Aligning Incentives to Balance Covariates in Experiments with Selection Bias: Experiment Report

ANONYMIZED

Contents

Contents 0
1 Introduction 1
2 Simulation Study 1
2.1 Without Unobserved Confounders 1
2.2 Robustness to Model Mis-specification and Unobserved Confounders 3
3 Real World Data 5

Manuscript submitted for review to the 26th ACM Conference on Economics & Computation (EC'25).

1 Introduction

This report proceeds in two main parts. In **Section 2**, we conduct a comprehensive suite of simulation experiments: Section 2.1 presents results under a no-confounder design with three nonlinear outcome models (polynomial, sigmoidal, and sinusoidal), comparing the naive difference-in-means (DIM), inverse-propensity-weighting (IPW) estimators, and doubly robust AIPW in terms of ATE bias and variance; Section 2.2 then evaluates estimator performance under unobserved confounders and deliberate choice-model misspecification using the same outcome specifications. In **Section 3**, we apply these methods to a real-world dataset of 683 Prolific participants who rated two AI-generated videos (fantasy vs. sci-fi). We first analyze estimator accuracy without incentives, and then introduce a genre-targeted incentive mechanism to rebalance covariates, again comparing bias, variance, and choice-model feature importance. All simulation code and empirical data are provided in the github repo: https://github.com/papersubmission1935/paper-code/tree/master

2 Simulation Study

We begin by evaluating our incentivized experiment through a series of simulation studies. Specifically, we examine two settings—one with unobserved confounders and one without—and impose a deliberately imbalanced choice probability to reflect the primary focus of this paper. Throughout, the underlying utility-based choice model remains fixed: each feature vector $X_i \in \{0,1\}^4$ has

$$X_{i1}, X_{i3} \sim \text{Bernoulli}(0.8), \quad X_{i2}, X_{i4} \sim \text{Bernoulli}(0.2).$$

Without loss of generality, we let the latent utilities for treatments 0 and 1 be

$$U_{i0} = \varepsilon_{i0}, \quad U_{i1} = X_i^\top \theta^* + \varepsilon_{i1},$$

where $\varepsilon_{i0}, \varepsilon_{i1}$ are i.i.d. standard Gumbel noise and $\theta^* = (3, -3, 3, -3)^\top$. Consequently, the probability of assignment to treatment 1 is

$$\mathbb{P}(W_i = 1 \mid X_i) = \frac{\exp\left(X_i^\top \theta^*\right)}{1 + \exp\left(X_i^\top \theta^*\right)}.$$

2.1 Without Unobserved Confounders

When incentive gaps are unbalanced, even in the simplest no-confounder setting, propensity-score–based estimators (e.g. IPW) become highly unstable and can underperform the Difference-in-Means (DIM) estimator. We now introduce three distinct outcome definitions:

(1) Polynomial:

$$Y_0 = (X^T\gamma_0)^2 + X^T\theta^* + \eta_0, \qquad Y_1 = (X^T\gamma_1)^2 + X^T\theta^* + \eta_1,$$

(2) Sigmoid:

$$Y_0 = \sigma(X^T\theta^*) + \sigma(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sigma(X^T\theta^*) + \sigma(X^T\gamma_1) + \eta_1,$$

(3) Sine:

$$Y_0 = \sin(X^T\theta^*) + \sin(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sin(X^T\theta^*) + \sin(X^T\gamma_1) + \eta_1,$$

where $\sigma(X) = \frac{1}{1+e^{-x}}$ is the sigmoid function, and η_1, η_0 are i.i.d. standard Gaussian noise. γ_1, γ_0 are uniformly generated from $[0,1]^4$.

We generate 20 independent pairs (γ_0, γ_1), each drawn uniformly from $[0,1]^4$. For each pair, we simulate an experiment with $n = 1000$ units, drawing treatments and outcomes as described above. We then repeat the entire sampling and estimation procedure 100 times per (γ_0, γ_1) to obtain empirical estimates of bias and variance for each estimator. Table 1 reports the aggregated results across all parameter realizations.

