Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
DOI: https://doi.org/10.51903/jtie. v4i3.500

Offline Counterfactual Evaluation for Advertising and Recommendation Slot Policies: A Reproducible Study on the Open Bandit Dataset (Small)

Jinyi Mu*[1], Tong Ye[2], Priya Patel[3]
Email: mjy072180@gmail.com
[1]*Computer Science and Engineering, UCSD, CA, USA*
[2]*Computer Science, Northeastern University, CA, USA*
[3]*Computer Science, Heriot-Watt University, Edinburgh, UK*
*Corresponding Author

Abstract
Offline or counterfactual evaluation is a critical capability for iterating advertising and recommender ranking strategies when online A/B testing is slow, expensive, or risky. Off-policy evaluation (OPE) estimates the expected reward of a candidate policy using logged interaction data from a different behavior policy. Still, it can suffer from high variance under poor overlap and can be misleading when the operational objective is choosing among candidate policies rather than minimizing point-estimation bias alone. This paper presents a fully reproducible empirical study of IPS, self-normalized IPS (SNIPS), doubly robust (DR), and Switch-DR estimators on the Open Bandit Dataset (OBD) small release. Using the Men and Women campaigns (10,000 logged item-impressions per campaign and behavior policy) collected by uniform random and Bernoulli Thompson Sampling (BTS), we construct a held-out oracle for stationary slot-wise policies from the random-traffic split and evaluate both value estimation and policy-ranking consistency on random-logged and BTS-logged test sets. Across 1,000 nonparametric bootstrap replications, IPS and SNIPS are accurate on randomly logged data, whereas BTS-logged data exhibit extreme importance weights and very small effective sample sizes (ESS), making IPS-based ranking unreliable under weak support. Switch-DR is most useful in moderate-overlap regimes, where it truncates high-variance corrections. Still, it introduces bias that depends on the switching threshold and must therefore be stress-tested rather than treated as a universally superior estimator. Finally, we provide a structured reporting template—based on oracle decomposition, overlap diagnostics, and estimator components—for explaining why a policy appears better and how reliable that conclusion is.
Keywords: *Off-Policy Evaluation, Counterfactual Evaluation, Contextual Bandits, Inverse Propensity Scoring.*

I. INTRODUCTION

Modern advertising and recommendation systems are driven by continuous policy iteration: ranking models are retrained, exploration rules are adjusted, and new constraints (e.g., diversity, fairness, inventory, or advertiser constraints) are added. In production, the gold standard for comparing two candidate strategies is an online randomized experiment. However, frequent online experimentation can be prohibitively costly, can delay iteration cycles, and can expose users and revenue to risk when a candidate policy underperforms. These constraints motivate offline counterfactual evaluation, also known as off-policy evaluation (OPE), which estimates the expected reward of an evaluation policy using historical logs collected by a different behavior policy (Horvitz & Thompson, 1952; Dudik, Langford, & Li, 2011).

Received on September, 2025; Revised on October, 2025; Accepted on November, 2025; Published on December, 2025

OPE is appealing because it can be run repeatedly without deploying a new policy, enabling faster and safer iteration. Yet OPE is fragile. Importance sampling estimators, such as inverse propensity scoring (IPS), are unbiased when the propensity scores are correct and the evaluation policy has support where the behavior policy acts, but IPS can have extremely high variance under weak overlap. This is common in real recommender logs where the behavior policy is optimized and assigns very small propensities to many actions (Swaminathan & Joachims, 2015; Wang, Agarwal, & Dudik, 2017). Doubly robust (DR) estimators combine IPS with a learned reward model to reduce variance and remain consistent when either the reward model or the propensity model is correct (Dudik et al., 2011). Nevertheless, DR can still suffer when importance weights are heavy-tailed, and policy selection can be unstable even when point estimates look plausible (Saito et al., 2021).

A persistent obstacle to empirical OPE research is the lack of publicly available, real-world datasets containing logs from multiple behavior policies on the same platform. Without multiple logging policies, it is difficult to quantify estimator error and to study whether an estimator ranks policies correctly. The Open Bandit Dataset (OBD) addresses this gap by releasing large-scale logged bandit feedback collected on a real fashion e-commerce platform under both a uniform random policy and a Bernoulli Thompson Sampling policy (Saito, Aihara, Matsutani, & Narita, 2021). The zr-obp repository also provides a small dataset variant with 10,000 records per campaign-behavior policy pair, enabling fast, reproducible experiments (st-tech/zr-obp, OBD README).

This paper studies how OPE can be used for strategy iteration in advertising and slot-based recommendation interfaces when policies are evaluated offline. Our focus is practical: (1) estimating policy value with confidence intervals (CIs), (2) selecting a better policy by ranking candidates, and (3) producing a structured, auditable report for why a policy appears better. The second objective deserves separate emphasis. In practice, teams act on the policy with the largest estimated value, not on an isolated point estimate, so the operational failure event is selecting the wrong policy rather than merely incurring point-estimation error. An estimator can therefore be nearly unbiased on average yet still be unsafe for decision-making if variance or overlap sensitivity destabilizes the ranking. To make these goals empirically testable, we construct stationary evaluation policies that specify a probability distribution over items for each slot position and use a held-out oracle computed from random-traffic logs to provide ground-truth reference values for these stationary policies.

Our experimental design reflects a real iteration loop in applied teams: collect logs under a behavior policy; define a set of candidate strategies; estimate each strategy's expected

performance; quantify uncertainty; and decide whether any candidate warrants online validation. A key theme in this paper is that this loop is only as reliable as the overlap between candidate policies and the logging policy, and that overlap diagnostics must be integrated into the loop rather than treated as an afterthought.

We do not propose a new OPE estimator. Rather, our contribution is a reproducible empirical protocol for studying estimator behavior in the context of policy selection. First, we construct a held-out oracle for stationary slot-wise policies on OBD small using randomized test traffic, enabling direct measurement of both estimation error and ranking consistency. Second, we quantify ranking stability—not only point-estimation error—through bootstrap top-1 accuracy and rank correlation across six policies with controlled overlap. Third, we connect overlap diagnostics, heavy-tailed weights, ESS, and DR decompositions to specific estimator failures so that ranking breakdowns can be interpreted rather than merely reported. Fourth, we provide a structured reporting template that turns these measurable artifacts into auditable practitioner-facing summaries. The contribution is therefore the evaluation protocol and reporting framework, not a methodological breakthrough in estimator design.

II. LITERATURE REVIEW

Off-policy evaluation in bandit and reinforcement learning settings has a long history rooted in importance sampling and survey sampling (Agisti & Ariani, 2025; Bai et al., 2026; Hidayat et al., 2025; Tolah & Malatji, 2025). The Horvitz-Thompson estimator (Horvitz & Thompson, 1952) and related inverse probability weighting ideas provide unbiased estimation when sampling probabilities are known. In contextual bandits, IPS estimates a target policy's value by reweighting observed rewards using the ratio of the target policy's value to the logging propensity (Dudik et al., 2011). Self-normalization (SNIPS) divides by the sum of importance weights to reduce variance at the cost of small bias (Swaminathan & Joachims, 2015). These estimators are widely used in counterfactual learning from logged bandit feedback, including applications in advertising, search, and recommendation (Bottou et al., 2013; Joachims, Swaminathan, & de Rijke, 2018).

Doubly robust estimation combines a direct method (DM) based on a reward model with an importance-weighted correction term, yielding consistency if either the propensity model or the reward model is correct (Dudik et al., 2011; Bang & Robins, 2005). In contextual bandits, DR can be interpreted as using the reward model as a control variate to decrease the variance of IPS. Subsequent work has developed variance-reduction and robustness techniques, such as weight clipping and shrinkage (Su, Dimakopoulou, Krishnamurthy, & Dudik, 2020), cross-fitting, and estimator-selection heuristics (Saito et al., 2021).

When overlap is poor, importance weights become heavy-tailed, and even DR can have unstable finite-sample behavior. Wang et al. (2017) propose the SWITCH estimator, which applies the importance-weighted correction only when the weight is below a threshold, interpolating between DR and DM to improve finite-sample mean squared error. Kallus and Uehara (2019) further argue for stability and boundedness properties and propose empirical likelihood estimators with favorable efficiency properties, illustrating a broader trend toward variance-controlled OPE in high-variance regimes.

