Three Tests for an Adoption-Aware Crop Recommender: Sensitivity, Permutation, Placebo

Vairaaj Bindal [1]

Abstract

We evaluated an adoption-aware crop recommender on Ethiopian Living Standards Measurement Study panel data (n = 8,744 cultivated plots) using three classical observational tests: Rosenbaum signed-rank sensitivity, a sign-exchangeable permutation test on matched-pair differences, and a stratified label-shuffle placebo. The recommender targets Expected Realized Benefit (ERB), the product of predicted yield gain and predicted adoption probability, against an accuracy-only baseline that maximizes yield alone. The aggregate ERB lift of +50 kg/ha (bootstrap 95% CI $[+24, +74]$, n = 800 EA-disjoint test plots) clears Rosenbaum $\Gamma^\star \geq 5$ on every education subgroup and clears the permutation test at $p < 10^{-3}$. The published per-education subgroup pattern does not clear the placebo: the real per-subgroup means fall inside the placebo distribution at $p \in [0.14, 0.87]$. The placebo undercuts a published equity-targeting reading, leaving descriptive concentration of an aggregate effect as the defensible report. None of the three tests is new, and none is yet standard in observational offline-policy ML. We argue all three should be.

1. Introduction

Adoption-aware recommendation systems for smallholder agriculture optimize a hybrid of predicted yield and predicted adoption probability. The argument for adoption-aware over accuracy-only is straightforward: a recommendation a farmer will not follow generates no realized benefit, however high its predicted yield. The Expected Realized Benefit objective (Ruzzante et al., 2021) formalizes this trade-off:

$$\mathrm{ERB}(a \mid x) = \mathbb{P}(\text{adopt}\, a \mid x) \cdot \big(\mathbb{E}[Y \mid a, x] - \mathbb{E}[Y \mid a_0, x]\big), \tag{1}$$

where a_0 is the farmer's status-quo action.

Evaluating an ERB-maximizing policy from observational panel data is hard. The farmer's actual choice over crop and inputs depends on private information no survey records: plot quality, idiosyncratic risk preferences, household-internal negotiation. The off-policy-evaluation literature — inverse-propensity scoring (Horvitz & Thompson, 1952), self-normalized IPS (Swaminathan & Joachims, 2015), doubly-robust estimators (Dudík et al., 2011; Jiang & Li, 2016) — requires a known or estimable propensity. Neither holds here. Direct-method estimators — predict yield, predict adoption, combine analytically — are the standard alternative, but they offer no built-in protection against unmeasured confounding, and the literature on how to test their downstream claims is thin. External validation of the recommender's adoption head against the Duflo et al. (2011) Kenya SAFI fertilizer experiment reveals a ~ 50 percentage-point absolute calibration gap on out-of-distribution actions, which bounds the magnitudes any test in this paper can claim are correct (§5). The tests below ask which qualitative claims about the recommender survive that uncertainty.

This paper is about three tests we ran on one such recommender. We pulled them from the agricultural-economics, statistics, and observational-causal-inference literatures, where each has decades of use. Together they catch three different failure modes of a direct-method offline-policy claim:

- Rosenbaum signed-rank sensitivity (Rosenbaum, 2002): whether the aggregate lift survives unmeasured confounding of bounded strength.
- Sign-exchangeable permutation: whether the matched-pair sign-pattern is real, conditional on the magnitudes.
- Stratified label-shuffle placebo: whether an across-subgroup pattern is evidence of targeting, or just de-

[1]Lynbrook High School, San Jose, California, USA. Correspondence to: Vairaaj Bindal <vairaajb@gmail.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

scriptive concentration (the aggregate effect concentrating in whichever stratum happens to be largest, with no subgroup-specific mechanism).

None of these is new, and none is standard in the observational offline-policy ML papers we have seen. On the recommender we study, two pass and one fails, and that failure changes how the aggregate result should be reported.

2. The Recommender

The framework is a three-head architecture optimizing equation 1. The yield head $\hat{f}(a, x)$ is a hybrid model: a histogram gradient-boosting regressor with Gamma loss as base, with a literature-anchored linear response component for counterfactual actions (per-crop fertilizer-response coefficients from Vanlauwe et al. (2014); Mueller et al. (2012), improved-seed multiplier from Khonje et al. (2015)). The base is wrapped because a pure HistGBM on EA-disjoint splits shows near-zero predicted-yield deltas under counterfactual action perturbations, which collapses the policy-comparison signal. The yield head is wrapped in Mondrian-by-crop split-conformal regression intervals at 90% nominal coverage (empirical 88.7%).

