Clarifying Uncertainty Quantification in Off-Policy Evaluation: Beyond Effective Sample Sizes, Towards Confidence Intervals

Aditya Dutta [1] **Kaixuan Liu** [1] **Shengpu Tang** [1]

Abstract

Off-policy evaluation (OPE) is often used to assess or compare whether a policy learned from previously collected behavior data is reliable enough to deploy, but good decisions require uncertainty diagnostics that reflect the actual error of the estimator being used. A common practice is to report the normalized-weight effective sample-size proxy $\widehat{\mathrm{ESS}} = 1/\sum_i \bar{w}_i^2$, and to treat it as evidence about the reliability of an OPE estimate. We argue that this interpretation is too broad: $\widehat{\mathrm{ESS}}$ is best understood as a practical approximation motivated by self-normalized importance sampling (SNIS), not as a universal uncertainty measure. First, even within importance sampling, it is incomplete because fixed normalized weights can correspond to different estimator variances when the reward variance changes; second, across direct, hybrid, and fitted evaluators such as DM, MRDR, and FQE, it is not estimator-agnostic at all. This leaves cross-estimator uncertainty comparison as an open problem. We therefore argue that interval-based summaries provide a more promising common language for comparison: confidence-interval width can be reported in practice, while empirical coverage provides a simulation-based calibration metric for evaluating whether intervals are valid.

1. Introduction

Offline reinforcement learning (RL) relies on OPE to assess a candidate policy using previously collected data. In high-stakes and large-scale settings such as healthcare, recommendation systems, education, and public-policy decision making, a point estimate alone is insufficient to select or reject a policy based on limited observational data (Gottesman et al., 2018; 2019; Li et al., 2010; Doroudi et al., 2019; Tang and Wiens, 2021). A practitioner also wants to know how uncertain the estimate is, whether one estimator is more reliable than another, and whether the same uncertainty summary can be compared meaningfully across estimator families. Two uncertainty summaries are especially common in this discussion. The first is effective sample size (ESS), typically reported for importance-sampling (IS)-based estimators. The second is a confidence interval (CI), obtained through asymptotic, concentration-based, or bootstrap procedures. These summaries are both widely used, but they come from different inferential logics: ESS originates inside weighted Monte Carlo estimation, whereas a CI is an inferential object that can in principle be defined for any estimator (Elvira et al., 2022; Hao et al., 2022).

That distinction matters because modern OPE consists of many estimator families. IS methods rely on empirical likelihood ratios between an evaluation policy and a behavior policy (Precup et al., 2000; Jiang and Li, 2016). Direct methods rely on learned value or reward models (Voloshin et al., 2021). Hybrid estimators such as DR, WDR, and MRDR combine model-based predictions with weighted correction terms (Jiang and Li, 2016; Farajtabar et al., 2018). Fitted evaluators such as FQE estimate policy value through iterative fitted value-function regression (Le et al., 2019; Hao et al., 2022; Voloshin et al., 2021). Since these families are driven by different stochastic objects, there is no obvious reason to expect a single quantity defined with IS weights to serve as a universal uncertainty summary across all of them.

This paper makes that point precise. Our first claim is internal to importance sampling: the common normalized-weight ESS proxy is a partial diagnostic of weight degeneracy rather than a complete measure of estimator uncertainty. Our second claim is cross-OPE family: because modern OPE estimators are driven by different sources of randomness and approximation error, a weight-based ESS is not an estimator-agnostic solution to cross-estimator uncertainty comparison. This does not solve the broader comparison problem; rather, it clarifies why that problem remains open. We therefore study confidence-interval width as a value-scale uncertainty summary, and empirical coverage as a simulation-based calibration metric for evaluating interval

[1]Emory University, Atlanta, Georgia, USA. Correspondence to: Aditya Dutta <aditya.dutta@emory.edu>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning. Seoul, South Korea.

procedures across estimator families. We do not propose a new method for constructing confidence intervals; instead, we study how interval-based summaries can be used as diagnostics for comparing uncertainty across OPE estimators.

2. Problem Formulation

We consider a finite-horizon discounted Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, R, d_1, \gamma, T)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P(\cdot \mid s, a)$ is the transition kernel, $R(\cdot \mid s, a)$ is the reward distribution, d_1 is the initial-state distribution, $\gamma \in (0, 1]$ is the discount factor, and T is the horizon. Let π_b denote the behavior policy and π_e the evaluation policy, both stochastic $\mathcal{S} \to \Delta(\mathcal{A})$.

Off-policy evaluation. We observe an offline dataset $D = \{\tau_i\}_{i=1}^N$ of N trajectories generated independently under π_b. Each trajectory has the form

$$\tau_i = (s_{i,0}, a_{i,0}, r_{i,0}, \ldots, s_{i,T-1}, a_{i,T-1}, r_{i,T-1}).$$

Let $\mathbb{P}^\pi$ denote the trajectory law induced by d_1, P, and policy π. Since the trajectories in D are independent and generated under π_b, the dataset law is $\mathcal{P}_{\pi_b}^N := \left(\mathbb{P}^{\pi_b}\right)^{\otimes N}$. The target quantity is the value of the evaluation policy,

$$V(\pi_e) = \mathop{\mathbb{E}}_{\tau \sim \mathbb{P}^{\pi_e}} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right].$$

We study four broad OPE families. The first consists of importance-sampling estimators, including ordinary IS, weighted IS (i.e., SNIS), and PDIS (Precup et al., 2000; Jiang and Li, 2016). The second consists of direct estimators, including DM and related model-based value estimators (Voloshin et al., 2021). The third consists of hybrid estimators, such as DR, WDR, and MRDR, that combine model-based predictions with weighted correction terms (Jiang and Li, 2016; Farajtabar et al., 2018). The fourth consists of fitted evaluators, most notably FQE, together with related fitted value-function methods (Le et al., 2019; Hao et al., 2022; Tang and Wiens, 2021).

For trajectory-wise IS, let

$$w_i = \prod_{t=0}^{T-1} \frac{\pi_e(a_{i,t} \mid s_{i,t})}{\pi_b(a_{i,t} \mid s_{i,t})}$$

denote the trajectory importance ratio for sample i, let $\bar{w}_i = w_i / \sum_{j=1}^N w_j$ denote its normalized version, and let $G_i = \sum_{t=0}^{T-1} \gamma^t r_{i,t}$ denote its discounted return. Then

$$\widehat{V}_{\text{IS}}(D) = \frac{1}{N} \sum_{i=1}^N w_i G_i, \qquad \widehat{V}_{\text{SNIS}}(D) = \sum_{i=1}^N \bar{w}_i G_i.$$

Per-decision IS instead applies the cumulative importance ratio only up to the time at which each reward is observed. Let

$$w_{i,0:t} = \prod_{\ell=0}^{t} \frac{\pi_e(a_{i,\ell} \mid s_{i,\ell})}{\pi_b(a_{i,\ell} \mid s_{i,\ell})}.$$

The per-decision IS estimator is

$$\widehat{V}_{\text{PDIS}}(D) = \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^{T-1} \gamma^t w_{i,0:t} r_{i,t}.$$

Thus, trajectory-wise IS multiplies the entire return by the full trajectory ratio w_i, whereas PDIS weights each reward only by the importance ratios accumulated up to that time step. This can reduce variance in sequential settings because early rewards are not multiplied by later importance ratios.

Evaluating OPE point estimate. For a generic OPE estimator $\widehat{V}(D)$, we write the dataset-level error as

$$\text{Err}(D) = \widehat{V}(D) - V(\pi_e).$$

We also consider absolute error $\text{AE}(D) = \left|\widehat{V}(D) - V(\pi_e)\right|$ and squared error $\text{SqErr}(D) = \left(\widehat{V}(D) - V(\pi_e)\right)^2$. The corresponding mean-squared error is

$$\text{MSE} = \mathop{\mathbb{E}}_{D \sim \mathcal{P}_{\pi_b}^N} \left[\left(\widehat{V}(D) - V(\pi_e)\right)^2 \right].$$

Uncertainty quantification in OPE. In Sections 3 and 5, we explore two popular ways of communicating OPE uncertainty, namely effective sample size (ESS) and confidence intervals (CI), which can serve as diagnostics for the quality of OPE. Before explaining the details, we first define what constitutes a "good" uncertainty diagnostic.