High-confidence OPE has been studied using concentration bounds and lower-confidence bounds on policy value, emphasizing that point estimates alone are insufficient for safe deployment decisions (Thomas et al., 2015; Thomas & Brunskill, 2016). In practice, teams often fall back on bootstrap intervals because they are simple to implement and can account for non-Gaussian variability (Efron & Tibshirani, 1993). However, bootstrap intervals do not solve bias; when estimators are systematically biased (e.g., due to poor overlap or model misspecification), intervals may have poor coverage even when they appear narrow.

In parallel with OPE, contextual bandit learning has been studied extensively in the fields of recommendation and advertising. Randomized traffic enables unbiased offline evaluation and was used to validate contextual bandit algorithms in large-scale deployments (Li et al., 2010; Li et al., 2011). Practical learning algorithms include reduction-based methods and oracle-efficient approaches (Beygelzimer & Langford, 2009; Agarwal et al., 2014). Thompson sampling is a particularly influential baseline in online decision-making systems and has been empirically validated in advertising settings (Chapelle & Li, 2011). These learning works motivate the importance of reliable offline evaluation for rapid iteration and safe experimentation.

Empirical evaluation of OPE estimators requires datasets where the performance of evaluation policies can be estimated reliably. The Open Bandit Dataset was designed to provide this capability by releasing real-world bandit feedback collected during an A/B test between a uniform random policy and a Thompson sampling policy on a large e-commerce platform, with true propensity scores and standardized preprocessing pipelines (Saito et al., 2021). Subsequent work emphasizes robustness protocols for OPE, arguing that practitioners should stress-test estimator choices and hyperparameters rather than rely on a single number (Saito et al., 2021).

Policy selection is a distinct objective from unbiased value estimation. In practice, teams often need to choose the best candidate policy from many options, and ranking stability under estimation noise becomes critical. Recent work highlights the gap between estimator accuracy and safe decision-making, suggesting that conservative bounds, sensitivity analysis, and overlap diagnostics serve as guardrails (Thomas et al., 2015; Saito et al., 2021). Explanations and

auditability are also increasingly important in ad/recommender iteration, where stakeholders want to understand why a strategy is predicted to improve performance.

III. RESEARCH METHOD

This section describes the dataset, experimental protocol, evaluation policies, OPE estimators, confidence interval construction, and the structured reporting template used in the analysis.

A. Dataset

We use the Open Bandit Dataset (OBD) small release, distributed in the zr-obp repository, which contains 10,000 logged item impressions per behavior-policy-campaign pair. We focus on two campaigns (Men and Women) and two behavior policies: uniform random and Bernoulli Thompson Sampling (BTS). Each record corresponds to a displayed item at a specific slot position (three slots per interface), with columns for timestamp, item_id, position, binary click outcome, and the true propensity score of the behavior policy choosing that item at that position. User context is represented by four hashed categorical user_feature fields and a vector of user-item affinity scores. Action (item) context is represented by one numeric and three categorical item features (Saito et al., 2021).

B. Preprocessing and splits

For each (campaign, behavior policy) dataset, we sort logs by timestamp and perform a time-ordered split into 70% training data (7,000 rows) and 30% test data (3,000 rows). The training split from the random-traffic dataset is used to fit the reward model required by DR and Switch-DR, and to estimate CTRs used to define several evaluation policies. The held-out random test split is used to construct an oracle policy value for stationary policies. OPE is evaluated on two test sets per campaign: the random test set (full overlap) and the BTS test set (potentially weak overlap).

C. Oracle construction for stationary slot-wise policies

To evaluate estimation error and ranking accuracy in a fully empirical manner, we restrict evaluation policies to stationary slot-wise distributions pi_e(a|position) that do not depend on user context and do not adapt over time. For such policies, V(pi_e)=E_position[E_{a~pi_e(.|position)}[E[r | a, position]]]. Because the random logging policy samples items uniformly, the empirical CTR for each (item, position) on the held-out random test split is an unbiased estimator of E[r | a, position] for that test period, provided that (i) the logged propensities are correct, (ii) impressions are treated as independent at the impression level, and (iii) the reward distribution is stationary within the held-out split. The CTR table used to construct the Greedy-CTR and Softmax-CTR policies is likewise unbiased with respect to the earlier training window

under random logging. Still, its transfer to the later test oracle depends on temporal stability. We therefore compute an oracle value for each evaluation policy by combining its action probabilities with the held-out CTR table and the empirical distribution of slot positions. This oracle is a common held-out reference for comparing estimators, not noise-free population truth. The time-ordered split reduces leakage from policy construction into evaluation. Still, temporal drift remains possible: if item attractiveness changes over time, the oracle should be interpreted as the policy value for the held-out test window rather than a timeless ground truth.

D. Evaluation policy set

We compare six stationary evaluation policies per campaign (Table 2): (1) Uniform, (2) BTS-marginal, defined as the empirical marginal distribution of items selected by BTS at each position on the BTS training split, (3) Greedy-CTR, a deterministic policy that selects the single item with the highest training CTR at each position, (4) epsilon-greedy-CTR with epsilon=0.1, and two Softmax-CTR policies using temperatures T=0.05 and T=0.2. The latter three policies are constructed from the CTR table estimated on the random training split. This policy set includes both high-overlap policies (Uniform, Softmax-CTR) and low-support policies (Greedy-CTR), enabling controlled overlap stress tests and reflecting common strategy variants used in practice (deterministic, softened, and smoothed policies). At the same time, the family is deliberately small and structured, so the ranking conclusions in this paper should be interpreted relative to this candidate set rather than arbitrary policy classes.

E. Reward model for DR

For each campaign, we fit an L2-regularized logistic regression click model on the random training split with solver SAGA, regularization strength C=0.5, and maximum iterations 8,000. Features include (i) one-hot encodings of the four user_feature fields (24 categories for Men and 23 for Women in the training split), (ii) numeric user-item affinity scores (34 for Men and 46 for Women), (iii) one-hot slot position, and (iv) item context features from item_context.csv. On the random test split, the model AUC equals 0.503 for Men and 0.495 for Women, which is essentially at chance level. This weak discrimination is plausible in the present benchmark because clicks are sparse, the sample size is small, and the policies studied here are stationary slot-wise rules whose average values are driven more by item/position averages than by rich per-impression personalization. We intentionally use this simple, transparent baseline and did not perform an extensive hyperparameter search; accordingly, the DR results should be interpreted as DR with a weak reward model, not as an upper bound on what DR could achieve with stronger nuisance estimation. The weak model helps explain why DR does not consistently outperform IPS/SNIPS with high overlap and why DR can become unstable when overlap is low.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

F. OPE estimators

Let $\mu(a_t \mid x_t)$ denote the logged propensity score, $\pi_e(a_t \mid x_t)$ the probability assigned by the evaluation policy to the logged action, and $w_t = \frac{\pi_e(a_t|x_t)}{\mu(a_t|x_t)}$ the corresponding importance weight. The Inverse Propensity Scoring (IPS) estimator computes the policy value as $\hat{V}_{\mathrm{IPS}} = \frac{1}{T}\sum_{t=1}^{T} w_t\, r_t$ (Horvitz & Thompson, 1952). The Self-Normalized IPS (SNIPS) estimator reduces variance through normalization, yielding $\hat{V}_{\mathrm{SNIPS}} = \frac{\sum_{t=1}^{T} w_t r_t}{\sum_{t=1}^{T} w_t}$ (Swaminathan & Joachims, 2015). The Doubly Robust (DR) estimator combines a direct reward model with an importance-weighted correction, defined as $\hat{V}_{\mathrm{DR}} = \frac{1}{T}\sum_{t=1}^{T}[\hat{q}(x_t, \pi_e) + w_t(r_t - \hat{q}(x_t, a_t))]$, where $\hat{q}(x_t, \pi_e) = \mathbb{E}_{a\sim\pi_e(\cdot|x_t)}[\hat{q}(x_t, a)]$ (Dudík et al., 2011). The Switch-DR estimator applies the correction term only when the importance weight does not exceed a threshold τ, namely $\hat{V}_{\mathrm{SwitchDR}}(\tau) = \frac{1}{T}\sum_{t=1}^{T}[\hat{q}(x_t, \pi_e) + w_t(r_t - \hat{q}(x_t, a_t))\mathbf{1}\{w_t \le \tau\}]$ (Wang et al., 2017). In this study, $\tau = 10$ is used as the primary setting in Tables 4–7, while sensitivity is further examined over $\tau \in \{0,1,2,5,10,20,50,100,200\}$, covering the full spectrum from pure direct modeling to near-DR behavior.