The adoption head $\hat{g}(a, x)$ is a logistic baseline propensity on farmer and plot features, plus a per-tier log-odds penalty $\{T_0 : 0, T_1 : -0.4, T_2 : -1.5, T_3 : -3.2, T_4 : -2.8\}$ on adoption probability. The penalty magnitudes are anchored to the Ruzzante et al. (2021) meta-analysis of 367 adoption studies across 41 SSA and Asian countries. They are not monotone in the tier index: the crop-switch-only tier T_3 carries a larger barrier (-3.2) than the bundled T_4 (-2.8, full inputs plus a crop switch). This is deliberate. A bare crop substitution with no complementary inputs is a behaviorally unusual, higher-risk change in the adoption literature, whereas the complete input-and-seed package is the bundle extension services demonstrate and promote, and it draws somewhat higher uptake despite changing more on paper. Raw outputs are Mondrian-isotonic recalibrated by education quintile.

The policy compares argmax-ERB to argmax-yield on 800 EA-disjoint test plots from LSMS-ISA Ethiopia Wave 4 (Central Statistical Agency of Ethiopia & World Bank LSMS-ISA Team, 2019). The action space has five tiers, ordered by how far each departs from the status quo (how many components change), though the adoption penalties above are not monotone in this order: T_0 status quo, T_1 small fertilizer nudge, T_2 full input package on current crop, T_3 crop switch only, T_4 full inputs plus crop switch. Define $\Delta\mathrm{ERB}(a \mid x_i) = \mathrm{ERB}(a \mid x_i) - \mathrm{ERB}(a_0 \mid x_i)$, and let $\hat{a}_i^{\mathrm{aware}} = \arg\max_a \widehat{\mathrm{ERB}}(a \mid x_i)$ and $\hat{a}_i^{\mathrm{acc}} = \arg\max_a \widehat{\mathbb{E}}[Y \mid a, x_i]$ be the two policies' picks. The matched-pair difference is $d_i = \Delta\mathrm{ERB}(\hat{a}_i^{\mathrm{aware}} \mid x_i) - \Delta\mathrm{ERB}(\hat{a}_i^{\mathrm{acc}} \mid x_i)$, evaluated through the same calibrated heads for both policies (so any head-level miscalibration cancels in the contrast). Bootstrap 95% confidence intervals on $\bar{d}$ are from 2,000 plot-level percentile resamples; an enumeration-area (EA) cluster version, reported in §3, gives a comparable interval.

The mean $\bar{d} = +50$ kg/ha (95% CI $[+24, +74]$). The CI excludes zero. Read alone, this is a positive aggregate policy claim. The contested framing is one paragraph deeper: the per-education-quintile breakdown shows the largest absolute lift on the no-education subgroup (+53 kg/ha, $n = 543$) and smaller but positive lifts on more-educated quintiles (+38 to +49). The natural reading is that the recommender lifts outcomes most for the least-educated farmers — an identified equity-targeting claim. The three tests below check both readings.

3. Three Tests

Rosenbaum signed-rank sensitivity (Rosenbaum, 2002). Compute Rosenbaum's signed-rank statistic on the matched-pair differences. Report the critical $\Gamma^\star$: the largest unmeasured-confounder odds-ratio at which the one-sided p-value of $H_1 : \bar{d} > 0$ stays below 0.05. The agricultural-economics literature treats $\Gamma = 1.5$ as the conventional robustness threshold. Scope of the bound. Rosenbaum bounds were developed for matched-pair studies of treatment effects on observed outcomes; here the matched pairs are two policies assigned to the same plot, and the d_i are model-based contrasts evaluated through identical calibrated heads, not direct causal effects on yields. The Γ-bound consequently asks how strong hidden confounding in the observational adoption-training data would have to be to flip the sign of the model-based contrast, not to flip an unconditional causal effect on yield. This is a narrower interpretation than the classical one, but it is the one relevant to policy-comparison claims like ours. Shared-head inflation. Because both policies are scored through the same calibrated heads, confounding or miscalibration common to the two arms cancels in the within-plot contrast d_i. The Γ-bound therefore speaks to residual, arm-differential confounding, and a large $\Gamma^\star$ certifies the robustness of the contrast, not of either policy's absolute level. Shared structure could still inflate $\Gamma^\star$ by suppressing the variance of d_i; we bound

this two ways. First, $\Gamma^\star$ is computed only on the n^{dis} disagreement plots, so plots where the shared heads force $d_i = 0$ are excluded by construction. Second, perturbing the shared adoption head itself (the SAFI recalibration of §5, a ~ 50 pp downward shift in adoption probabilities) leaves the sign of the aggregate contrast intact, so the bound is not an artifact of the head's uncalibrated probabilities.

