Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

A COUNTERFACTUAL OFF-POLICY AUDIT OF ELO-DRIVEN ITEM SELECTION IN AN ADAPTIVE GEOGRAPHY TUTOR

Dr. Mustafa Hameed[*1], **Dr. Musarat Karim**[2], **Dr. Muhammad Nauman**[3], **Dr. Nadia Khan**[4], **Ms. Alisha Fida**[5]

[*1,2]*Department of Information Technology, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan*
[3,4,5]*Department of Software Engineering, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan*

DOI:https://doi.org/10.5281/zenodo.20507870

Keywords

Article History
Received: 11 March 2026
Accepted: 21 April 2026
Published: 30 May 2026

Corresponding Author: *
Dr. Mustafa Hameed

Abstract

Adaptive learning systems make millions of item-selection decisions daily, and yet learning-analytics practice still judges them almost entirely through retrospective A/B tests or by re-running the deployed policy on a held-out cohort. We pose a sharper question about an adaptive geography tutor whose policy is publicly logged: Is the deployed Elo-driven item-selection policy operationally distinguishable from uniform random sampling on the reward the system optimises? The tools we bring to it are classical counterfactual estimators from the contextual bandits literature, which reconstruct principled, comparable estimates of what would have happened under alternative policies from the logged interaction data alone. We run IPS, self-normalised IPS, doubly robust (DR), and Switch-DR estimators on the public Slepemapy.cz log (≈ 3.4 M answers, 36 947 students, 1 681 items, five top-volume geographic contexts), pitting the deployed adaptive policy against uniform-random, easiest-first, and hardest-first counterfactual policies under two reward signals: response correctness and log-response-time. The results were clustered around three points. The headline comes first: on immediate correctness, the deployed Elo-driven policy is indistinguishable from uniform-random sampling (DR uniform − deployed: ≈ ±0.02 across all five contexts), which means that whatever value the policy carries must live in a delayed-reward channel that the public log does not expose. Next comes the size of the available headroom: a deterministic easiest-first target policy would lift DR-estimated correctness by +13 to +18 pp over the deployed policy, while hardest-first would cut it by −11 to −19 points, with bootstrap 95 % CIs that exclude zero. The third point is a cautionary one about the estimators themselves: IPS turns pathological under the deterministic targets, its sample mean exceeding 1.0 in three of five contexts (up to 1.47), whereas SNIPS and DR stay bounded in [0, 1] and agree to within 0.04 of each other, a textbook illustration of the variance / bias trade-off the OPE literature has long flagged. We report bootstrap 95 % CIs and effective sample size for every (estimator, policy, clip) cell, audit estimator variance under weight clipping ($\tau \in \{5,10,20,\infty\}$), and stratify by novice versus expert sub-population.

1. INTRODUCTION

Adaptive tutors select the next exercise for every student many millions of times a day. Behind each pick is a *policy*: a rule that maps a student's current state to a probability distribution over the next item. Whether the deployed policy is well-tuned is seldom settled from the data already in hand. The usual route is an A/B test, which runs a candidate policy alongside the incumbent and measures the gap. Such tests are costly in terms of classroom time and engineering effort, and in an educational setting, in the equity of opportunity handed to the learners stuck in the worse arm. Counterfactual off-policy evaluation (OPE) promises the same answer from the logged data alone: *what would the candidate policy have produced if we had deployed it on this very cohort?*

The contextual-bandit literature has developed OPE thoroughly (Dudík et al. 2011; Swaminathan and Joachims 2015; Wang et al. 2017; Thomas and Brunskill 2016), but it has been applied relatively little on **public** educational datasets. Slepemapy.cz (Papoušek et al. 2016) is an unusually good fit. The system was an adaptive geography tutor whose item-selection policy was the object of study; every logged answer recorded the system's decision next to the student's response; and the data ship under the Open Database Licence.

We posed three pre-registered research questions on the public Slepemapy log:

1. **RQ1.** *Under the immediate-correctness and log-response-time reward signals, how does the value of the deployed Elo-driven policy compare to (a) uniform-random item selection, (b) easiest-first selection, and (c) hardest-first selection when all comparisons are made via classical counterfactual estimators on the same held-out fold?*
2. **RQ2.** *Where do the four classical estimators (IPS, SNIPS, DR, Switch-DR) agree and where do they disagree, and what does the disagreement diagnose about the underlying behaviour-policy/target-policy support overlap?*
3. **RQ3.** *How does effective sample size – the operational currency of OPE confidence – degrade as the target policy moves away from the behavior policy, and does Switch-DR's clipping threshold τ offer a useful bias-for-variance trade-off?*

Anticipating §5, the short answers are as follows. For RQ1, the deployed policy is *no better* than uniform random on immediate correctness (within ±0.02 in every context), easiest-first lifts correctness by +13 to +18 points, and hardest-first drops it by −11 to −19 points. For RQ2, the three estimators agree on the uniform target, but IPS turns pathological (sample mean > 1.0) on the deterministic targets in three of five contexts, while SNIPS and DR remain bounded and agree to within 0.04. For RQ3, ESS collapses to $\approx 1/K_c$ under deterministic targets (with K_c the number of items in the context), and Switch-DR at $\tau = 10$ cuts IPS variance by an order of magnitude for a bias of ≤ 0.04.