G. Confidence intervals and bootstrap

For each dataset, policy, and estimator, we compute 95% confidence intervals using the percentile method over 1,000 nonparametric bootstrap resamples of the test logs (Efron & Tibshirani, 1993). Bootstrap CIs capture sampling variability of the estimator conditional on the observed test log, but they do not correct systematic bias or account for oracle uncertainty. We therefore report both error relative to the oracle reference and CI coverage of the oracle, using coverage as a diagnostic for whether an estimator's uncertainty quantification is compatible with its bias in this benchmark. Coverage should be read as compatibility with the held-out reference in this setting, not as a guarantee of population coverage. We fix the bootstrap random seed for reproducibility.

H. Implementation and computational efficiency

We implement all estimators in a vectorized manner and precompute per-policy quantities (importance weights, direct-method values, and DR corrections) so that each bootstrap replicate is reduced to indexing and aggregation. On the execution environment used for this paper, computing all estimator results with 1,000 bootstrap replicates for a single dataset (3,000 rows and six policies) takes 10.8 seconds. Running the full benchmark across the four test datasets takes 42.7 seconds end-to-end, after the reward model is fitted and the prediction matrices are prepared. This runtime supports repeated offline iteration in practice.

I. Overlap diagnostics

We quantify overlap using importance-weighted statistics (mean, standard deviation, 99th percentile, and maximum) and the effective sample size ESS = (sum w)^2 / sum w^2. ESS approximates the number of independent, uniformly weighted samples contributing to the estimate and is widely used as a stability heuristic in importance sampling (Swaminathan & Joachims, 2015). In this paper we treat ESS <= 10 as a practical red-flag rather than a theorem-backed cutoff: empirically, policies with ESS in the 1–5 range on Women-BTS are extremely unstable, while Men-BTS Greedy-CTR with ESS ≈ 11 remains borderline and still fails for IPS-based ranking. We therefore use the threshold only as a warning signal that IPS-style estimates may be too fragile for decision-making.

J. Policy ranking metrics

For each dataset and estimator, we rank the six evaluation policies by their estimated values and compare them with the oracle ranking. We report Spearman rank correlation and top-1 selection accuracy over bootstrap resamples. Because only six policies are compared and several oracle values are close, we interpret rank correlation as a coarse stability diagnostic rather than a precise summary of decision quality.

K. Structured reporting template

To support practical decision-making, we summarize measurable artifacts rather than offer free-form speculation. For each campaign, we (i) identify the oracle-best policy and its margin over the runner-up, (ii) compute CTR-weighted contributions of (item,position) pairs to the oracle value, (iii) report overlap diagnostics (weight distribution and ESS) on each logged test set, and (iv) decompose DR-family estimators into the reward-model term and the importance-weight correction. We then convert these facts into a fixed report that answers the questions of what changed, why it matters, and how reliable the estimate is. This is intended as a practical reporting template rather than a new method of explanation. If desired, the same structured evidence could later be verbalized by an LLM under deterministic decoding, but the analysis in this paper is fully template-based and reproducible.

IV. RESULT AND DISCUSSION

This section presents the experimental results of offline policy evaluation using the Open Bandit Dataset (OBD) Small. The descriptive statistics for the test set, including the number of rows, actions, click rate, and propensity distribution for each campaign and behavior policy, are summarized in Table 1. The experimental protocol for offline evaluation and policy ranking is illustrated in Figure 1. The stationary slot-wise evaluation policies compared in the experiments

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

are listed in Table 2. In contrast, the oracle policy values and ranking for the six evaluation policies on the held-out random test split are reported in Table 3. Finally, the observed click rates in each test log dataset are shown in Figure 2.

Table 1. Test-Set Descriptive Statistics for the Open Bandit Dataset (Small) Used in This Study (3,000 Rows per Dataset After Time Split)

Campaign	Behavior Policy (Test)	Rows	Actions	Click Rate	Propensity Min	Propensity Mean	Propensity Max
men	random	3000	34	0.006000	0.0294118	0.0294118	0.0294118
men	bts	3000	34	0.005333	0.000165	0.168582	0.72529
women	random	3000	46	0.006000	0.0217391	0.0217391	0.0217391
women	bts	3000	46	0.007000	1e-06	0.11312	0.701295

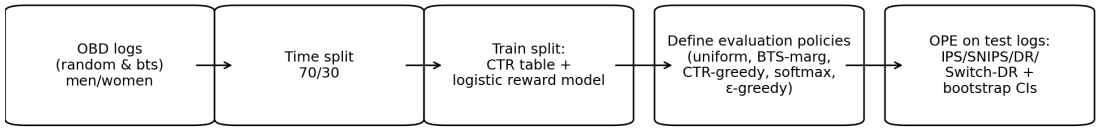

Figure 1. Experimental Protocol for Offline Evaluation and Policy Ranking on OBD Small

Table 2. Stationary Slot-Wise Evaluation Policies Compared in the Experiments

Policy	Definition	Parameters	Support
Uniform	pi(a|pos)=1/K for all items a at each position pos	K=n_actions	Full
BTS-marginal	Empirical item frequency under BTS training logs per position	Smoothing=1e-6	Full (after smoothing)
Greedy-CTR	Deterministic: choose argmax_a CTR_train(a,pos) at each position	None	1 item/position
Epsilon-greedy-CTR	(1-eps) Greedy-CTR + eps Uniform	eps=0.1	Full
Softmax-CTR (T=0.05)	pi(a|pos) ∝ exp(CTR_train(a,pos)/T)	T=0.05	Full
Softmax-CTR (T=0.2)	pi(a|pos) ∝ exp(CTR_train(a,pos)/T)	T=0.2	Full

Table 3. Oracle Policy Values (Held-Out Random Test Split) and Oracle Ranking for the Six Evaluation Policies

Campaign	Rank	policy	oracle_value
men	1	greedy_ctr	0.009759
men	2	eps0.1_ctr	0.009380
men	3	bts_marg	0.006097
men	4	softmax_T0.05	0.006034
men	5	softmax_T0.2	0.005980
men	6	uniform	0.005966
women	1	bts_marg	0.006762
women	2	uniform	0.005718
women	3	softmax_T0.2	0.005636
women	4	softmax_T0.05	0.005357
women	5	eps0.1_ctr	0.000572
women	6	greedy_ctr	0.000000

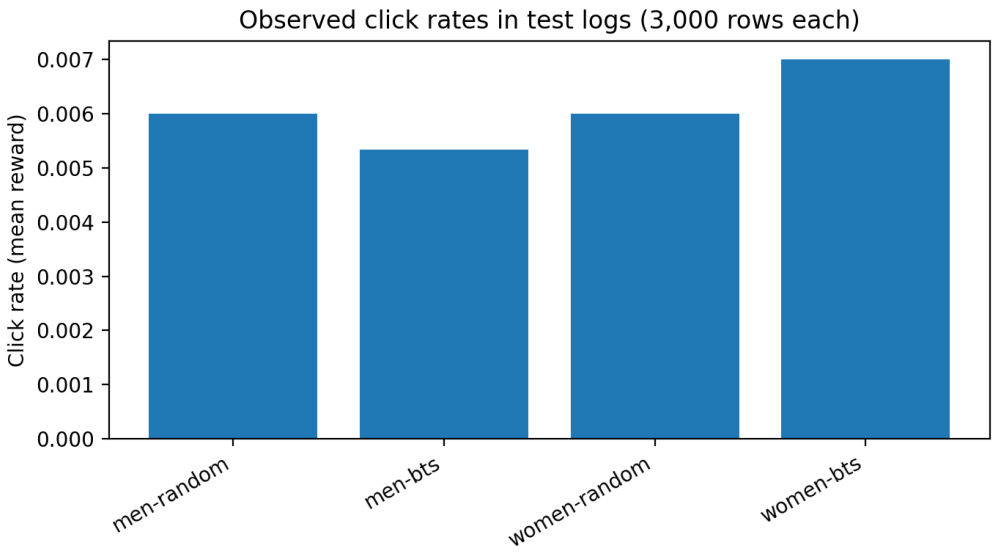

Figure 2. Observed Click Rates in Each Test Log Dataset

Practitioner framework used throughout Section IV. We organize the empirical results around a simple decision workflow: (1) compare candidate policies against the held-out oracle when randomized traffic is available; (2) inspect overlap diagnostics on the logs actually available for OPE (max weight, 99th percentile, and ESS); (3) use IPS/SNIPS when overlap is strong; (4) use SNIPS and Switch-DR as variance-controlled stress tests when overlap is moderate; and (5) if overlap is weak, avoid go/no-go decisions without additional randomized exploration traffic. This framing anticipates the later results and highlights why maintaining a small stream of random exploration traffic is a central operational takeaway rather than an afterthought.