Sign-exchangeable permutation test. The aggregate is positive at bootstrap 95%, and the bootstrap interval is symmetric around the mean. A complementary question: under the null that the d_i are sign-exchangeable around zero, what is the probability of observing the realized $\bar{d}$ or larger? We randomly flip the signs of $\{d_i\}$ ($B = 2{,}000$ replications), compute $\bar{d}^{\text{perm}}$ each time, and report the two-sided p-value. Clustering. The sign-flip is plot-level, but plots within the same enumeration area (EA) share peer-adoption rates, micro-climate, and local extension capacity, so they are not independent. Rather than argue informally that the margin absorbs this, we ran the EA-level check directly: resampling and sign-flipping at the EA level (44 EAs, mean 18 plots each) give a cluster-robust aggregate of +50 kg/ha, 95% CI $[+33, +70]$, and a cluster-level sign-flip $p < 5 \times 10^{-4}$. Clustering does not change the conclusion.

Stratified label-shuffle placebo. Neither Rosenbaum nor permutation tells you whether the differences across education strata mean the recommender is actually keying on education, or whether the same pattern would show up under any relabeling of farmers that kept the subgroup sizes fixed. The placebo holds $\{d_i\}$ fixed, permutes the household-head education-quintile labels with a stratified shuffle that preserves marginal subgroup sizes, recomputes per-subgroup means $\bar{d}_s$ under the shuffle, and repeats $B = 2{,}000$ times. We report two-sided p-values against the resulting placebo distribution.

The placebo tests the null that education labels are exchangeable across plots given the observed d_i pattern. A $p < 0.05$ on at least one subgroup would be evidence of education-specific targeting beyond what structural correlates of education (landholding, market access, geographic clustering) already explain. A $p > 0.05$ leaves that null un-rejected: the across-strata pattern is then consistent with descriptive concentration of an aggregate effect, though failing to reject does not by itself establish it. A stronger version would restrict the shuffle to within EA clusters, preserving geographic and community-level composition; the marginal-size-preserving shuffle we use is conservative for the descriptive direction, since a stronger placebo can only fail to clear in more subgroups, not fewer.

The three catch different failure modes: Rosenbaum, hidden confounding; permutation, chance; the placebo, the descriptive-versus-targeting question at the subgroup level. A subgroup result has to clear all three before any one of them means what its authors want it to.

4. What the Tests Revealed

Table 1 summarizes the three tests applied to the recommender.

Subgroup	n^{dis}	$\Gamma^\star$	Placebo p
Aggregate	368	≥ 8.0	—
No education	257	≥ 8.0	0.14
Early primary	17	5.15	0.86
Late primary	47	≥ 8.0	0.87
Secondary	25	7.85	0.50

Table 1. The three tests on the recommender (hybrid heads, 800 EA-disjoint test plots). n^{dis} is the number of plots where the two policies disagree on the recommendation. $\Gamma^\star$ is the Rosenbaum critical odds-ratio at $\alpha = 0.05$; conventional agricultural-economics threshold is 1.5. The four strata account for 257+17+47+25 = 346 of the 368 aggregate disagreement plots; the remaining 22 are plots with missing household-head education, and a singleton post-secondary stratum ($n = 1$) produces no disagreements. All 368 enter the Aggregate row. Plot-level sign-exchangeable permutation p-values are $< 10^{-3}$ on aggregate and every subgroup (omitted from the table). Placebo p-values are two-sided against the distribution of per-subgroup means under stratified shuffles of education labels.

Aggregate. Rosenbaum clears $\Gamma \geq 5$ on the aggregate matched-pair sample. The conventional ag-econ robustness bar is $\Gamma = 1.5$. To flip the aggregate sign, an unobserved confounder would need to shift the matched-pair d_i-sign probability by an odds-ratio of at least 5 — which the technology-adoption literature does not document for any plausible omitted variable. The aggregate also clears the sign-exchangeable permutation at $p < 10^{-3}$. The aggregate +50 kg/ha is real, in the sense any observational claim can be real: it survives the two tests for hidden confounding and chance.

Per subgroup. The same two tests pass on each subgroup separately. But the placebo fails on all four. The actual per-subgroup means fall inside the bulk of the placebo distribution at $p \in [0.14, 0.87]$. In words: if you take the realized d_i pattern and shuffle the education labels randomly while keeping subgroup sizes

fixed, the per-subgroup means you would get by chance are indistinguishable from the per-subgroup means the recommender actually produced.

This is a non-trivial reframe of what the per-subgroup numbers say. The reading the table invites (the recommender targets less-educated farmers by treating education as a friction-reducer) is not supported by the placebo evidence. What the data does support is that (i) the aggregate is positive and unusually robust, and (ii) the no-education stratum is where most of the absolute lift falls, because that stratum is $543/800 = 68\%$ of the test set. The across-strata pattern is consistent with average per-plot behavior plus structural correlates of education (landholding, market access, geographic clustering), rather than with the education feature itself.

