.