Definition 1 (Uncertainty diagnostic)**.** An uncertainty diagnostic is any statistic or inferential object reported alongside an OPE point estimate in order to summarize the reliability, dispersion, or calibration of that estimate.

Definition 2 (Useful diagnostic)**.** A useful uncertainty diagnostic should be:

1. well-defined for the estimator under consideration,
2. informative about the uncertainty quantity of interest, and
3. comparable across estimators when cross-estimator comparison is the goal.

Definition 2 separates two questions that are often conflated. A diagnostic may be meaningful within one estimator family but fail as a cross-estimator comparison tool. For example, a normalized-weight diagnostic is well-defined for importance-sampling estimators because those estimators explicitly use importance ratios, but the same object is not naturally defined for DM or FQE. Conversely, an interval-based diagnostic can be reported for any estimator equipped with a valid interval-construction procedure, but its interpretation still depends on the validity of that procedure.

3. Effective Sample Size: From Monte Carlo to Self-Normalized Importance Sampling

3.1. Classical motivation

Effective sample size (ESS) originates as a variance-ratio concept inside importance sampling, especially in the self-normalized setting. Informally, it asks how many direct Monte Carlo samples from the target distribution would be needed to match the variance of an importance-sampling estimate. Early derivations trace back to Kong and coauthors; see Elvira et al. (2022) for a detailed reconstruction and critique of the standard practical approximation.

Consider the expectation $I = \mathbb{E}_{X\sim\pi}[h(X)]$, where π is the target distribution and h is the integrand. If samples can be drawn directly from π, the Monte Carlo (MC) estimator is $\widehat{I}_{\text{MC}} = N^{-1}\sum_{i=1}^{N} h(X_i)$, with i.i.d. $X_i \sim \pi$. If samples are instead drawn from a proposal distribution q, the self-normalized importance-sampling estimator is $\widetilde{I}_{\text{SNIS}} = \sum_{i=1}^{N} \bar{w}_i h(X_i)$, where $X_i \sim q$, $w_i = \pi(X_i)/q(X_i)$, and $\bar{w}_i = w_i/\sum_{j=1}^{N} w_j$. The conceptual definition of effective sample size is the variance ratio $\text{ESS} = N \cdot \text{Var}(\widehat{I}_{\text{MC}})/\text{Var}(\widetilde{I}_{\text{SNIS}})$. This definition is useful conceptually, but it is computable in practice.

3.2. The standard normalized-weight approximation

The exact ESS is intractable because it requires the variances of both the MC estimator and the SNIS estimator. The quantity most commonly reported is instead $\widehat{\text{ESS}} = 1/(\sum_{i=1}^{N} \bar{w}_i^2) \approx \text{ESS}$. As reviewed by Elvira et al. (2022); Kong (1992), this expression is obtained through a sequence of assumptions and approximations applied to the conceptual definition, which we summarize below.

Start from $\text{ESS} = N \cdot \text{Var}_\pi(\widehat{I}_{\text{MC}})/\text{Var}_q(\widetilde{I}_{\text{SNIS}})$. Since $\widetilde{I}_{\text{SNIS}}$ is biased at finite N, an exact error comparison would involve mean-squared error rather than only variance. The derivation nevertheless treats $\text{MSE}_q(\widetilde{I}_{\text{SNIS}}) \approx \text{Var}_q(\widetilde{I}_{\text{SNIS}})$, thereby ignoring finite-sample bias.

Let $W = \pi(X)/q(X)$ and $H = h(X)$. If π is a normalized probability density, then $\mathbb{E}_q[W] = \int\{\pi(x)/q(x)\}q(x)\,dx = 1$. The SNIS estimator is the ratio estimator $\widetilde{I}_{\text{SNIS}} = (N^{-1}\sum_{i=1}^{N} W_i H_i)/(N^{-1}\sum_{i=1}^{N} W_i)$. This identity does not require the weights to be normalized sample-by-sample; the denominator is random, while $\mathbb{E}_q[W] = 1$ is a population-level normalization used in the approximation.

A delta-method expansion of this ratio gives an approximation to $\text{Var}_q(\widetilde{I}_{\text{SNIS}})$. At this point, the variance still depends on both the weights W and the integrand $H = h(X)$. The common derivation then applies additional moment approximations that reduce this dependence to $\text{Var}_q(\widetilde{I}_{\text{SNIS}}) \approx \text{Var}_\pi(\widehat{I}_{\text{MC}})(1 + \text{Var}_q(W))$. This is the key simplification: the effect of h is absorbed into the direct Monte Carlo variance term, leaving the remaining inflation factor to depend only on the variability of the weights. This loss of explicit dependence on h is one reason the final proxy should be interpreted cautiously; when the variability of $h(X)$ differs across regions emphasized differently by q and π, weight concentration alone cannot describe the estimator variance.

Substituting the approximation into the variance-ratio definition gives $\text{ESS} \approx N/(1+\text{Var}_q(W))$. Since $\mathbb{E}_q[W] = 1$, we have $1 + \text{Var}_q(W) = \mathbb{E}_q[W^2]$, so $\text{ESS} \approx N/\mathbb{E}_q[W^2]$. Replacing population moments by empirical plug-in estimates using unnormalized weights w_i gives

$$\widehat{\text{ESS}} = N\frac{\left(N^{-1}\sum_{i=1}^{N} w_i\right)^2}{N^{-1}\sum_{i=1}^{N} w_i^2} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2} = \frac{1}{\sum_{i=1}^{N} \bar{w}_i^2},$$

where $\bar{w}_i = w_i/\sum_{j=1}^{N} w_j$.

Thus, the practical quantity $\widehat{\text{ESS}}$ should be interpreted as a normalized-weight proxy motivated by SNIS. It is useful for summarizing weight concentration, but it is not an exact variance ratio and should not be treated as a universal uncertainty measure.

3.3. What the approximation measures

Within importance sampling, $\widehat{\text{ESS}}$ measures weight concentration. If all normalized importance weights are equal (to $1/N$), then $\widehat{\text{ESS}} = 1/(N/N^2) = N$. If a small number of normalized importance weights dominate, then $\sum_i \bar{w}_i^2$ is large and $\widehat{\text{ESS}}$ is small. In that sense, $\widehat{\text{ESS}}$ is a useful summary of weight degeneracy (Elvira et al., 2022).

At the same time, once the final approximation is written entirely in terms of normalized empirical weights, it can only reflect this single aspect of uncertainty due to importance sampling reweighting. It does not, by itself, encode every source of variation that affects the estimate.

Remark 1. Throughout the remainder of the paper, $\widehat{\text{ESS}}$ refers specifically to the normalized-weight proxy

$$\widehat{\text{ESS}} = \frac{1}{\sum_{i=1}^{N} \bar{w}_i^2},$$

with $\bar{w}_i = w_i/\sum_{j=1}^{N} w_j$. This is the practical approximation associated with the self-normalized importance-sampling derivation above. We do not claim that it is the unique or canonical ESS notion for every importance-sampling estimator (as was explored in Elvira et al. (2022)).

4. Inherent Limitations of Weight-Based ESS Even Within Importance Sampling

Our critique begins inside the IS OPE family itself. Even there, a diagnostic based only on realized importance ratios is necessarily incomplete. Although the practical proxy $\widehat{\mathrm{ESS}}$ is motivated by the self-normalized setting, the underlying criticism is broader: any uncertainty summary that depends only on realized importance ratios can at most summarize ratio concentration, not the full uncertainty of the estimator.

The most direct OPE example is reward variance. Even if the IS ratios are the same, estimator uncertainty can still change as rewards become more or less noisy.

To illustrate this point, consider a contextual bandit with an IS estimate $\widehat{V}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} \rho_i r_i$, where $\rho_i = \pi_e(a_i|s_i)/\pi_b(a_i|s_i)$. A tempting interpretation would be that, if two OPE problems have the same realized importance-ratio structure, then a weight-based ESS proxy should represent the same uncertainty. The following proposition shows that this interpretation is false.