In settings with large class imbalance, penalized logistic regression introduces substantial shrinkage bias, which can destabilize estimates of the underlying choice model. This reflects a classic bias–variance trade-off: although IPW estimators are unbiased in theory when all confounders are correctly specified, their high variability can lead them to underperform even a simple DIM estimator in practice. Augumented-IPW (AIPW), or "doubly robust," estimators mitigate this volatility by combining outcome regression with propensity-score weighting, yielding more reliable estimates under model misspecification. Nonetheless, when the true outcomes exhibit strong nonlinearity (e.g., polynomial relationships), AIPW still incurs nontrivial bias in estimating Y_0, and—across all three simulation scenarios—its variance remains markedly larger than that of the DIM estimator.

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	ATE variance
Polynomial	DIM	4.044098	0.485364	-3.558735	0.051760
	IPW	2.063882	0.048878	-2.015004	1.217703
	AIPW	0.345920	0.039557	-0.306363	0.166380
Sigmoid	DIM	0.449822	0.054365	-0.395457	0.007994
	IPW	0.544242	-0.012095	-0.556337	0.112538
	AIPW	-0.020972	0.003569	0.024541	0.098769
Sine	DIM	-0.013673	0.007796	0.005877	0.008633
	IPW	0.187496	-0.011561	-0.199057	0.042525
	AIPW	-0.079291	0.001283	0.080574	0.109859

Table 1. Comparison of three estimators under three outcome models. Red text highlights, for each outcome, the smallest absolute bias in ATE, Y_1, and Y_0, and the smallest ATE variance.

In the next phase, we deploy our low-switching incentivization policy to achieve substantially more accurate ATE estimates. We implement a simple two-stage design with a single policy switch. As established in our theoretical analysis, the optimal policy is characterized by a threshold λ: given estimates $\hat{p}(X)$ of the propensity score and $\hat{\theta}$ of the utility parameters, we assign a bonus $s^*(X)$ such that

$$\begin{cases} \text{Give bonus } s^*(X) \text{ to treatment 1, such that } \frac{e^{X^T\hat{\theta}+s^*(X)}}{1+e^{X^T\hat{\theta}+s^*(X)}} = \lambda, & \text{if } \frac{e^{X^T\hat{\theta}}}{1+e^{X^T\hat{\theta}}} \le \lambda, \\ \text{Give bonus } s^*(X) \text{ to treatment 1, such that } \frac{e^{X^T\hat{\theta}}}{1+s^*(X)+e^{X^T\hat{\theta}}} = 1-\lambda, & \text{if } \frac{e^{X^T\hat{\theta}}}{1+e^{X^T\hat{\theta}}} \ge 1-\lambda, \\ s^*(X) = 0, & \text{otherwise.} \end{cases}$$

In our simulations, we fix the threshold $\lambda = 0.4$ (other choices yield qualitatively similar results) and draw a total of $n = 1000$ units. We allocate $n_1 = 300$ to the first stage: after observing these 300 samples, we fit the choice model via penalized logistic regression and use the fitted parameters to construct the incentive function $s^*(X)$. Because penalized logistic regression tends to understate the true assignment imbalance, the post-incentive allocation will not be perfectly balanced at λ, but it still suffices to dramatically reduce selection bias. As Table 2 shows (with the best results highlighted in red), our incentivized design enables the AIPW estimator to achieve substantially lower bias—and markedly smaller variance—than in the no-incentive setting, across all three outcome models. From Table 2, we observe that the naive DIM estimator continues to exhibit nontrivial selection bias even under incentivization, although its bias is substantially reduced compared to the no-incentive case.