We organize the discussion in three parts: estimation accuracy, CI calibration, and ranking stability. We begin by characterizing the logged data and overlap conditions that drive all three outcomes. Table 1 confirms that each campaign and behavior policy contributes 10,000 records in the OBD small release, and that our time split yields 3,000 test records per dataset. The click rate is low (0.0053–0.0070 in the four test sets), consistent with sparse click outcomes in real recommendation interfaces.

The most important difference between logging policies is the propensity score distribution. Under uniform random logging, propensity scores are constant by design: 1/34 = 0.029412 for the Men campaign and 1/46 = 0.021739 for the Women campaign. In contrast, BTS logging produces highly non-uniform and sometimes extremely small propensities. In the Men-BTS test set, propensity scores range from 0.000165 to 0.72529, with a mean of 0.16858; in the Women-BTS test set, propensity scores range from 1e-6 to 0.701295, with a mean of 0.11312. Figure 3 visualizes these distributions on a log-count scale.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

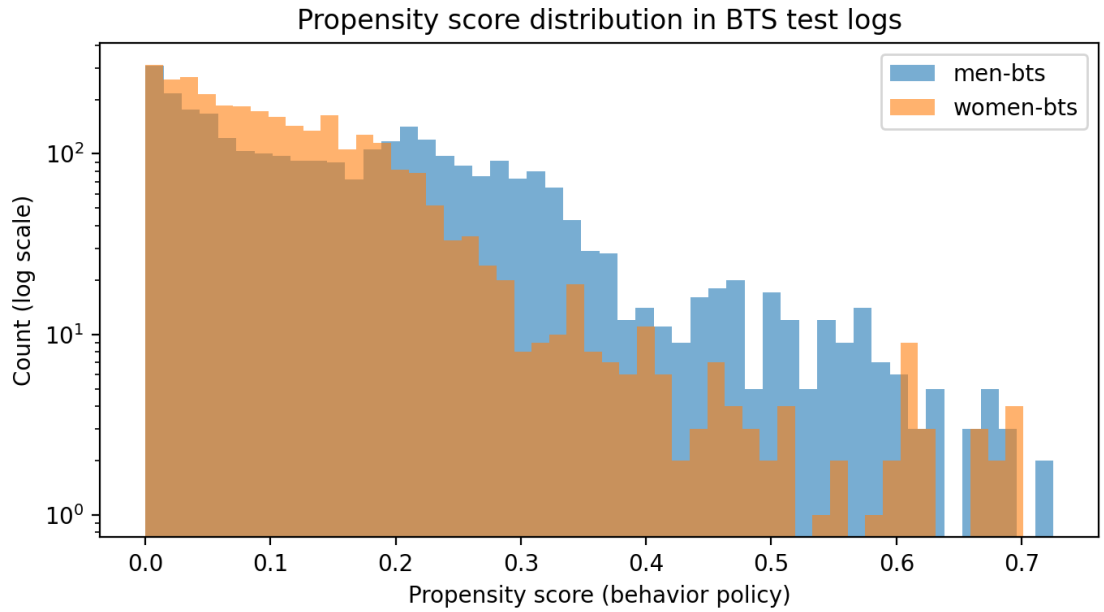

Figure 3. Propensity Score Distribution in BTS-Logged Test Sets (Men vs Women)

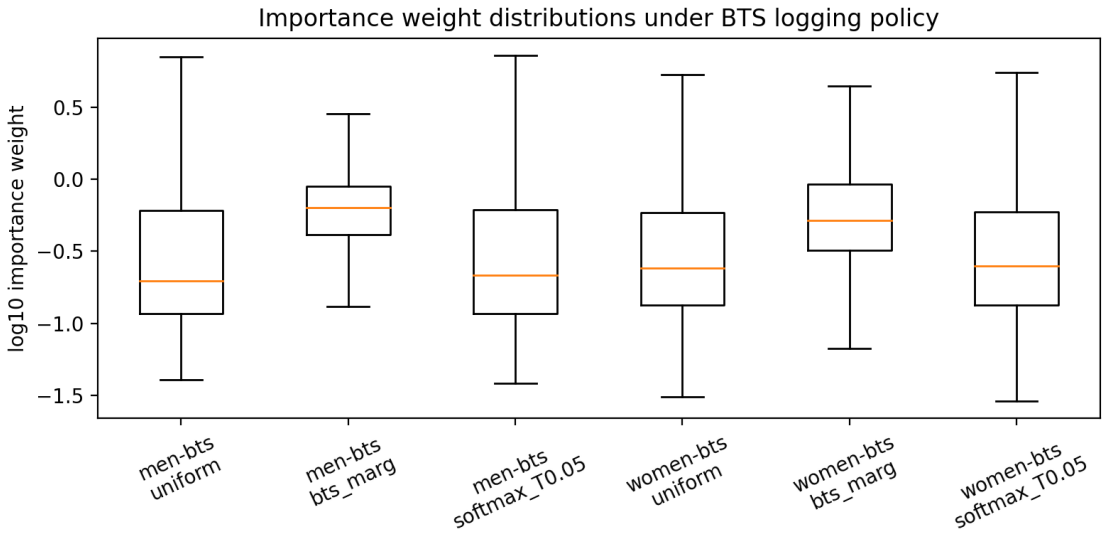

Figure 4. Importance Weight Distributions Under BTS Logging (Log10 Scale)

These propensity distributions imply large importance weights for many evaluation policies. Figure 4 shows log10 weight boxplots for three representative policies. The Women-BTS case is especially pathological: for the Uniform policy, the maximum weight is 21,739 and the mean weight is 8.116, so a single observation accounts for about 21,739/(8.116×3000) ≈ 89% of the total weight mass. If that one observation had reward r=1, its IPS contribution alone would be 21,739/3000 ≈ 7.25, which is more than three orders of magnitude larger than the oracle scale (~0.006). This makes clear how a single sample can dominate IPS under severe support mismatch.

Table 10 shows the same pathology through ESS ≈ 1.25, effectively indicating that only about one uniformly weighted sample is driving the estimate.

Table 4. Men Campaign: OPE Point Estimates and 95% Bootstrap CIs on the Random-Logged Test Set (Tau=10 for Switch-DR)

policy	oracle	IPS	SNIPS	DR	SwitchDR_tau10
uniform	0.005965714079117671	0.006000 [0.003333, 0.009000]	0.006000 [0.003333, 0.009000]	0.006026 [0.003361, 0.009027]	0.006026 [0.003361, 0.009027]
bts_marg	0.006096865315925377	0.006243 [0.001925, 0.011771]	0.006248 [0.001895, 0.011707]	0.006879 [0.002483, 0.012345]	0.006879 [0.002483, 0.012345]
greedy_ctr	0.009759259259259259	0.011333 [0.000000, 0.034000]	0.011236 [0.000000, 0.038471]	0.012267 [-0.000194, 0.036063]	0.006185 [0.005495, 0.007126]
eps0.1_ctr	0.009379904741412451	0.010800 [0.000367, 0.031500]	0.010716 [0.000353, 0.034850]	0.011643 [0.000366, 0.033115]	0.006151 [0.005465, 0.007075]
softmax_T0.05	0.006033824848742337	0.006110 [0.003324, 0.009249]	0.006128 [0.003341, 0.009253]	0.006193 [0.003362, 0.009325]	0.006193 [0.003362, 0.009325]
softmax_T0.2	0.005980412119364177	0.006024 [0.003327, 0.009034]	0.006028 [0.003330, 0.009039]	0.006063 [0.003376, 0.009060]	0.006063 [0.003376, 0.009060]

A. Estimator accuracy on random-logged data

On the random-logged test sets (Tables 4 and 6), all actions have equal propensity and overlap is maximized. In this setting, IPS and SNIPS behave as expected: for full-support policies, their point estimates are close to the oracle values and their bootstrap CIs cover the oracle values for all six evaluation policies. For example, in the Men-Random test set, IPS estimates the BTS-marginal policy at 0.006243 with 95% CI [0.001925, 0.011771], while the oracle value is 0.006097. Similarly, in the Women-Random test set, SNIPS estimates the BTS-marginal policy at 0.007104 with CI [0.002768, 0.012268] versus an oracle value of 0.006762.