Proposition 1. *There exist two contextual-bandit OPE problems with identical realized importance ratios, and hence identical normalized-weight $\widehat{\mathrm{ESS}}$, but different IS variances.*

Proof. Consider a contextual bandit with i.i.d. samples $(s_i, a_i, r_i)_{i=1}^n$, where $s_i|d,\ a_i \sim \pi_b(\cdot|s_i)$, and $r_i \sim R(\cdot|s_i, a_i)$. Let $\rho(s,a) = \pi_e(a|s)/\pi_b(a|s)$, $\mu(s,a) = \mathbb{E}_{r\sim R(\cdot|s,a)}[r]$, and $\sigma_R^2(s,a) = \mathrm{Var}_{r\sim R(\cdot|s,a)}(r)$. For the ordinary importance-sampling estimator $\widehat{V}_{\mathrm{IS}} = n^{-1}\sum_{i=1}^{n} \rho(s_i,a_i) r_i$, independence gives $\mathrm{Var}(\widehat{V}_{\mathrm{IS}}) = n^{-1}\mathrm{Var}(\rho(s,a)r)$, where the variance is over $s \sim d$, $a \sim \pi_b(\cdot|s)$, and $r \sim R(\cdot|s,a)$. Applying the law of total variance over the state, action, and reward randomness:

$$\mathrm{Var}(\widehat{V}_{\mathrm{IS}}) = \frac{1}{n}\Bigg\{ \mathrm{Var}_{s\sim d}\left(\sum_a \pi_e(a \mid s)\mu(s,a)\right) + \mathbb{E}_{s\sim d}\left[\mathrm{Var}_{a\sim\pi_b(\cdot|s)}(\rho(s,a)\mu(s,a))\right] + \mathbb{E}_{s\sim d}\left[\mathbb{E}_{a\sim\pi_b(\cdot|s)}\left[\rho(s,a)^2\sigma_R^2(s,a)\right]\right] \Bigg\}.$$

The final term shows that the IS variance depends explicitly on the conditional reward variance $\sigma_R^2(s,a)$. Now construct two worlds $\mathcal{W}_1$ and $\mathcal{W}_2$ with the same state distribution d, behavior policy π_b, evaluation policy π_e, and reward mean function $\mu(s,a)$, but with different conditional reward variances on at least one state-action pair with positive probability. The first two terms in the displayed expression are identical across the two worlds, but the final reward-variance term differs, so the two worlds have different IS estimator variances. However, because d, π_b, and π_e are the same in both worlds, a coupled dataset with the same sampled state-action pairs (s_i, a_i) has identical realized ratios $\rho(s_i, a_i)$. Therefore any diagnostic depending only on the realized ratios, including normalized-weight $\widehat{\mathrm{ESS}}$, is identical across the two worlds. □

Proposition 1 makes the basic limitation explicit. A normalized-weight ESS proxy only summarizes concentration of realized importance ratios. It does not account for reward variance and therefore cannot be a complete uncertainty measure even for estimators in the importance-sampling family.

Remark 2. Our point is not that $\widehat{\mathrm{ESS}}$ is useless within importance sampling; it can still be informative about one specific failure mode, namely weight degeneracy. The point is narrower and more important: $\widehat{\mathrm{ESS}}$ is incomplete as a summary of estimator uncertainty, even where it is most naturally motivated.

5. Non-Portability of Weight-Based ESS Across OPE Families

Section 4 showed that normalized-weight ESS is already incomplete inside the importance-sampling family. The broader cross-family question is harder. Modern OPE includes direct, hybrid, and fitted estimators whose dominant uncertainty sources differ substantially, so cross-estimator uncertainty comparison remains open. The point of this section is therefore not to claim a complete solution, but to show why a weight-based ESS is not estimator-agnostic and why interval-based summaries are a more promising candidate common language.

The key distinction is between estimators whose randomness is fundamentally organized around empirical importance ratios and estimators whose randomness comes primarily from model fitting or approximation error. For ordinary IS, SNIS, and PDIS, the estimator is explicitly built from ratios between π_e and π_b. It is therefore natural that one might inspect a diagnostic constructed from those same ratios. Even then, as Section 4 showed, a ratio-only diagnostic is incomplete. But at least it is attached to a central stochastic object in the estimator.

That is no longer true once one moves beyond IS estimators. The direct method (DM) estimates policy value by fitting a reward model or value model and then evaluating that model under the target policy. Its uncertainty is therefore driven by regression error, misspecification, and extrapolation rather than by a native vector of realized importance ratios (Voloshin et al., 2021). Hybrid estimators such as DR, WDR, and MRDR combine a model-based term with a weighted correction, so their uncertainty depends jointly on model-estimation error and the correction term (Jiang and Li, 2016; Farajtabar et al., 2018). Fitted Q-evaluation (FQE) is different again: it is produced by iterative fitted

value-function regression, so its uncertainty is governed by fitted Bellman error, approximation error, and finite-sample error propagation through the learned value function (Le et al., 2019; Hao et al., 2022). These are not minor variations of the same estimator; they are different procedures with different dominant sources of uncertainty. We make the cross-family requirement in Definition 2 precise with the following notion of estimator-agnosticity.

Definition 3 (Estimator-agnostic diagnostic)**.** For a class of estimators $\mathcal{E}$, a diagnostic D_e is estimator-agnostic over $\mathcal{E}$ if it is defined from the stochastic objects that determine each estimator $e \in \mathcal{E}$, and if the resulting quantity has the same target interpretation across all $e \in \mathcal{E}$.

Proposition 2. *A diagnostic that depends only on empirical importance ratios cannot be estimator-agnostic across IS, DM, MRDR, and FQE.*

Proof. Let a ratio-only diagnostic have the form

$$D_\rho(D) = \phi(\rho_1, \ldots, \rho_N),$$

where ρ_i denotes an empirical importance ratio, or a trajectory-wise product of per-step ratios, and ϕ is any measurable function of these ratios alone. For IS-type estimators, the same ratios appear in the estimator itself, e.g.

$$\widehat{V}_{\mathrm{IS}}(D) = \frac{1}{N}\sum_{i=1}^{N} \rho_i G_i, \qquad \widehat{V}_{\mathrm{SNIS}}(D) = \sum_{i=1}^{N} \bar{\rho}_i G_i.$$

Thus D_ρ is defined from a native stochastic object used directly in the estimator.

For DM, the estimator has the generic form

$$\widehat{V}_{\mathrm{DM}}(D) = \mathop{\mathbb{E}}_{s\sim d_1}\left[\sum_a \pi_e(a \mid s)\widehat{Q}(s,a)\right],$$

or an analogous plug-in value based on a fitted reward or transition model. Its defining stochastic object is the fitted model $\widehat{Q}$, $\widehat{r}$, or $\widehat{P}$ (and crucially depends on the chosen hypothesis classes when fitting these models), not the vector $(\rho_1, \ldots, \rho_N)$. Therefore two datasets can induce the same empirical importance-ratio vector but different fitted models, and hence different DM uncertainty.

For FQE, the estimator is determined by the fitted sequence of value-function regressions

$$\widehat{Q}_H, \widehat{Q}_{H-1}, \ldots, \widehat{Q}_1,$$

so its error depends on regression error, approximation error, and propagation through the fitted Bellman recursion. Again, $(\rho_1, \ldots, \rho_N)$ is not a sufficient stochastic object for the estimator. For MRDR and related hybrid estimators, ratios enter only through correction terms, while the estimator also depends on fitted nuisance functions. Hence D_ρ can describe a native object for IS-type estimators and at most one component of hybrid estimators, but it is not defined from the determining stochastic objects of DM or FQE. By Definition 3, it is therefore not estimator-agnostic across these families. □

Proposition 2 identifies a portability failure rather than a correctness failure. The problem is not that normalized-weight ESS is inherently wrong. The problem is that it is derived from one estimator family and often interpreted as though it were a universal cross-family uncertainty summary.

Our discussion is also related to estimator selection for OPE, where the goal is to choose among estimators using logged data, as in policy-adaptive estimator selection (Udagawa et al., 2023). We view this as complementary: our goal is not to select as estimator, but to clarify which uncertainty diagnostics are portable when multiple estimators are compared.

5.1. Interval-based summaries as a candidate common language

If a weight-based ESS is not estimator-agnostic, what kind of uncertainty summary can be compared across OPE families? The relevant requirement is not merely that the quantity can be computed for every dataset. A diagnostic can be computable while still failing to summarize the uncertainty of the estimator being reported. For cross-family comparison, the diagnostic should have a common target interpretation across estimators: it should describe uncertainty in $\widehat{V}(D)$, rather than only the behavior-evaluation policy mismatch.