Small-n caveat. The placebo is under-powered on the early-primary stratum ($n^{\text{dis}} = 17$) and the secondary stratum ($n^{\text{dis}} = 25$). A larger test set might let one of those subgroups clear the placebo. But under-powered falsification reinforces the descriptive framing rather than weakening it: a targeting claim that cannot be falsified at this n should not be reported as targeting. If the pattern is real, a larger test set will catch it; until then, the safer report is descriptive.

Why the placebo matters here in particular. An identified equity-targeting claim and a descriptive-concentration claim imply different deployments. The first says deploy to populations like the Ethiopian no-education stratum specifically to reduce inequity. The second says deploy where you can; whoever in your test set is in the modal stratum will see most of the absolute lift. Both readings are compatible with the same +53 kg/ha number in a table. The placebo is the test that separates them, and skipping it would have left the stronger reading in the paper.

5. Discussion

The three tests are cheap. Rosenbaum and permutation are seconds of compute on 800 plots, and the placebo took under an hour on a laptop. None of them needs new theory, new estimators, or new data.

The most useful change is at the per-subgroup level. Bootstrap CIs and Rosenbaum bounds get reported in many observational ML papers already. Placebos on subgroup-targeting claims do not. The cost of skipping the placebo is the kind of overclaim we describe in §4: a real aggregate effect plus a descriptive concentration get presented as an identified targeting mechanism. Reviewers should be able to ask "did the placebo run?" of any observational offline-policy paper with a subgroup breakdown.

The three tests do not address every threat. They say nothing about the calibration of the heads themselves out-of-distribution. Validated against the Kenya SAFI fertilizer experiment (Duflo et al., 2011), the recommender's adoption head preserves the directional ordering of treatment effects but mis-calibrates the absolute magnitudes by ~ 50 percentage points on counterfactual actions. Rather than assert that the comparison's sign survives this gap, we recompute the lift under the correction: capping the adoption probabilities of treated actions at the SAFI-observed uptake rates (the treated-max, treated-mean, and control rates of 0.51, 0.43, and 0.33) shrinks the aggregate from +50 to +44, +38, and +26 kg/ha, and even at the most pessimistic control cap the interval excludes zero ($[+13, +39]$). The sign is robust to the calibration gap; the absolute kg/ha numbers are model-believed rather than calibrated against ground-truth realized yields. Four other limitations bear mention. First, the per-tier penalty magnitudes are anchored to a meta-analytic prior; a $\pm 25\%$ perturbation sweep in the parent work leaves the aggregate significant, while halving the penalties collapses it. Second, ERB beats a yield-minus-$\lambda \cdot$tier heuristic baseline at every λ tested (parent work). Third, the cluster-robust bootstrap and permutation are reported in §3; the within-cluster placebo restriction remains a clean follow-up. Fourth, the recommender is evaluated on a single country (Ethiopia); a Tanzania replication in the parent work does not reproduce the per-subgroup pattern, consistent with the descriptive framing.

6. Conclusion

The recommender's aggregate +50 kg/ha claim survives Rosenbaum and permutation; the per-subgroup targeting reading does not survive the placebo. Running all three tests costs little, and it changes what the paper is entitled to claim.

References

Central Statistical Agency of Ethiopia and World Bank LSMS-ISA Team. Ethiopia Socioeconomic Survey 2018-2019, Wave 4. World Bank Microdata Library, Catalog 3823, 2019. URL https://microdata.worldbank.org/index.php/catalog/3823.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In International Conference on Machine Learning, 2011.

Duflo, E., Kremer, M., and Robinson, J. Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya. American Economic Review, 101(6):2350–2390, 2011.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260):663–685, 1952.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In International Conference on Machine Learning, 2016.

Khonje, M., Manda, J., Alene, A. D., and Kassie, M. Analysis of adoption and impacts of improved maize varieties in eastern Zambia. World Development, 66: 695–706, 2015.

Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., and Foley, J. A. Closing yield gaps through nutrient and water management. Nature, 490(7419):254–257, 2012.

Rosenbaum, P. R. Observational Studies. Springer, 2nd edition, 2002.

Ruzzante, S., Labarta, R., and Bilton, A. Adoption of agricultural technology in the developing world: A meta-analysis of the empirical literature. World Development, 146:105599, 2021.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In Advances in Neural Information Processing Systems, 2015.

Vanlauwe, B., Coyne, D., Gockowski, J., Hauser, S., Huising, J., Masso, C., Nziguheba, G., Schut, M., and Van Asten, P. Sustainable intensification and the african smallholder farmer. Current Opinion in Environmental Sustainability, 8:15–22, 2014.