B. Finite-sample overlap under deterministic policies

Low-support policies reveal a finite-sample phenomenon even under random logging. Greedy-CTR chooses a single item per position, so the expected match rate between logged impressions and the greedy action at a given position equals 1/K under uniform random logging. In the Women-Random test set (K=46), 46 logged impressions match the Greedy-CTR action at their position and all 46 matched impressions have reward r=0; therefore IPS and SNIPS output 0.000000 with degenerate confidence intervals. This should not be interpreted as evidence that the true policy value is exactly zero. Rather, the oracle itself is noisy for such deterministic policies because it is built from a small number of matched random impressions. In the Men-Random test set (K=34), 89 impressions match Greedy-CTR and 1 click is observed among the

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

matched impressions, leading to a nonzero IPS estimate. DR produces a positive estimate (0.001755) and Switch-DR with tau=10 produces 0.007547 on Women-Random, both reflecting reliance on the reward model rather than logged matches; these estimates overstate the held-out oracle and demonstrate model bias under policy extrapolation. In practice, deterministic policies should be evaluated with caution unless the logged dataset is large enough to yield stable matches or the policy is softened (e.g., by epsilon-greedy exploration).

Table 5. Men Campaign: OPE Point Estimates and 95% Bootstrap CIs on the BTS-Logged Test Set (Tau=10 for Switch-DR)

policy	oracle	IPS	SNIPS	DR	SwitchDR_tau10
uniform	0.005965714079117671	0.001991 [0.000639, 0.003753]	0.002037 [0.000625, 0.003893]	0.002394 [0.000919, 0.004306]	0.003321 [0.001954, 0.005094]
bts_marg	0.006096865315925377	0.006506 [0.001818, 0.014361]	0.006468 [0.001793, 0.014485]	0.006806 [0.002074, 0.014595]	0.003755 [0.002261, 0.005469]
greedy_ctr	0.009759259259259259	0.000000 [0.000000, 0.000000]	0.000000 [0.000000, 0.000000]	0.003487 [0.001890, 0.004829]	0.005448 [0.005321, 0.005578]
eps0.1_ctr	0.009379904741412451	0.000199 [0.000064, 0.000375]	0.000418 [0.000121, 0.001043]	0.003378 [0.001890, 0.004591]	0.005174 [0.004992, 0.005403]
softmax_T0.05	0.006033824848742337	0.001954 [0.000675, 0.003645]	0.001970 [0.000649, 0.003755]	0.002367 [0.000890, 0.004154]	0.003347 [0.002008, 0.005036]
softmax_T0.2	0.005980412119364177	0.001983 [0.000651, 0.003726]	0.002021 [0.000629, 0.003849]	0.002386 [0.000907, 0.004275]	0.003309 [0.001944, 0.005073]

Table 6. Women Campaign: OPE Point Estimates and 95% Bootstrap CIs on the Random-Logged Test Set (Tau=10 for Switch-DR)

policy	oracle	IPS	SNIPS	DR	SwitchDR_tau10
uniform	0.005717822234826301	0.006000 [0.003333, 0.008667]	0.006000 [0.003333, 0.008667]	0.006010 [0.003392, 0.008693]	0.006010 [0.003392, 0.008693]
bts_marg	0.006761934907566886	0.006865 [0.002722, 0.011925]	0.007104 [0.002768, 0.012268]	0.007026 [0.002927, 0.011999]	0.007026 [0.002927, 0.011999]
greedy_ctr	0.0	0.000000 [0.000000, 0.000000]	0.000000 [0.000000, 0.000000]	0.001755 [-0.000347, 0.003641]	0.007547 [0.007370, 0.007722]
eps0.1_ctr	0.000571782234826303	0.000600 [0.000333, 0.000867]	0.000817 [0.000466, 0.001326]	0.002180 [0.000265, 0.003885]	0.007405 [0.007120, 0.007736]
softmax_T0.05	0.005357238337253451	0.005635 [0.003338, 0.008243]	0.005694 [0.003355, 0.008345]	0.005706 [0.003322, 0.008312]	0.005706 [0.003322, 0.008312]
softmax_T0.2	0.005635596586914397	0.005917 [0.003337, 0.008591]	0.005930 [0.003348, 0.008614]	0.005940 [0.003405, 0.008620]	0.005940 [0.003405, 0.008620]

C. Estimator accuracy under BTS logging

Under BTS logging (Tables 5 and 7), the propensity distributions are highly non-uniform and overlap can be weak. Table 10 shows that Women-BTS contains extremely small propensities (down to 1e-6), inducing massive importance weights and driving ESS to values between 1.254 and 1.286 for several full-support policies (Uniform and Softmax variants). These ESS values already foreshadow the ranking instability in Table 9: once only one or a few effectively weighted samples remain, small reward perturbations can reorder policies arbitrarily. In the Men-BTS test set, IPS and SNIPS produce estimates far below the oracle for the Uniform policy (0.00199–0.00204 vs oracle 0.00597) and assign zero value to Greedy-CTR because the few matched samples under BTS logging contain no clicks despite ESS being only borderline acceptable. DR partially corrects for this by using the reward model, but its CIs remain wide for policies with large weights.

Table 7. Women Campaign: OPE Point Estimates and 95% Bootstrap CIs on the BTS-Logged Test Set (Tau=10 for Switch-DR)

policy	oracle	IPS	SNIPS	DR	SwitchDR_tau10
uniform	0.005717822234826301	0.015088 [0.001330, 0.042237]	0.001859 [0.000095, 0.046246]	-0.014409 [-0.087629, 0.041684]	0.003812 [0.002714, 0.005082]
bts_marg	0.006761934930756686	0.012413 [0.003395, 0.029586]	0.005921 [0.001156, 0.027730]	0.007634 [-0.008051, 0.026831]	0.006405 [0.003697, 0.009533]
greedy_ctr	0.0	0.000000 [0.000000, 0.000000]	0.000000 [0.000000, 0.000000]	0.000816 [-0.002144, 0.002941]	0.005223 [0.004712, 0.005700]
eps0.1_ctr	0.000571782234826303	0.001509 [0.000133, 0.004224]	0.000923 [0.000060, 0.004720]	-0.000706 [-0.009058, 0.005349]	0.006061 [0.004413, 0.009139]
softmax_T0.05	0.005357238337253451	0.019373 [0.001351, 0.055054]	0.002605 [0.000110, 0.059319]	-0.007283 [-0.078915, 0.054209]	0.004032 [0.002784, 0.005520]
softmax_T0.2	0.005635596586914397	0.016091 [0.001345, 0.045215]	0.002021 [0.000099, 0.049366]	-0.012767 [-0.085662, 0.044600]	0.003842 [0.002713, 0.005129]

D. DR decomposition under weak overlap

To make model reliance explicit, we decompose DR into its direct-method component and its importance-weight correction. For Women-BTS, the mean absolute correction term equals 0.049299 for the Uniform policy and 0.050814 for Softmax(T=0.05), while the corresponding DM terms are 0.004840 and 0.004943. The correction is strongly negative for several policies (e.g., -0.019249 for Uniform), producing negative DR estimates (Table 7). This behavior is consistent with heavy-tailed weights: a small number of samples can dominate the correction term, and when the reward model is weak, the correction does not stabilize. In other words,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Women-BTS is a regime where both overlap and reward modeling are weak: the DM term is biased because the reward model is uninformative, while the correction term has enormous variance, so DR can perform worse than either component alone. In contrast, for Men-BTS the correction magnitude is smaller (Table 11), and Switch-DR can trade bias for variance more effectively.