Confidence intervals are more natural from this perspective. For any estimator $\widehat{V}(D)$, one may attempt to construct an interval $C(D) = [L(D), U(D)]$ intended to contain $V(\pi_e)$ with a prescribed probability under repeated sampling. The interval width, $\mathrm{Width}(D) = U(D) - L(D)$, is computable in practice and summarizes the estimator's reported uncertainty on the value scale. Unlike $\widehat{\mathrm{ESS}}$, this quantity is tied directly to the reported uncertainty of the value estimate rather than only to the concentration of IS ratios.

This perspective addresses the shortcomings of weight-based ESS emphasized in this paper. First, CI width can respond to reward variance, which is invisible to a ratio-only proxy. Second, interval width can be reported for direct, hybrid, and fitted evaluators, not only for importance-weighted estimators. Third, the interval is expressed on the same value scale as $\widehat{V}(D)$, making it easier to compare reported uncertainty across estimator families, provided the interval construction is justified for each estimator.

Meanwhile, the problem of cross-estimator uncertainty comparison remains open; interval validity is still estimator-specific. In practice, intervals are usually obtained through one of three routes: an asymptotic normal approximation when a limiting distribution is available; a concentration-

based construction, often conservative but finite-sample motivated (Thomas et al., 2015; Liu et al., 2018); or a bootstrap procedure, when bootstrap validity has been established for the estimator in question (Hao et al., 2022). Thus, interval-based summaries should be understood here not as a final theoretical solution, but as a more promising practical framework for estimator-agnostic comparison.

5.2. A concrete positive example: bootstrap inference for FQE

Among the estimators discussed here, FQE provides the clearest currently available positive example of estimator-specific interval inference. Under linear function approximation and suitable regularity conditions, the FQE estimation error is asymptotically normal, its asymptotic variance is efficient, and an episode-level bootstrap consistently estimates its sampling distribution (Hao et al., 2022).

Theorem 1 (Bootstrap inference for FQE, after Hao et al. (2022))**.** *Under the regularity conditions studied by Hao et al. (2022), the FQE estimation error is asymptotically normal, and an episode-level bootstrap consistently estimates its sampling distribution. Consequently, the resulting bootstrap confidence intervals and bootstrap variance estimates are asymptotically valid for FQE under those assumptions.*

The episode-level qualifier is important. In sequential data, transitions within an episode are dependent, so resampling individual transitions does not in general preserve the dependence structure relevant for inference. The available bootstrap theory for FQE therefore works by resampling episodes rather than transitions (Hao et al., 2022).

The role of Theorem 1 in this paper is illustrative. It does not show that uncertainty quantification is equally well understood across all OPE families. Rather, it shows the kind of estimator-specific inferential result that is needed once one leaves the importance-sampling family. The uncertainty summary must be tied to the actual estimator and its own stochastic structure, not forced back into a weight-based proxy that was derived elsewhere.

6. Simulation Study

We evaluate the paper's claims with controlled tabular simulations. The simulation study answers three questions. First, within importance sampling, can $\widehat{\mathrm{ESS}}$ fail to capture uncertainty when the importance-ratio structure is fixed but reward noise changes? Section 6.2 answers this with a contextual-bandit reward-noise intervention, with an additional within-family scatterplot deferred to Appendix F. Second, across estimator families, do CI width and coverage provide a more useful comparison layer than a shared weight diagnostic? Sections 6.4 and 6.5 answer this with short-horizon tabular MDP experiments. Third, what happens in a harder long-horizon regime with stronger behavior-evaluation mismatch? Section 6.3 answers this with a long-horizon mismatch stress test, and Appendix G gives an additional FQE-specific bootstrap calibration check.

Across all experiments, ground-truth values $V(\pi_e)$ are computed exactly by dynamic programming from the known tabular model. Sequential intervals use episode-level bootstrap resampling. The reported practical simulation uses 200 bootstrap resamples for bootstrap intervals. The short-horizon coverage and width summaries in Figure 3 use 1000 repeated datasets per plotted condition, while the FQE bootstrap calibration study in Appendix G uses 200 repeated datasets per plotted condition.

6.1. Experimental design

The study uses three environment classes. The first is a contextual bandit with $S = 10$ states, $A = 5$ actions, and horizon $H = 1$. The second is a short-horizon tabular MDP with $S = 30$, $A = 3$, and horizon $H = 5$. The third is a long-horizon tabular MDP with $S = 40$, $A = 3$, and horizon $H = 20$. The estimator set consists of IS, SNIS, PDIS, DM, DR, MRDR, and FQE, though each experiment uses only the estimators relevant to its question.

The policy-mismatch parameter α controls the separation between the behavior policy π_b and evaluation policy π_e. For each environment, we first generate state-, action-, and, for MDPs, time-dependent policy logits $L_{t,s,a}$, and define

$$\pi_e(a \mid s, t) = \mathrm{softmax}(L_{t,s,\cdot}/0.85)_a.$$

We then define an anti-target policy

$$\pi_{\mathrm{anti}}(a \mid s, t) \propto \max_{a'} \pi_e(a' \mid s, t) - \pi_e(a \mid s, t) + 10^{-3}.$$

The behavior policy is the mixture

$$\pi_b(\cdot \mid s, t) = (1 - \alpha)\pi_e(\cdot \mid s, t) + \alpha\pi_{\mathrm{anti}}(\cdot \mid s, t),$$

followed by a behavior-probability floor of 0.02 and renormalization. The evaluation policy is also floored at 10^{-4} and renormalized. Thus $\alpha = 0$ makes behavior nearly identical to evaluation, while larger α moves behavior toward actions assigned lower probability by the evaluation policy. The parameter α is therefore a policy-mismatch interpolation parameter, not a bound on total variation distance, maximum probability difference, or maximum importance ratio.

For each estimator $\widehat{V}(D)$, we report the OPE absolute error $\mathrm{AE}(D) = |\widehat{V}(D) - V(\pi_e)|$, the empirical estimator variance across repeated datasets, the normalized-weight proxy $\widehat{\mathrm{ESS}}$ when applicable, and the mean 90% confidence-interval width. We also report empirical 90% coverage as a simulation calibration metric, not as a deployable uncertainty diagnostic, since it requires the known ground-truth

value $V(\pi_e)$. When studying diagnostic usefulness, we compute Spearman correlations between $\widehat{\text{ESS}}$ and absolute error and between CI width and absolute error, then aggregate those correlations across conditions.

The sample-size grids are experiment-specific. The reward-noise intervention uses $n \in \{100, 500, 1000\}$. The within-family scatterplot uses $n = 500$. The short-horizon cross-family coverage, width, and diagnostic-strength experiments use $n \in \{50, 100, 200, 300, 500, 1000\}$. The long-horizon mismatch stress test fixes $n = 300$. The FQE bootstrap calibration study uses $n \in \{50, 100, 200, 300, 500, 1000\}$. Full data-generating details for all three environment classes are given in Appendix A.

6.2. Bandit reward-noise sanity check

We begin with the cleanest empirical test of the internal limitation discussed in Section 4. In this experiment, the behavior policy, evaluation policy, and mismatch level are held fixed in a contextual bandit, while only the reward-noise scale changes. The reward-noise scale c ranges over $\{0.1, 0.25, 0.5, 1.0, 2.0, 4.0\}$, with sample sizes $\{100, 500, 1000\}$. In the implementation this parameter is named `reward_variance_scale`, but it multiplies the reward standard deviation; therefore the conditional reward variance scales as c^2. Figure 1 reports the $n = 500$ condition for IS and SNIS, the two estimators most directly tied to the ESS argument.

Figure 1 separates the weight diagnostic from reward-driven uncertainty. Mean $\widehat{\text{ESS}}$ remains essentially flat across the reward-noise grid, while estimator variance, mean absolute error, and mean CI width all increase sharply. Since the behavior-evaluation mismatch structure is fixed, the normalized-weight proxy remains governed by the empirical ratio structure. The uncertainty of the value estimate still increases because the conditional reward noise increases. This is the within-family failure mode predicted by theory.

The experiment does not imply that $\widehat{\text{ESS}}$ is useless inside importance sampling. It shows that $\widehat{\text{ESS}}$ cannot stand in for full uncertainty, because it is blind to changes in reward variance that do not materially alter the realized weights. Appendix F gives an additional within-family scatterplot analysis showing that, in a representative IS-family condition, the dataset-level association between $\widehat{\text{ESS}}$ and realized absolute error can also be weak.