More importantly, all three estimators achieve uniformly better performance once incentives are introduced. This improvement stems from our low-switching incentivization policy to correct the original treatment-assignment imbalance, and we can stabilize the propensity-score weights and shrink both bias and variance across all estimators.

Outcome	Method	ATE bias	Y1 bias	Y0 bias	Variance
Polynomial	DIM	2.083975	0.505646	-1.578329	0.105427
	IPW	1.295498	0.063204	-1.232294	0.084574
	AIPW	0.023256	0.021915	-0.001342	0.013452
Sigmoid	DIM	0.232880	0.052415	-0.180465	0.006985
	IPW	0.327998	-0.006939	-0.334937	0.010912
	AIPW	0.003361	0.001430	-0.001930	0.012843
Sine	DIM	-0.057995	0.004677	0.062673	0.006599
	IPW	0.086523	-0.010082	-0.096605	0.007499
	AIPW	0.001363	-0.000660	-0.002023	0.013045

Table 2. Comparison of three estimators under three outcome models. Red text highlights, for each outcome, the smallest absolute bias in ATE, Y_1, and Y_0, and the smallest ATE variance.

Comparing Tables 1 and 2 reveals that our two-stage incentivization policy substantially reduces both bias and variance across all estimators, with the largest gains for IPW and AIPW. In the polynomial outcome, AIPW's ATE bias falls from 0.3459 to 0.0233 and its variance from 0.1664 to 0.0135; IPW's bias decreases from 2.0639 to 1.2955 and its variance from 1.2177 to 0.0846; by contrast, DIM's bias is halved (from 4.0441 to 2.0840) while its variance increases only modestly. Similar improvements occur for the sigmoid model and for the sine model. These comparisons confirm that aligning treatment-assignment incentives to improve covariate distribution sharply enhances the stability and accuracy of propensity-based estimators, driving AIPW biases to near zero with variances comparable to or below those of DIM.

2.2 Robustness to Model Mis-specification and Unobserved Confounders

In the previous section, we studied estimator behavior when all confounders are observed. To assess performance under hidden bias, we now introduce an unobserved confounder. Consider the same choice model, for each covariate we let

$$U_{i0} = \varepsilon_{i0}, \quad U_{i1} = X_i^\top \theta^* + \varepsilon_{i1},$$

with $\varepsilon_{i0}, \varepsilon_{i1} \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0, 1)$, and assign

$$W_i = \mathbf{1}\{U_{i1} > U_{i0}\}.$$

We then generate outcomes using the same three functional forms as before—continuous, sigmoidal, and sinusoidal:

(1) Continuous:

$$Y_0 = U_0 + (X^T\gamma_0)^2 + X^T\gamma_0 + \eta_0, \qquad Y_1 = U_1 + (X^T\gamma_1)^2 + X^T\gamma_1 + \eta_1,$$

(2) Sigmoid:

$$Y_0 = \sigma(U_0) + \sigma(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sigma(U_1) + \sigma(X^T\gamma_1) + \eta_1,$$

(3) Sine:

$$Y_0 = \sin(U_0) + \sin(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sin(U_1) + \sin(X^T\gamma_1) + \eta_1,$$

Here, $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the logistic function, and $\eta_0, \eta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are Gaussian noise terms. The vectors γ_0, γ_1 are drawn uniformly from $[0,1]^4$.

We first repeat the no-incentive experiments under this confounded design, employing the same estimators as in the previous section. Table 3 reports the results. In all three outcome scenarios, the simple DIM estimator attains the lowest variance and the smallest ATE bias, outperforming both IPW and AIPW. This result reflects that when confounders drive imbalanced assignment probabilities, propensity-score weights become highly variable, inflating the variance while the DIM estimator avoids weight instability.