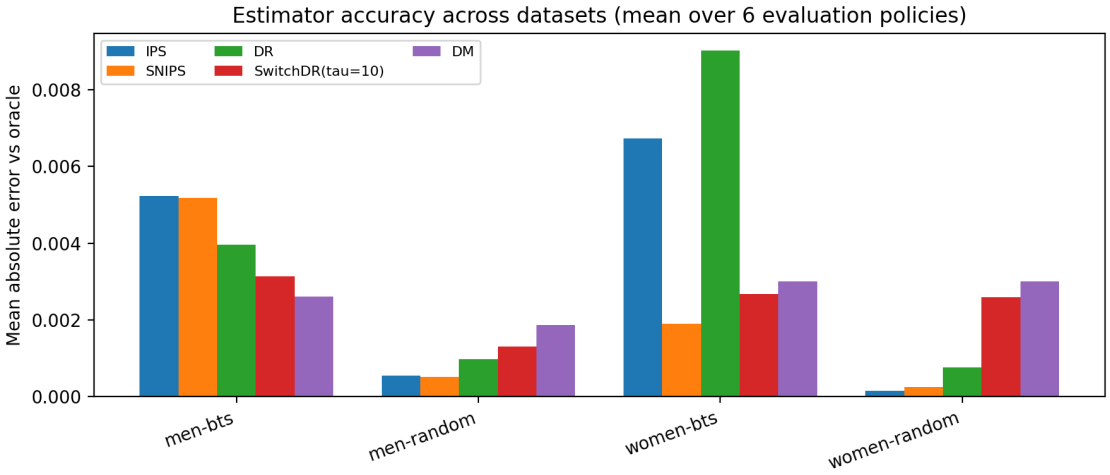

Figure 5. Estimator Accuracy Across Datasets (Mean Absolute Error Averaged Over Six Evaluation Policies)

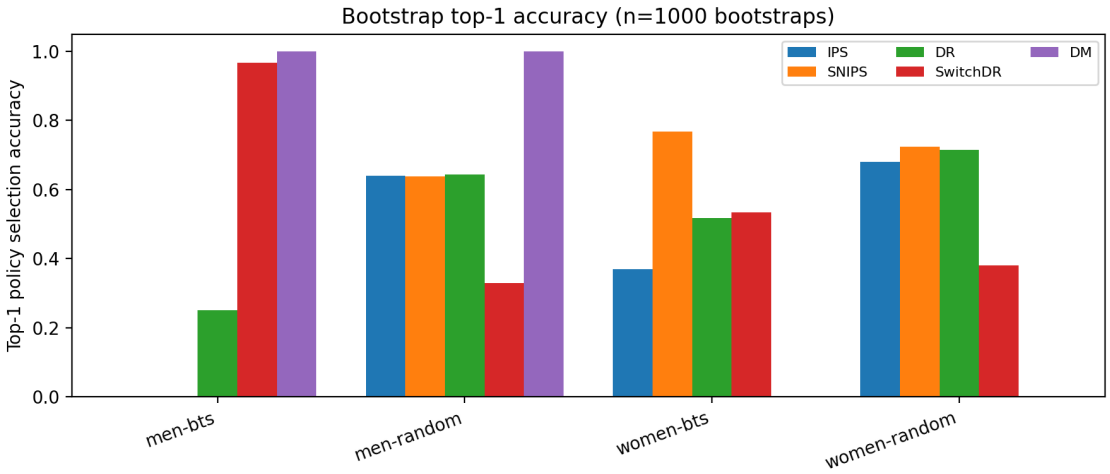

Figure 6. Bootstrap Top-1 Policy Selection Accuracy for Each Estimator and Dataset (1,000 Bootstrap Resamples)

E. Estimator comparison by regime

Table 8 summarizes the main pattern. Under random logging, where overlap is maximal, IPS and SNIPS have the lowest MAE and perfect CI coverage of the oracle. Under Men-BTS, the problem shifts from bias control to variance control: DM has the lowest MAE on average, and Switch-DR improves over DR by truncating unstable corrections. Under Women-BTS, where ESS collapses to 1–5 for several policies, no estimator is uniformly trustworthy; SNIPS attains the lowest

average MAE among IPS-family methods, but the result should be read as the least unstable option in an information-poor regime rather than a universally reliable winner. Overall, the results suggest three regimes: strong overlap favors near-unbiased estimators, moderate overlap can benefit from bias-variance tradeoffs such as Switch-DR, and severe overlap failure leaves too little information for confident offline selection.

F. Confidence intervals and calibration

Table 8 also reports mean CI width and CI coverage of the oracle. On random-logged data, IPS, SNIPS, and DR cover the oracle across all six policies. Under BTS logging, coverage deteriorates sharply. This is especially clear for DM and Switch-DR: both can return narrow intervals that miss the oracle because truncation or model bias shifts the center of the interval. The practical implication is that narrow CIs are only useful when read together with overlap and bias diagnostics; CI coverage of the oracle in this benchmark is therefore a calibration diagnostic, not a stand-alone quality measure.

Table 8. Summary of Estimator Performance: Mean Absolute Error vs Oracle, Mean CI Width, and CI Coverage of the Oracle (Averaged Over Six Evaluation Policies)

Dataset	Estimator	Mean abs. error	Mean CI width	CI coverage rate
men-bts	DM	0.002603	0.000195	0.000
men-bts	DR	0.003969	0.004697	0.167
men-bts	IPS	0.005234	0.003669	0.167
men-bts	SNIPS	0.005174	0.003868	0.167
men-bts	SwitchDR(tau=10)	0.003144	0.002196	0.000
men-random	DM	0.001860	0.001607	0.167
men-random	DR	0.000976	0.016030	1.000
men-random	IPS	0.000549	0.015380	1.000
men-random	SNIPS	0.000523	0.016678	1.000
men-random	SwitchDR(tau=10)	0.001315	0.005069	0.667
women-bts	DM	0.003011	0.000255	0.000
women-bts	DR	0.009023	0.074512	1.000
women-bts	IPS	0.006738	0.028127	1.000
women-bts	SNIPS	0.001903	0.030977	1.000
women-bts	SwitchDR(tau=10)	0.002683	0.003178	0.333
women-random	DM	0.003005	0.000239	0.000
women-random	DR	0.000762	0.005364	1.000
women-random	IPS	0.000162	0.004205	1.000
women-random	SNIPS	0.000250	0.004325	1.000
women-random	SwitchDR(tau=10)	0.002598	0.004258	0.667

G. Policy ranking consistency

Figure 6 and Table 9 show that estimator choice materially changes offline policy selection. The main driver is overlap. On Men-BTS, the oracle-best policy is Greedy-CTR, but under BTS logging this policy has ESS ≈ 11.1 and the few matched samples contain no clicks; IPS and SNIPS therefore assign it zero value, yielding top-1 accuracy 0.000 and negative Spearman correlation. By contrast, on Women-BTS the full-support policies Uniform and Softmax have ESS close to 1,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

which explains why rank correlation becomes unstable even though the policies are not deterministic. Several Spearman intervals span nearly the full [-1, 1] range because only six policies are ranked, several oracle values are very close (especially in Men among Uniform, Softmax(T=0.2), Softmax(T=0.05), and BTS-marginal), and bootstrap resamples of sparse clicks can easily flip their order. In practical terms, such intervals mean that the estimator does not support fine-grained ranking among near-tied candidates. The DM results further show that ranking behavior is policy-set dependent. DM attains perfect top-1 accuracy in Men-Random and Men-BTS not because the reward model is generally strong—the AUC is still ≈0.5—but because this particular six-policy family is largely generated from average CTR structure, and DM preserves that coarse ordering on the Men campaign. The Women campaign shows the opposite pattern, with DM selecting the wrong policy consistently. We therefore interpret the Men DM success as alignment with this restricted policy family, not as evidence that DM is robust for arbitrary candidate sets.