6.3. Long-horizon mismatch stress test

The long-horizon tabular MDP experiment studies a regime where ratio-based methods visibly deteriorate under increasing behavior-evaluation policy mismatch. The sample size is fixed at $n = 300$, and the policy-mismatch interpolation parameter varies over $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. As defined in Section 6.1, larger α moves π_b away from π_e toward an anti-target policy, increasing the concentration of cumulative importance ratios. The estimators compared are PDIS, DM, DR, MRDR, and FQE.

Figure 2 shows the sharpest regime distinction in the simulations. As mismatch increases, PDIS deteriorates dramatically: panel A shows a sharp rise in mean absolute error, panel B shows the corresponding collapse in $\widehat{\text{ESS}}$, and panel C shows a dramatic increase in CI width. By contrast, DM, FQE, and to a lesser extent DR and MRDR, remain much more stable over the same mismatch range.

This is the appropriate interpretation of $\widehat{\text{ESS}}$. Within the IS-like family, and especially under large mismatch, the normalized-weight proxy is doing something real: it detects weight collapse. But the same scalar does not provide a common uncertainty scale for non-IS estimators, whose dominant uncertainty mechanisms are different. The stress test therefore reinforces both parts of the paper's argument: $\widehat{\text{ESS}}$ is meaningful in its native regime, yet non-portable across estimator families.

6.4. Coverage and width as cross-family interval summaries

If weight-based ESS is not estimator-agnostic, a natural alternative is to compare interval procedures through quantities defined on the value scale. Figure 3 reports empirical 90% coverage, used as a calibration metric, and mean 90% CI width, used as a reported uncertainty summary, over sample sizes $n \in \{50, 100, 200, 300, 500, 1000\}$ for DM, DR, FQE, MRDR, and PDIS.

Figure 3a uses the known simulation ground truth to compare calibration across estimator families. Coverage is poor for several estimators at very small sample sizes but generally improves as the number of episodes increases. PDIS starts closest to nominal coverage and remains relatively well calibrated across the grid. DR also approaches nominal coverage as n grows. DM, MRDR, and especially FQE under-cover at small sample sizes before improving substantially at larger n. Thus, empirical coverage exposes calibration differences that cannot be assessed from interval width alone.

Figure 3b provides the companion efficiency picture. PDIS produces the widest intervals at every sample size, although its widths shrink sharply as n increases. DR is also relatively wide, while DM, MRDR, and FQE are narrower. The width curves are not uniformly monotone: for example, FQE width initially increases before decreasing at larger sample sizes. This reinforces the main point that width should be interpreted together with coverage; a narrow interval is attractive only if it is also reasonably calibrated.

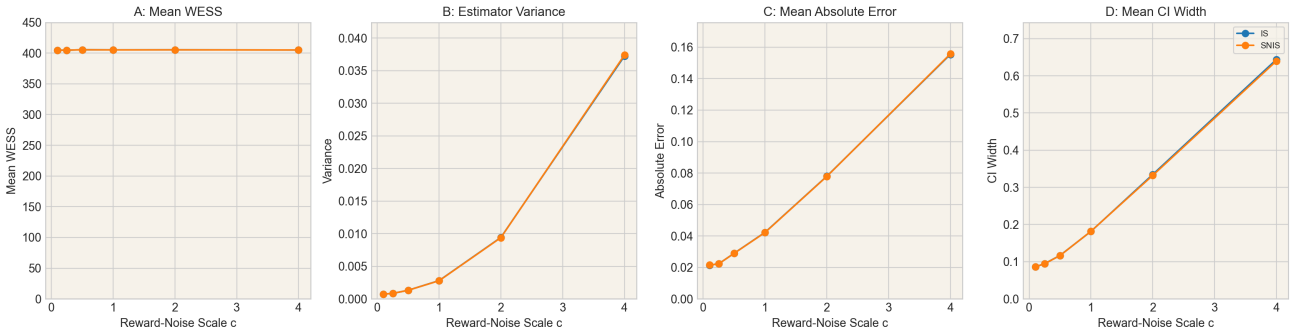


Figure 1. Bandit reward-noise sanity check. Panel A shows that mean normalized-weight ESS is nearly unchanged as the reward-noise scale c increases. Panels B–D show that estimator variance, mean absolute error, and mean CI width increase over the same range. The IS and SNIS curves are nearly indistinguishable in this experiment.

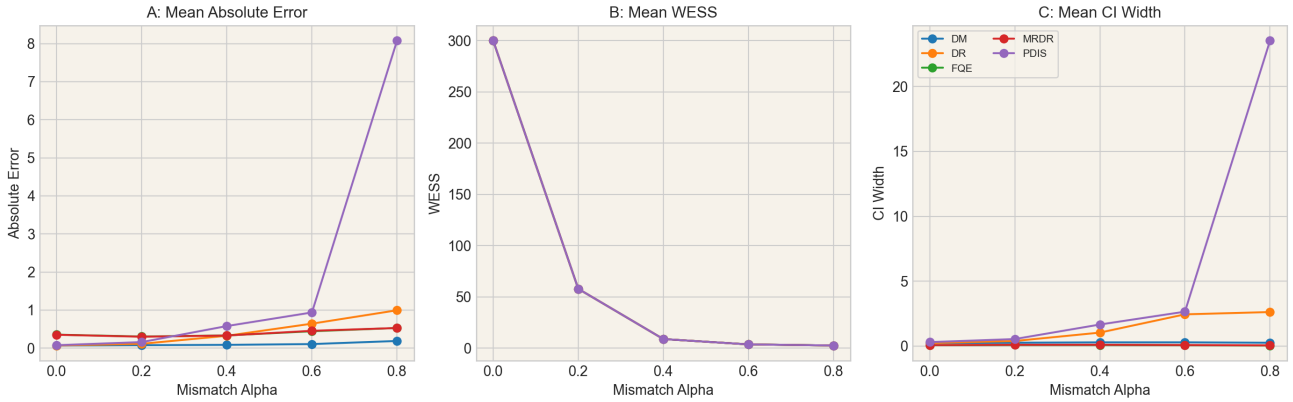
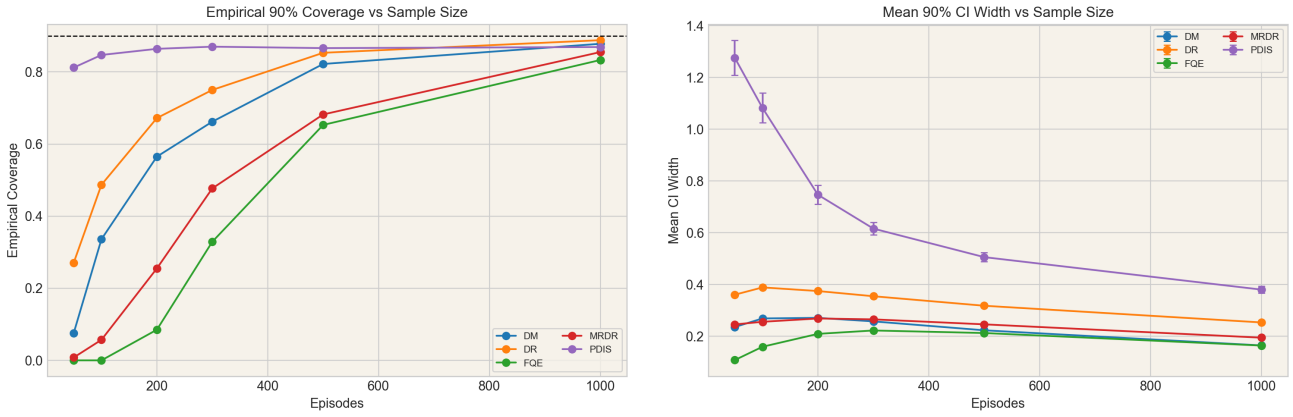


Figure 2. Long-horizon mismatch stress test. Panel A plots mean absolute error versus mismatch α. Panel B plots mean normalized-weight ESS. Panel C plots mean 90% CI width. As mismatch increases, PDIS shows a sharp error blow-up, corresponding collapse in normalized-weight ESS, and dramatic CI-width inflation. The model-based and fitted estimators remain much more stable.



(a) Empirical 90% coverage.

(b) Mean 90% CI width.