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	ATE variance
Continuous	DIM	-0.645	0.681	1.326	0.061
	IPW	-1.240	0.177	-1.417	1.633
	AIPW	-3.361	0.196	3.557	0.324
Sigmoid	DIM	-0.089	0.067	0.156	0.008
	IPW	0.228	0.012	-0.216	0.118
	AIPW	-0.287	0.035	0.322	0.097
Sine	DIM	-0.107	0.027	0.134	0.013
	IPW	0.366	0.023	-0.343	0.063
	AIPW	0.272	0.053	-0.219	0.131

Table 3. Comparison of three estimators across outcome models. For each outcome, the smallest absolute ATE bias, Y_1 bias, Y_0 bias, and the smallest ATE variance are highlighted in red.

Next, we apply our two-stage, single-switch incentivization policy to the confounded scenario. As before, we use $n_1 = 300$ initial observations to fit the penalized logistic model and then construct the bonus function $s^*(X)$ at threshold $\lambda = 0.4$. Table 4 presents the resulting ATE bias and variance for each estimator across all three outcome models.

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	ATE variance
Continuous	DIM	-0.163	0.013	0.176	0.020
	IPW	0.311	-0.034	-0.345	0.028
	AIPW	-0.022	0.011	0.033	0.010
Sigmoid	DIM	0.013	0.044	0.031	0.004
	IPW	0.187	-0.004	-0.556	0.009
	AIPW	-0.025	0.008	0.033	0.010
Sine	DIM	-0.014	0.001	0.015	0.007
	IPW	0.155	-0.005	-0.159	0.010
	AIPW	0.019	0.012	-0.007	0.015

Table 4. Comparison of three estimators across outcome models. For each outcome, the smallest absolute ATE bias, Y_1 bias, Y_0 bias, and the smallest ATE variance are highlighted in red.

Comparing Table 3 (no incentives) with Table 4 (with incentives) demonstrates the dramatic impact of our two-stage policy on bias and variance under confounding. In particular, it highlights the exceptional robustness of our incentivization policy when paired with the AIPW estimator. In the continuous outcome, AIPW's ATE bias plummets

from −3.361 to −0.022(variance 0.324 → 0.010). For the sigmoidal model, bias improves from −0.287 to −0.025 (variance 0.097 → 0.010), and in the sinusoidal case from 0.272 to 0.019 (variance 0.131 → 0.015). These improvements confirm that incentivization restores covariate overlap, stabilizes propensity-score weights, and enables all estimators—especially AIPW—to achieve near-oracle accuracy even with hidden confounding.

Overall, these results underscore two key points. First, in the presence of unobserved confounders and an imbalanced covariate distribution, the simple DIM estimator exhibits very large bias, and propensity-score–based methods (IPW and even doubly robust AIPW) become highly unstable—often performing worse than DIM. Second, introducing our targeted incentivization to restore overlap significantly reduces both bias and variance for all three estimators.

In our final experiment, we assess the robustness of the incentivization policy under deliberate choice-model misspecification. Specifically, we generate treatment assignments according to

$$U_{i0} = \varepsilon_{i0}, \quad U_{i1} = X_i^\top \theta^* + \varepsilon_{i1}, \quad W_i = \mathbf{1}\{U_{i1} > U_{i0}\},$$

where $\varepsilon_{i0}, \varepsilon_{i1} \sim \mathcal{N}(0,1)$ are standard Gaussian noise (instead of the Gumbel noise assumed by the logistic model). We nevertheless fit a logistic regression to estimate the propensity scores, thereby introducing misspecification. Remarkably, thanks to the doubly-robust property of AIPW and our incentive-driven covariate balancing, the AIPW estimator continues to achieve the lowest bias and variance in almost every scenario. Table 5 reports the detailed results.