Table 9. Policy-Ranking Consistency: Bootstrap Top-1 Selection Accuracy and Spearman Rank Correlation (Mean and 95% Bootstrap Interval)

Dataset	Estimator	Top-1 acc.	Spearman mean	Spearman CI low	Spearman CI high
men-random	IPS	0.640	0.248	-1.000	1.000
men-random	SNIPS	0.638	0.253	-1.000	1.000
men-random	DR	0.644	0.301	-1.000	1.000
men-random	SwitchDR	0.330	0.159	-0.829	1.000
men-random	DM	1.000	1.000	1.000	1.000
men-bts	IPS	0.000	-0.578	-0.657	-0.429
men-bts	SNIPS	0.000	-0.599	-0.657	-0.429
men-bts	DR	0.251	0.564	-0.657	1.000
men-bts	SwitchDR	0.967	0.783	0.429	1.000
men-bts	DM	1.000	1.000	1.000	1.000
women-random	IPS	0.681	0.896	0.657	1.000
women-random	SNIPS	0.725	0.903	0.657	1.000
women-random	DR	0.715	0.886	0.429	1.000
women-random	SwitchDR	0.380	-0.209	-0.771	1.000
women-random	DM	0.000	-0.657	-0.657	-0.657
women-bts	IPS	0.369	0.567	0.429	1.000
women-bts	SNIPS	0.768	0.698	0.429	0.771
women-bts	DR	0.518	0.072	-0.657	0.771
women-bts	SwitchDR	0.535	-0.222	-0.657	0.314
women-bts	DM	0.000	-0.657	-0.657	-0.657

H. Switch-DR sensitivity to the switching threshold

Switch-DR interpolates between DR and DM via tau: tau=0 yields DM, while tau→∞ yields DR (Wang et al., 2017). Figure 7 reports mean absolute error (MAE) on the BTS-logged test sets over

a grid tau ∈ {0, 1, 2, 5, 10, 20, 50, 100, 200}. This grid was chosen to span the full transition from pure DM to near-DR behavior, with intermediate thresholds around the empirical 99th-percentile weights (roughly 7–15 in Table 10) and larger thresholds that admit most Men-BTS weights while still stress-testing the extreme Women-BTS tail. On Men-BTS, the minimum MAE is 0.002532 at tau=1 (tau=0 yields MAE 0.002603 and tau=10 yields MAE 0.003144). On Women-BTS, the minimum MAE is 0.002433 at tau=20 (tau=0 yields 0.003011 and tau=10 yields 0.002683). In practice, tau cannot be tuned against an oracle. A defensible strategy is to use a validation split or cross-fitting on randomized traffic when available, and otherwise report a sensitivity sweep and choose the smallest tau that stabilizes the estimate, CI width, and ranking over neighboring values. If the conclusions change sharply across nearby tau values, that instability should itself be treated as a warning that the logs do not support confident policy selection.

Table 10. Overlap Diagnostics on BTS-Logged Test Sets: Importance Weight Statistics and Effective Sample Size (ESS) by Evaluation Policy

Campaign	Policy	Mean w	Std w	99th pct w	Max w	ESS
men	uniform	0.977	4.420	11.602	178.253	139.815
men	bts_marg	1.006	3.384	7.261	150.631	243.567
men	greedy_ctr	0.420	6.906	0.000	192.678	11.062
men	eps0.1_ctr	0.476	6.245	1.645	173.977	17.314
men	softmax_T0.05	0.992	5.119	10.994	219.977	108.552
men	softmax_T0.2	0.981	4.583	11.398	188.135	131.443
women	uniform	8.116	396.834	10.533	21739.131	1.254
women	bts_marg	2.097	62.615	6.719	3423.193	3.360
women	greedy_ctr	0.914	11.326	15.052	484.262	19.424
women	eps0.1_ctr	1.634	40.961	13.997	2173.913	4.769
women	softmax_T0.05	7.436	359.200	10.305	19677.516	1.285
women	softmax_T0.2	7.963	388.370	10.866	21275.494	1.261

Table 11. DR Decomposition on BTS-Logged Test Sets: Direct-Method Term and Importance-Weight Correction (Tau Not Applied)

Campaign	Policy	DM term mean	Correction mean	Mean |correction|	DR mean
men	uniform	0.003933	-0.001539	0.005496	0.002394
men	bts_marg	0.004861	0.001944	0.010973	0.006806
men	greedy_ctr	0.005468	-0.001981	0.001981	0.003487
men	eps0.1_ctr	0.005315	-0.001937	0.002333	0.003378
men	softmax_T0.05	0.004058	-0.001691	0.005575	0.002367
men	softmax_T0.2	0.003962	-0.001576	0.005518	0.002386
women	uniform	0.004840	-0.019249	0.049299	-0.014409
women	bts_marg	0.005360	0.002274	0.022420	0.007634
women	greedy_ctr	0.007732	-0.006916	0.006916	0.000816
women	eps0.1_ctr	0.007443	-0.008149	0.011154	-0.000706
women	softmax_T0.05	0.004943	-0.012226	0.050814	-0.007283
women	softmax_T0.2	0.004863	-0.017630	0.049679	-0.012767

I. Implications for policy iteration

The practitioner framework introduced above can now be summarized empirically. When overlap is strong (e.g., ESS in the hundreds as in Men-BTS for BTS-marginal with ESS 243.6),

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 521-543
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

IPS/SNIPS with bootstrap CIs are adequate screening tools. When overlap is moderate, SNIPS and Switch-DR can be useful stress tests, but the sensitivity to tau must be reported. When overlap is weak (e.g., ESS 1.25–1.29 for Uniform and Softmax policies on Women-BTS), the offline logs contain too little support for reliable IPS-style ranking, and the correct action is to collect or preserve randomized exploration traffic rather than over-interpret a point estimate. This is why we emphasize maintaining a small stream of random exploration traffic as a core operational requirement, not a secondary recommendation.

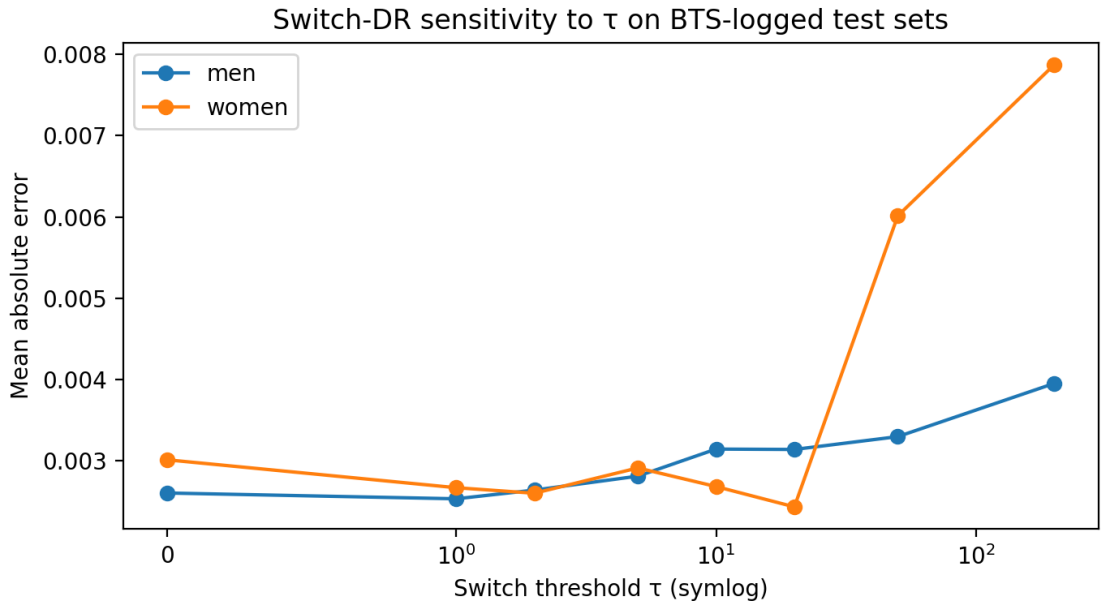

Figure 7. Switch-DR Sensitivity to the Switching Threshold Tau on BTS-Logged Test Sets

J. Structured reporting examples

The oracle ranking provides a ground-truth target for policy iteration, but stakeholders also need to understand why a policy is better and whether the evidence is reliable. We illustrate the reporting template with two cases.

a. Men campaign (oracle-best: Greedy-CTR)

The oracle value of Greedy-CTR is 0.009759, exceeding the runner-up Epsilon-greedy-CTR (0.009380) by 0.000379. Greedy-CTR places all probability mass on a single item per slot: item 11 at position 1, item 33 at position 2, and item 19 at position 3. In the held-out random test split, only the position-2 choice contributes materially: CTR_test(item 33, position 2)=0.027778 and the empirical probability of position 2 is 0.3513, yielding 0.3513×0.027778=0.009759. This explanation is auditable because it is a direct decomposition of the oracle computation. Reliability depends on overlap: on Men-BTS logs, Greedy-CTR has ESS ≈11.1 (Table 10), so IPS assigns it value 0.000000 with a degenerate CI. Switch-DR restores correct top-1 selection (Table 9) by

relying on the reward model when weights exceed tau, but its estimate (0.005448) underestimates the oracle value, quantifying the bias introduced by truncation.

b. Women campaign (oracle-best: BTS-marginal)

The oracle value of BTS-marginal is 0.006762 versus 0.005718 for Uniform, a margin of 0.001044. The top CTR-weighted contributors to BTS-marginal's value are (item 2, position 3) with probability 0.0813 and CTR_test 0.0476 (contribution 0.001284), (item 24, position 1) with probability 0.0800 and CTR_test 0.0417 (contribution 0.001098), and (item 16, position 2) with probability 0.0351 and CTR_test 0.0714 (contribution 0.000850). Reliability is the primary concern under Women-BTS logging: the uniform policy induces maximum weight 21,739 and ESS ≈1.25, so IPS and DR are dominated by a few samples and exhibit large, unstable CIs (Table 7). Under this overlap regime, policy decisions should be based on random-traffic evaluation or on conservative estimates with explicit uncertainty, rather than on BTS-logged IPS.