Figure 3. Coverage and width as cross-family interval summaries in the short-horizon MDP. Panel A reports empirical 90% coverage versus sample size, with the dashed line marking nominal 0.9 coverage. Panel B reports mean 90% CI width versus sample size, with standard-error bars over plotted interval widths pooled across mismatch conditions. Width should be interpreted together with coverage: narrow intervals are useful only when they are also reasonably calibrated.

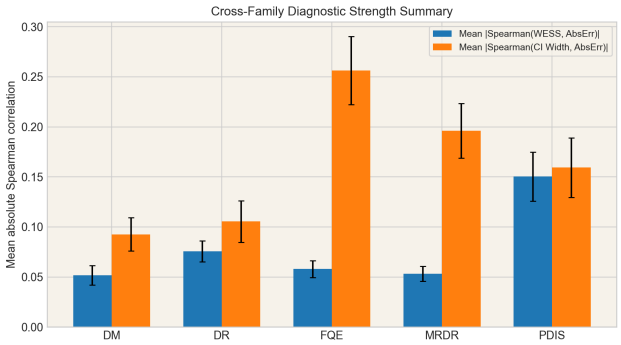


Figure 4. Cross-family diagnostic strength summary. Bars show condition-matched mean absolute Spearman correlations with realized absolute error. Error bars denote standard errors across condition-level correlations. CI width is more strongly associated with absolute error than normalized-weight ESS for DR, FQE, and MRDR, while PDIS is the main case where the two diagnostics are competitive.

Together, the two panels in Figure 3 show why interval-based reporting is more informative than a shared weight diagnostic, but also why it is not a complete solution. CI width gives a value-scale uncertainty summary that can be reported for all interval-equipped estimators. Empirical coverage gives a calibration check in simulation.

6.5. Cross-family diagnostic strength

We next summarize diagnostic usefulness directly by comparing dataset-level rank associations with realized absolute error. The short-horizon tabular MDP experiment varies the behavior-evaluation mismatch parameter $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$ and sample size $n \in \{50, 100, 200, 300, 500, 1000\}$, and compares PDIS, DM, DR, MRDR, and FQE. For each condition, we compute the absolute Spearman correlation between $\widehat{\text{ESS}}$ and absolute error, and the analogous correlation between CI width and absolute error.

Figure 4 provides an aggregate version of the comparison. For DR, FQE, and MRDR, CI width is more informative about realized absolute error than $\widehat{\text{ESS}}$. For DM, both summaries are weak. PDIS is the main case where $\widehat{\text{ESS}}$ remains competitive with CI width, consistent with its IS-like structure. This supports the portability claim: $\widehat{\text{ESS}}$ can be useful in its native regime, but its usefulness is family-dependent, whereas CI width is a more portable value-scale uncertainty summary.

6.6. Bootstrap-based FQE inference

We also include an FQE-specific bootstrap calibration study, since the theory section identifies FQE as a concrete example where estimator-specific bootstrap inference has existing asymptotic support. Appendix G reports empirical 90% coverage and mean CI width for bootstrap-based FQE in short- and long-horizon settings over sample sizes $n \in \{50, 100, 200, 300, 500, 1000\}$. The short-horizon setting uses the $S = 30, A = 3, H = 5$ MDP, and the long-horizon setting uses the $S = 40, A = 3, H = 20$ MDP. Both use $\alpha = 0.4$.

The results support a careful positive claim rather than a sweeping one. In the short-horizon case, bootstrap-based FQE coverage improves with sample size and approaches the nominal level by the largest sample sizes. In the long-horizon case, coverage remains poor through small and moderate sample sizes and improves only at $n = 1000$. The width behavior is also not monotone. In the short-horizon setting, mean CI width increases at small sample sizes before decreasing at larger n. In the long-horizon setting, mean CI width increases across the sample-size grid. A diagnostic audit confirmed that this pattern is not due to plotting, aggregation, bootstrap-quantile, stale-cache, or sample-labeling errors: bootstrap resampling is episode-level, FQE is refit on each bootstrap resample, and widths are computed as upper minus lower after quantile computation. We therefore interpret Appendix G as evidence of finite-sample instability in bootstrap-based FQE inference, especially in long-horizon settings, rather than as evidence of uniformly reliable finite-sample calibration.

6.7. Main empirical takeaway

Taken together, the simulations give four main conclusions. First, within the importance-sampling family, normalized-weight ESS tracks weight concentration but fails to track all uncertainty sources: when reward variance changes while the mismatch structure is held fixed, actual uncertainty changes sharply while $\widehat{\text{ESS}}$ barely moves. Second, under long-horizon mismatch, $\widehat{\text{ESS}}$ correctly detects the collapse of IS-like estimators, but this does not make it a common uncertainty scale for non-IS estimators. Third, CI width is reportable for all interval-equipped estimators, while empirical coverage provides a simulation-based calibration check; the two must be interpreted jointly because narrow intervals can under-cover. Fourth, FQE supplies a concrete case where estimator-specific bootstrap inference can be studied directly, but the results also show that finite-sample calibration and width behavior remain strongly regime-dependent.

7. Conclusion

This paper argued that effective sample size should be interpreted relative to the estimator family from which it arises. Within SNIS and related importance-sampling estimators, the normalized-weight proxy $\widehat{\text{ESS}} = 1/\sum_i \bar{w}_i^2$ remains a meaningful diagnostic of weight degeneracy, but it does not capture all relevant sources of uncertainty even there.

Across direct, hybrid, and fitted evaluators such as DM, MRDR, and FQE, a single weight-based ESS cannot serve as an estimator-agnostic uncertainty summary. This leaves cross-estimator uncertainty comparison as an open problem. CI width provides a more portable value-scale uncertainty summary, while empirical coverage provides a simulation-based calibration check for the interval procedure; both must still be justified for the estimator under study.

Acknowledgments

This work was supported in part by an Emory URC Research Award to ST. The authors thank the RLC 2024 "I Can't Believe It's Not Better" Workshop participants for helpful discussions, as well as the anonymous reviewers for constructive feedback.

Data and Code Availability

The code for all experiments is available at https://github.com/aditya1909-bit/ess-ope-diagnostics.

References

Shayan Doroudi, Vincent Aleven, and Emma Brunskill. Where's the reward? a review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education*, 29:568–620, 2019.

Víctor Elvira, Luca Martino, and Christian P. Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2022.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christoph Mosch, Li-Wei Lehman, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1): 16–18, 2019.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvári, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. *arXiv preprint arXiv:2102.03607*, 2022.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Augustine Kong. A note on importance sampling using standardized weights. Technical report, The University of Chicago, 1992. https://d3qi0qp55mx5f5.cloudfront.net/stat/docs/tech-rpts/tr348.pdf.

Hoang M. Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

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

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Proceedings of Machine Learning Research*, volume 149, pages 1–34, 2021.

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

Takuma Udagawa, Haruka Kiyohara, Yusuke Narita, Yuta Saito, and Kei Tateno. Policy-adaptive estimator selection for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10025–10033, 2023.

Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *NeurIPS Datasets and Benchmarks*, 2021.

A. Simulation Data-Generating Processes

This appendix gives the complete data-generating specification for the simulations. All environments are finite-horizon tabular environments with uniform initial-state distribution. Conditional on (t, s, a), rewards are Gaussian:

$$R_t \mid S_t = s, A_t = a \sim \mathcal{N}(\mu_{t,s,a}, \sigma^2_{t,s,a}).$$

The transition tensor, reward means, reward standard deviations, and policies are generated from a fixed environment seed. Repeated datasets within a condition are then generated independently from that fixed condition.

Policy construction. For each environment, evaluation-policy logits $L_{t,s,a}$ are generated first. The evaluation policy is

$$\pi_e(a \mid s, t) = \frac{\exp(L_{t,s,a}/0.85)}{\sum_{a'} \exp(L_{t,s,a'}/0.85)}.$$

The anti-target probabilities are

$$\pi_{\text{anti}}(a \mid s, t) = \frac{\max_{a'} \pi_e(a' \mid s, t) - \pi_e(a \mid s, t) + 10^{-3}}{\sum_b \{\max_{a'} \pi_e(a' \mid s, t) - \pi_e(b \mid s, t) + 10^{-3}\}}.$$

The behavior policy is

$$\widetilde{\pi}_b(\cdot \mid s, t) = (1 - \alpha)\pi_e(\cdot \mid s, t) + \alpha \pi_{\text{anti}}(\cdot \mid s, t).$$

A minimum probability floor of 0.02 is applied to $\widetilde{\pi}_b$, followed by renormalization. A floor of 10^{-4} is applied to π_e, also followed by renormalization.