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	Variance
Polynomial	DIM	2.277	0.520	-1.757	0.156
	IPW	1.467	0.063	-1.403	0.030
	AIPW	0.013	0.020	0.006	0.012
Sigmoid	DIM	0.232	0.052	-0.179	0.008
	IPW	0.384	0.007	-0.391	0.007
	AIPW	0.004	0.002	0.006	0.011
Sine	DIM	-0.035	0.007	0.042	0.007
	IPW	0.107	0.009	-0.116	0.006
	AIPW	-0.006	0.001	0.006	0.011

Table 5. Comparison of three estimators across outcome models. For each outcome, the smallest absolute ATE bias, Y_1 bias, Y_0 bias, and the smallest ATE variance are highlighted in red.

3 Real World Data

In our real-world study, we recruited 683 participants on Prolific and asked them to evaluate two AI-generated short videos—one in the fantasy genre and one in sci-fi—on overall quality. Before viewing, we collected five key covariates for each person:

(1) Age range,
(2) Gender,
(3) Enjoyment of imaginative or emotional content,
(4) Curiosity about scientific or technological topics,
(5) Preferred movie genre.

Participants then indicated which video they would choose to watch; however, we had each participant view and rate both videos, thereby observing both factual and counterfactual outcomes. As expected, stated genre preference strongly predicted choice, so we implemented a genre-targeted

incentive: participants preferring sci-fi watched an extra 20-second advertisement before the sci-fi video, and similarly for those preferring fantasy.

First, we assess covariate balance before and after introducing this incentive.

Table 6. Preference Balance in Incentive and No-Incentive Data

Dataset	Choice	Fantasy	Sci-Fi	Total	Pr(Sci-Fi)	Pr(Fantasy)
With Incentive	Fantasy	96	59	155	0.381	0.619
	Sci-Fi	40	124	164	0.756	0.244
No Incentive	Fantasy	108	54	162	0.333	0.667
	Sci-Fi	43	159	202	0.787	0.213

The table shows that, under the incentivized design, of the 155 participants assigned to Fantasy, 96 (61.9%) actually preferred Fantasy and 59 (38.1%) preferred Sci-Fi. In the no-incentive condition, the corresponding proportions were approximately 65% vs. 35%, so our genre-targeted bonus improved overlap by roughly 3–5 percentage points. Although this increase in balance is modest, the reductions in treatment-control imbalance can yield substantially more stable and accurate average treatment effect estimates, as we demonstrate below.

We begin by examining the no-incentive condition, comparing six estimators:

(1) **DIM**: Naive difference-in-means estimator.
(2) **PSM**: Propensity-score matching estimator.
(3) **IPW–Logit**: Inverse-propensity-weighted estimator using a logistic regression choice model.
(4) **IPW–RF**: IPW estimator using a random forest choice model.
(5) **AIPW–Logit**: Augmented IPW (doubly robust) with logistic choice model and random forest outcome model.
(6) **AIPW–RF**: AIPW with random forest choice and outcome models.

For all AIPW variants, outcome regressions are fit by random forest. Table 7 presents the estimated average treatment effects along with variance estimates obtained from 200 bootstrap replications.

Table 7. Estimator Results with Bias and ATE Variance Without Incentive

Method	Y1 (bias)	Y0 (bias)	ATE (bias)	Variance
True	3.563 (+0.000)	3.684 (+0.000)	-0.120 (+0.000)	—
Naive	3.638 (+0.074)	3.869 (+0.185)	-0.231 (-0.111)	0.0173
PSM_ATT	3.638 (+0.074)	4.136 (+0.452)	-0.498 (-0.378)	0.0180
IPW_Logit	2.248 (-1.315)	1.659 (-2.024)	0.589 (+0.709)	0.0493
AIPW_Logit	3.572 (+0.009)	4.028 (+0.345)	-0.456 (-0.336)	0.0131
IPW_RF	2.933 (-0.631)	2.856 (-0.827)	0.076 (+0.196)	0.0411
AIPW_RF	3.578 (+0.014)	3.792 (+0.109)	-0.215 (-0.095)	0.0175

The “true” rating refers to the overall mean across all 364 participants. We observe a clear positive selection bias for both videos: participants who favor a given genre are not only more likely to watch that video but also tend to award it higher scores than the general population. Pure IPW methods suffer from excessive variability and often over-adjust, leading to unstable estimates. For the sci-fi video, the AIPW estimator with a logistic-regression propensity model achieves the lowest

bias and variance. However, it does not fully correct the bias for fantasy ratings. In contrast, the AIPW estimator that uses a random-forest propensity model is the only method to substantially reduce bias in the fantasy arm—though some residual bias remains.