K. Structured reporting template in practice

The structured facts used above can be represented as a compact table or JSON object: oracle values, CI bounds, overlap diagnostics, and top CTR-weighted contributors per policy. These summaries are sufficient for a reproducible human-written explanation and can optionally be verbalized by an LLM under deterministic decoding. We do not claim this as a new explanation method; the contribution is the reporting template that keeps any narrative grounded in measurable evidence.

V. CONCLUSION AND RECOMMENDATION

This paper presented a reproducible empirical protocol for offline evaluation of advertising and slot-based recommendation strategies using the Open Bandit Dataset (small). By restricting evaluation policies to stationary slot-wise distributions and constructing a held-out oracle from random traffic, we were able to compare estimators on both point-estimation error and policy-selection stability. The experiments show that estimator reliability is governed primarily by overlap: when propensities are uniform and support is broad, IPS and SNIPS are accurate and their confidence intervals align with the oracle reference; when propensities become highly non-uniform, importance weights become heavy-tailed, ESS collapses, and both value estimation and ranking can become unstable. These conclusions are drawn for the six-policy family studied here and should not be assumed to transfer unchanged to arbitrary policy classes.

Recommendations for practitioners

First, maintain a small but sustained fraction of uniform random (or high-entropy) exploration traffic. In our experiments, access to random traffic is what makes the oracle possible and is the

most reliable safeguard against catastrophic overlap failure. Second, treat overlap diagnostics (weight distributions and ESS) as first-class metrics; in this study, ESS <= 10 is best read as a red-flag rather than a universal cutoff, because ESS in the 1–5 range was clearly unusable and ESS around 11 was still borderline. Third, for moderate overlap settings, use SNIPS and Switch-DR as stress tests rather than default winners, and tune Switch-DR thresholds using validation splits, cross-fitting on randomized traffic, and sensitivity analysis rather than oracle access. Fourth, when using DR-family estimators, report the reward-model quality and the DR decomposition to make model reliance explicit. Fifth, interpret bootstrap intervals as estimator-conditional uncertainty; failure of CI coverage of the oracle in this benchmark indicates bias or misspecification, not merely lack of data. Finally, summarize decisions with a structured reporting template—CTR-weighted contribution shifts, overlap diagnostics, and estimator components—so stakeholders can judge both the likely improvement and the reliability of the evidence.

REFERENCES

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., & Schapire, R. E. (2014). Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, *32*(2), 1638–1646. https://proceedings.mlr.press/v32/agarwalb14.html

Agisti, S., & Ariani, D. N. (2025). An Evaluative Study and Implementation Analysis of Disaster Management Information Systems in Local Governments Within the Context of Cities with High Disaster Risk Levels. *JUISI: Jurnal Ilmiah Sistem Informasi*, *4*(1), 100–111. https://doi.org/10.51903/je353h34

Bai, J., Wang, H., Wu, Q., & Zhang, B. (2025). Privacy-Robust Incrementality Estimation in Cookieless Settings via Uplift Modeling: Reproducible Evidence from the Hillstrom E-Mail Experiment. *Journal of Technology Informatics and Engineering*, *5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

Bang, H., & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Beygelzimer, A., & Langford, J. (2009). The Offset Tree for Learning with Partial Labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 129–138. https://doi.org/10.1145/1557019.1557033

Bottou, L., Peters, J., Quinonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., & Snelson, E. (2013). Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research, 14*(1), 3207–3260. http://jmlr.org/papers/v14/bottou13a.html

Chapelle, O., & Li, L. (2011). An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems (NEURIPS,* 2249–2257.

https://papers.nips.cc/paper_files/paper/2011/hash/e53a0a2978c28872a4505bdb51db06dc-Abstract.html

Dudík, M., Langford, J., & Li, L. (2011). Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 1097–1104. https://proceedings.mlr.press/v15/dudik11a.html

Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. *Chapman & Hall/CRC, 57*(1), 1-436. https://doi.org/10.1201/9780429246593

Hidayat, M. S., Muhammad, W., & Isdayanti, P. L. (2025). Digital Marketing Ethics in the Age of AI: A Comparative Analysis of Transparency and Consumer Trust in E-Commerce Platforms. *Journal of Management and Informatics*, *4*(1), 723–740. https://doi.org/10.51903/jmi.v4i1.178

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association, 47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Joachims, T., Swaminathan, A., & de Rijke, M. (2018). Deep Learning With Logged Bandit Feedback. In *International Conference on Learning Representations (ICLR),* 1-12. https://openreview.net/forum?id=SyS2zZ-C-

Kallus, N., & Uehara, M. (2019). Intrinsically Efficient, Stable, and Bounded Off-Policy Evaluation for Reinforcement Learning. In *Advances in Neural Information Processing Systems (NEURIPS), 32*, 1-10. https://papers.nips.cc/paper_files/paper/2019/hash/1ebcecc6ff5d2caa8c36a316ac3a73b7-Abstract.html

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 661–670. https://doi.org/10.1145/1772690.1772758

Li, L., Chu, W., Langford, J., & Wang, X. (2011). Unbiased Offline Evaluation of Contextual-Bandit-Based News Article Recommendation Algorithms. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, 297–306. https://doi.org/10.1145/1935826.1935861

Tolah, A., & Malatji, M. (2025). Evaluating Digital Transformation Within Integration Limitations Using Desk-Based Analytical Case Study. *Journal of Technology Informatics and Engineering*, *4*(2), 289–299. https://doi.org/10.51903/jtie.v4i2.365

Saito, Y., Aihara, S., Matsutani, N., & Narita, Y. (2021). Open Bandit Dataset and Pipeline: Toward Realistic and Reproducible Off-Policy Evaluation. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*. https://datasets-benchmarks-neurips21.github.io/openbandit/

Saito, Y., Udagawa, T., Kiyohara, H., Mogi, K., Narita, Y., & Tateno, K. (2021). Evaluating Off-Policy Evaluation: Sensitivity and Robustness. In *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys),* 181–190. https://doi.org/10.1145/3460231.3474243

Strehl, A. L., Langford, J., Li, L., & Kakade, S. M. (2010). Learning From Logged Implicit Exploration Data. In *Advances in Neural Information Processing Systems, 23*, 1-9. https://papers.nips.cc/paper_files/paper/2010/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Su, Y., Dimakopoulou, M., Krishnamurthy, A., & Dudík, M. (2020). Doubly Robust Off-Policy Evaluation With Shrinkage. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 9167–9176. http://proceedings.mlr.press/v119/su20a.html

Swaminathan, A., & Joachims, T. (2015). Counterfactual Risk Minimization: Learning From Logged Bandit Feedback. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 814–823. http://proceedings.mlr.press/v37/swaminathan15.html

Swaminathan, A., & Joachims, T. (2015). The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems (NEURIPS)*, *28*, 1-9. https://papers.nips.cc/paper_files/paper/2015/hash/3b2d8f0b0c580c873d1a0b9c9f4ffb42-Abstract.html

Thomas, P. S., & Brunskill, E. (2016). Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2139–2148. http://proceedings.mlr.press/v48/thomas16.html

Thomas, P. S., Theocharous, G., & Ghavamzadeh, M. (2015). High-Confidence Off-Policy Evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 3000–3006. https://ojs.aaai.org/index.php/AAAI/article/view/9746

Wang, Y.-X., Agarwal, A., & Dudík, M. (2017). Optimal and Adaptive Off-Policy Evaluation in Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 3589–3597. http://proceedings.mlr.press/v70/wang17f.html