For the contextual bandit, logits are

$$L_{s,a} = X_s + Y_a + 0.20 Z_{s,a},$$

where $X_s \sim \mathcal{N}(0, 0.45^2)$, $Y_a \sim \mathcal{N}(0, 0.55^2)$, and $Z_{s,a} \sim \mathcal{N}(0, 1)$ independently. For the MDPs,

$$L_{t,s,a} = B_{t,s,a} + u_t + v_a + W_s,$$

where $B_{t,s,a} \sim \mathcal{N}(0, 0.6^2)$, u_t is evenly spaced on $[-0.4, 0.4]$, v_a is evenly spaced on $[-0.6, 0.6]$, and $W_s \sim \mathcal{N}(0, 0.3^2)$.

Contextual bandit. The contextual bandit has $S = 10$, $A = 5$, and $H = 1$. The transition is degenerate:

$$P(S_1 = s \mid S_0 = s, A_0 = a) = 1.$$

Let x_s be evenly spaced on $[-1, 1]$, y_a evenly spaced on $[-0.6, 0.6]$, u_s evenly spaced on $[0.8, 1.2]$, and v_a evenly spaced on $[0.9, 1.1]$. The reward mean is

$$\mu_{s,a} = 0.45 \sin(1.7 x_s) + 0.25 x_s + 0.35 y_a + 0.20 \cos(1.2 x_s) y_a.$$

The reward standard deviation is

$$\sigma_{s,a} = c\, u_s v_a,$$

where $c \in \{0.1, 0.25, 0.5, 1.0, 2.0, 4.0\}$ in the reward-noise sanity check. The policy mismatch parameter is fixed at $\alpha = 0.4$ in that experiment.

Short-horizon MDP. The short-horizon MDP has $S = 30$, $A = 3$, and $H = 5$. For each (t, s, a), a support set of four next states is sampled uniformly without replacement. If the current state s is not in the sampled support, it replaces the first support element. Independent weights are then sampled from $\text{Unif}(0.1, 1.0)$, the weight assigned to state s is increased by 1.0, and the row is normalized to define $P_t(\cdot \mid s, a)$.

Let τ_t be evenly spaced on $[0, 1]$, x_s evenly spaced on $[-1, 1]$, and y_a evenly spaced on $[-0.8, 0.8]$. The reward mean is

$$\mu_{t,s,a} = 0.60 \sin\{2.2\pi(x_s + 0.18\tau_t)\} + 0.25 y_a + 0.10 \epsilon_{t,s,a},$$

where $\epsilon_{t,s,a} \sim \mathcal{N}(0, 1)$ independently. The reward standard deviation is $\sigma_{t,s,a} = 0.20$ for all (t, s, a).

Long-horizon MDP. The long-horizon MDP has $S = 40$, $A = 3$, and $H = 20$. Transitions are generated as in the short-horizon MDP, except that the support size is five and the self-transition weight increase is 0.8.

Rewards are sparse and mostly terminal. A set G of $\max(2, \lfloor S/8 \rfloor) = 5$ goal states is sampled uniformly without replacement. Each goal state $g \in G$ is assigned a preferred action a_g sampled uniformly from $\{1, \dots, A\}$. At the final time $H - 1$,

$$\mu_{H-1,g,a_g} \leftarrow 6.0, \qquad \mu_{H-1,g,a} \leftarrow \mu_{H-1,g,a} + 0.08 \cdot 6.0 \quad \text{for all } a.$$

The penultimate reward table receives 0.15 times the final reward table. A small dense component equal to 0.08 times the short-horizon dense reward function is then added at all times. The reward standard deviation is $\sigma_{t,s,a} = 0.18$ for all (t, s, a).

B. Summary of the Standard ESS Approximation

As discussed in Section 3, the practical normalized-weight proxy $\widehat{\mathrm{ESS}} = 1/\sum_{i=1}^N \bar{w}_i^2$ arises from a sequence of approximations applied to the self-normalized importance-sampling variance-ratio definition $\mathrm{ESS} = N\,\mathrm{Var}(\widehat{I}_{\mathrm{MC}})/\mathrm{Var}(\widetilde{I}_{\mathrm{SNIS}})$, where $\widehat{I}_{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^N h(X_i)$ with $X_i \sim \pi$, and $\widetilde{I}_{\mathrm{SNIS}} = \sum_{i=1}^N \bar{w}_i h(X_i)$ with $X_i \sim q$, $w_i = \pi(X_i)/q(X_i)$, and $\bar{w}_i = w_i/\sum_{j=1}^N w_j$.

The SNIS estimator can be written as the ratio $\widetilde{I}_{\mathrm{SNIS}} = \big(\frac{1}{N}\sum_{i=1}^N w_i h(X_i)\big)/\big(\frac{1}{N}\sum_{i=1}^N w_i\big)$. A delta-method expansion around the population means of the numerator and denominator then gives a variance approximation under i.i.d. sampling from a single proposal. Together with the approximation $\mathrm{MSE}(\widetilde{I}_{\mathrm{SNIS}}) \approx \mathrm{Var}(\widetilde{I}_{\mathrm{SNIS}})$, which ignores finite-sample bias, and subsequent moment simplifications, this leads to the practical form $\mathrm{ESS} \approx N/(1 + \mathrm{Var}_q(W))$ with $W = \pi(X)/q(X)$. Replacing population moments by empirical plug-in estimates yields

$$\widehat{\mathrm{ESS}} = N\frac{\Big(\frac{1}{N}\sum_{i=1}^N w_i\Big)^2}{\frac{1}{N}\sum_{i=1}^N w_i^2} = \frac{1}{\sum_{i=1}^N \bar{w}_i^2}.$$

Thus, the final proxy depends only on normalized empirical weights, which explains both its convenience and its limits.

C. Reward-Variance Counterexample

We expand the argument in Proposition 1. Consider a contextual bandit with i.i.d. samples (s_i, a_i, r_i), where actions are drawn under π_b, evaluation is with respect to π_e, and $\rho_i = \pi_e(a_i \mid s_i)/\pi_b(a_i \mid s_i)$. The ordinary importance-sampling estimator is $\widehat{V}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^n \rho_i r_i$.

Construct two worlds $\mathcal{W}_1$ and $\mathcal{W}_2$ with the same state distribution, behavior policy, and evaluation policy, and couple them so that they share the same sampled state-action pairs (s_i, a_i). Then the realized importance ratios are identical sample-by-sample. Consequently, any weight-based proxy computed from those ratios is also identical in the two worlds.

Now let the reward distribution differ: in $\mathcal{W}_1$, $r_i \mid (s_i, a_i) \sim R_1(\cdot \mid s_i, a_i)$, while in $\mathcal{W}_2$, $r_i \mid (s_i, a_i) \sim R_2(\cdot \mid s_i, a_i)$, with $\mathrm{Var}_{R_1}(r \mid s, a) \neq \mathrm{Var}_{R_2}(r \mid s, a)$ for at least one state-action pair of positive probability. Since the samples are i.i.d., $\mathrm{Var}(\widehat{V}_{\mathrm{IS}}) = \frac{1}{n}\mathrm{Var}(\rho r)$. By the law of total variance, $\mathrm{Var}(\rho r) = \mathrm{Var}(\mathbb{E}_{r|s,a}[\rho r]) + \mathbb{E}_{(s,a)}[\mathrm{Var}(\rho r \mid s, a)]$. Because ρ is determined by (s, a), we have $\mathrm{Var}(\rho r \mid s, a) = \rho(s, a)^2 \mathrm{Var}(r \mid s, a)$. Hence changing $\mathrm{Var}(r \mid s, a)$ changes the estimator variance even when the entire realized ratio structure is fixed. Therefore the two worlds can have identical weight-based ESS proxies but different estimator variances.

D. Additional Discussion of Cross-Family Non-Portability

The proof of Proposition 2 is given in the main text. Here we provide further intuition. A ratio-only diagnostic is a function of the behavior-evaluation mismatch observed in the dataset. This is a native object for IS-like estimators, because those estimators are themselves built from importance ratios. However, it is not a native object for model-based or fitted estimators.