Table 8 presents the choice-model estimates under both logistic-regression and random-forest specifications. In each case, stated genre preference emerges as the strongest predictor of video selection, followed by curiosity about science and technology.

Table 8. Choice Model Estimates: Logistic Regression vs. Random Forest

Feature	**Logit Coef.**	**RF Importance**
Age 26–30	−0.0242	0.0051
Age 31–40	−0.0498	0.0050
Age 40+	−0.0292	0.0054
Age Under 18	+0.0249	0.0005
Sex: Male	+0.2928	0.0239
Enjoyment of emotional materials	+0.0529	0.0252
Curiosity about science/technology progress	+0.3266	0.0457
Preference (Sci-Fi=1)	**+5.9469**	**0.8892**

Next, we evaluate estimator performance under the incentivized design. Of the 319 participants, 155 chose Fantasy and 164 chose Sci-Fi. Again true rating refers to the overall mean across all 319 participants. We apply the same six estimators as in the no-incentive case, and report results in Table 9; variances are again estimated via 200 bootstrap replications.

Compared to the no-incentive condition in Table 7, the incentivized design in Table 9 yields uniformly lower variance and reduced bias across all estimators. For example, IPW_Logit's ATE bias falls from +0.709 to +0.253, variance from 0.0493 to 0.0147. And even the naive DM estimator's variance decreases from 0.0173 to 0.0125. **These dramatic, real-world gains mirror our simulation findings—where AIPW paired with a low-switching incentive policy consistently delivered the most accurate and stable ATE estimates—and confirm that aligning incentives to rebalance covariates can robustly stabilize propensity-based estimators in practice.**

Table 10 presents the fitted choice models under logistic regression and random forest. While genre preference remains the dominant predictor, its relative importance falls sharply after incentivization—from 0.88 to 0.25 in the random forest and from 5.95 to 1.53 in the logistic model—confirming that our policy effectively attenuates its influence on treatment assignment.

Table 9. Estimator Results with Bias and ATE Variance (Incentive Data)

Method	Y1 (bias)	Y0 (bias)	ATE (bias)	Variance
True	3.377 (+0.000)	3.627 (+0.000)	-0.250 (+0.000)	—
Naive	3.517 (+0.140)	3.654 (+0.027)	-0.137 (+0.113)	0.0125
PSM_ATT	3.517 (+0.140)	3.531 (-0.096)	-0.014 (+0.236)	0.0285
IPW_Logit	3.479 (+0.102)	3.476 (-0.151)	0.003 (+0.253)	0.0147
IPW_RF	3.150 (-0.227)	3.185 (-0.442)	-0.035 (+0.215)	0.0345
AIPW_Logit	3.452 (+0.075)	3.591 (-0.037)	-0.139 (+0.111)	0.0151
AIPW_RF	3.474 (+0.097)	3.604 (-0.023)	-0.130 (+0.120)	0.0157

Table 10. Choice Model Estimates: Logistic Regression Coefficients & Random Forest Importances

Feature	**Logit Coefficient**	**RF Importance**
Age 26–30	−0.4374	0.0405
Age 31–40	−0.9849	0.0560
Age 40+	−1.2820	0.0701
Sex: Male	+0.9047	0.1558
Enjoyment of imaginative or emotional materials	+0.0411	0.2155
Curiosity about scientific or technological progress	+0.3802	0.2116
Preference (Sci-Fi = 1)	**+1.5336**	**0.2506**