For example, two datasets may have the same empirical importance-ratio vector but lead to different fitted value models because their rewards, transitions, or feature realizations differ. A direct estimator or FQE estimator can therefore have different uncertainty across those datasets even when the ratio-only diagnostic is unchanged. This illustrates why the issue is not merely that $\widehat{\mathrm{ESS}}$ is imperfect, but that its target object is not shared across estimator families.

E. Additional Cross-Family $\widehat{\text{ESS}}$ Scatter

Figure 5 plots shared $\widehat{\text{ESS}}$ against absolute error across estimator families in the short-horizon MDP. The clustered structure reflects that $\widehat{\text{ESS}}$ is a property of the behavior-evaluation mismatch and dataset, not a native uncertainty diagnostic for every estimator. The plot is therefore used as a diagnostic visualization rather than as the main cross-family comparison.

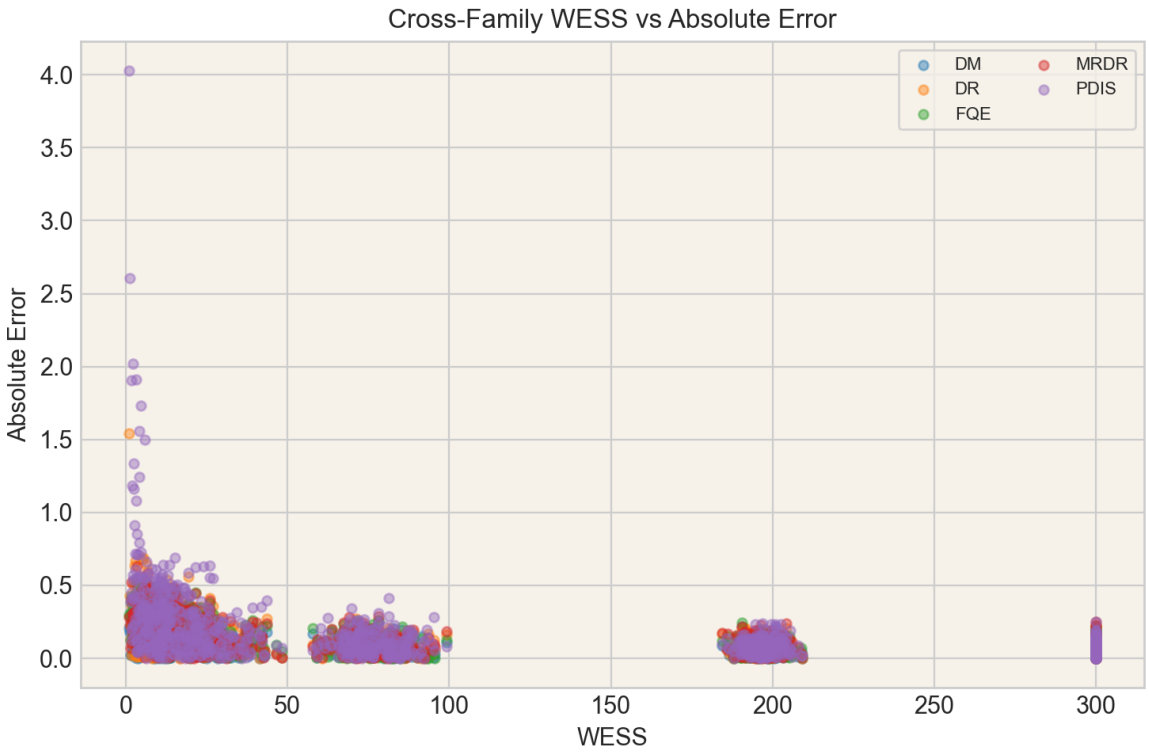


Figure 5. Cross-family relationship between shared $\widehat{\text{ESS}}$ and absolute error. The scatter shows that the same $\widehat{\text{ESS}}$ values can correspond to different error behavior across estimator families.

F. Additional Within-Family Diagnostic Plot

Figure 6 reports the dataset-level relationship between $\widehat{\text{ESS}}$ and absolute error for IS and SNIS in a representative within-family condition. The condition is the contextual bandit with $n = 500$, $\alpha = 0.4$, and reward-noise scale $c = 1.0$. The reported Spearman correlations are close to zero in both panels, showing a weak and noisy relationship between normalized-weight ESS and realized absolute error.

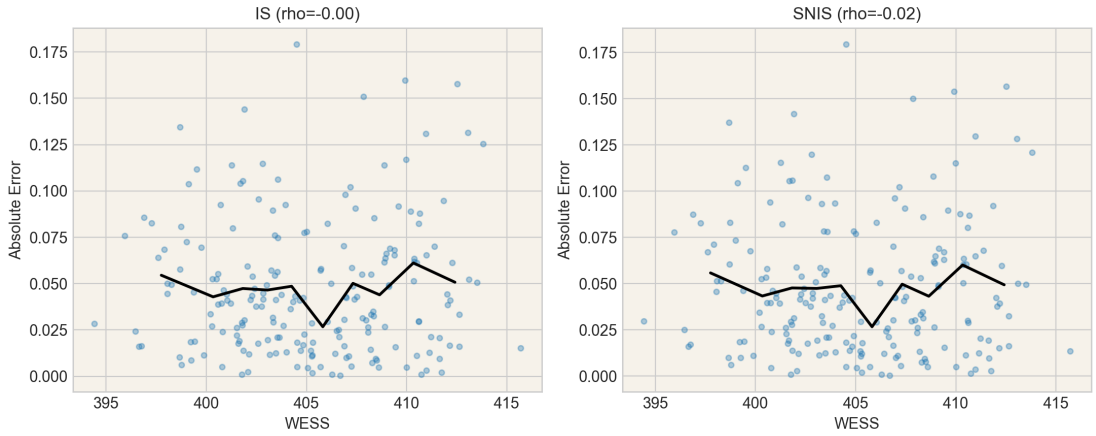


Figure 6. Within-family relationship between $\widehat{\mathrm{ESS}}$ and absolute error for IS and SNIS.

G. Additional FQE Bootstrap Calibration Plot

Figure 7 reports empirical 90% coverage and mean CI width for bootstrap-based FQE in short- and long-horizon settings. Coverage improves with sample size, but the long-horizon case remains substantially harder. The corresponding widths are not monotone: short-horizon width increases at small sample sizes before decreasing, while long-horizon width increases across the grid. This behavior was verified as genuine under the regenerated practical simulation rather than a plotting, aggregation, sample-labeling, or bootstrap-implementation bug.

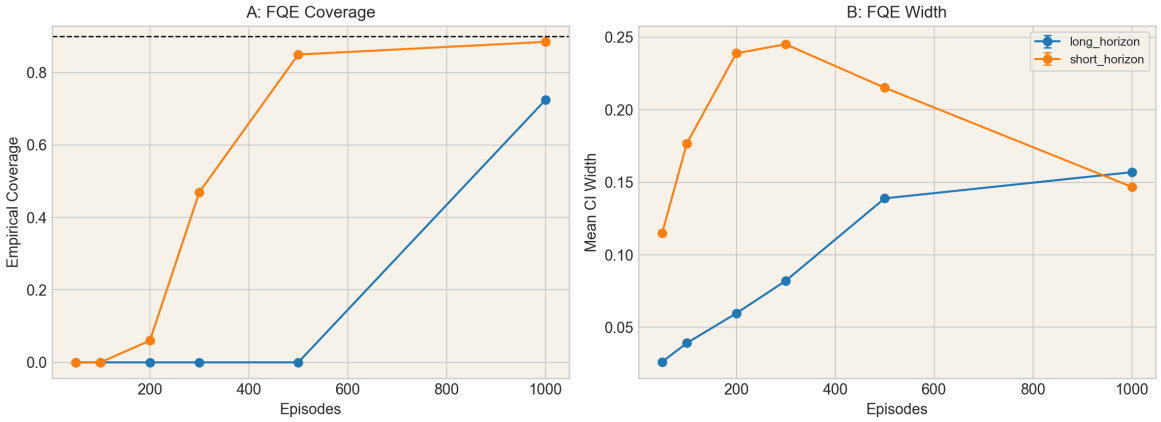


Figure 7. Bootstrap-based FQE calibration. Panel A shows empirical 90% coverage for short- and long-horizon settings. Panel B shows the corresponding mean CI width. The nonmonotone width behavior, especially the increasing long-horizon width, reflects finite-sample instability of the bootstrap interval procedure in this harder regime.