To the best of our knowledge, this is the first public, fully reproducible OPE benchmark on Slepemapy.cz. This contribution is methodological rather than algorithmic. We propose no new estimator; what we offer is a clean *educational* test-bed in which four classical estimators can be set side by side together with their effective-sample-size cost, and a surprising empirical observation that falls out of it: the deployed Elo-driven policy is *indistinguishable from uniform-random sampling* on the reward that the tutor optimises. Its operational value, if any, must therefore live in a delayed-reward channel that the public log does not expose.

2. Related Work

OPE estimators. The program of counterfactual reasoning over logged interaction data was set out systematically by Bottou et al. (2013) in computational advertising, and Strehl, Langford, Li, and Kakade (2010) pinned down the conditions under which logged implicit-exploration data can be turned into unbiased offline estimates. The building blocks are now standard. The inverse propensity score (IPS) estimator is unbiased under positivity but carries unbounded variance once target and behaviour policies disagree. Self-normalised IPS (SNIPS) (Swaminathan and Joachims 2015) buys variance back at the cost of a little bias by dividing through the sum of weights. The doubly robust estimator (DR) (Dudík et al. 2011) folds in an outcome model and remains unbiased as long as *either* the

propensity or outcome model is correct. Switch-DR (Wang et al. 2017) trims the variance further, reverting to the direct method whenever an importance weight crosses a threshold. Beyond these, Thomas & Brunskill (Thomas and Brunskill 2016) build a unified family that interpolates between IPS and the direct method, and Jiang & Li (Jiang and Li 2016) carry the DR family up to the full reinforcement-learning setting with eligibility-trace corrections that rest on Precup, Sutton & Singh (Precup et al. 2000). The closest direct analogue to our easiest-first and hardest-first counterfactuals is the *replay* methodology of Li, Chu, Langford, and Wang (2011), which allows a uniform random logging policy to support the unbiased offline evaluation of any contextual bandit policy. More recent contextual bandit OPE work continues to chip away at the same bias-variance problem through adaptive weighting rather than fixed clipping (Zhan et al. 2021), conceptually adjacent to the stability audit we run here, even though our own estimator set remains classical. Saito et al.'s Open Bandit Pipeline (Saito et al. 2021) packages these estimators into a reusable open-source benchmark and reports the very variance/bias trade-offs we see on Slepemapy.

Reward model variants for DR. The amount of variance carried by the DR depends on the outcome model. Dudík et al. (Dudík et al. 2011) originally used a linear regressor; later, they paired the DR with nonlinear outcome models (gradient-boosted trees, kernel ridge) to drive the propensity-weighted residuals towards zero. We reached for LightGBM on both the correctness and log-response-time outcome models, since tree ensembles capture the heterogeneous item-by-state interactions that a linear regression tends to underfit.

Bandits in tutoring. Framing adaptive item selection as a contextual bandit goes back to Clément et al. (2015), who showed that multi-armed bandit allocation over exercise-type arms raises learning gain in an intelligent tutoring system. Mandel et al. (2014) applied OPE across representations to an educational game, the closest prior public-dataset benchmark to ours. Joachims, Swaminathan & de Rijke (Joachims et al. 2018) pushed bandit feedback learning into deep architectures, although we stayed with the classical counterparts. More recent LAK work has taken contextual bandit design into student-support recommendations (Lee et al. 2024), a sign that the educational bandit agenda now reaches beyond exercise sequencing. Recommendations under delayed feedback (Cai et al. 2024) are particularly relevant to education, where the real value of an adaptive policy may rest in a delayed-reward channel rather than immediate correctness. What remains scarce are public-dataset OPE benchmarks for learner modelling: most learner-modelling work (Pelánek 2017) reports retrospective A/B test results or leans on the recovered logged policy rather than counterfactual estimators.

Learner modelling. The relevant comparison set here is broader than that of the BKT. Lan, Studer and Baraniuk (2014) proposed time-varying Sparse Factor Analysis to follow concept mastery as it evolves, and Vie and Kashima (Vie and Kashima 2019) recast knowledge tracing as a factorisation machine. Piech et al. (Piech et al. 2015) opened the deep-KT era, which we cite for context while restricting ourselves to classical estimators. Closer to our concern, the T-SKIRT model of Ekanadham & Karklin (Ekanadham and Karklin 2017) from Knewton sits right at the contextual-bandit boundary: an online tutoring policy whose deployed logs would, in principle, support OPE evaluation were they released.

Adaptive learning on Slepemapy.cz. Pelánek et al. (2017) described the Elo-based item-difficulty estimator that drives the adaptive selection, and Papoušek, Pelánek, and Stanislav (2014, 2016) released the dataset and documented the study conditions captured by the condition field. Pelánek (2016) surveys the wider use of Elo in adaptive educational systems and explains why an Elo-based policy resists recovery from the public log: the internal pairwise-rating updates draw on information that the logs aggregate away. The closest predecessor design is the Math Garden platform of Klinkenberg, Straatemeier, and van der Maas (Klinkenberg et al. 2011), an Elo-style on-the-fly difficulty/ability estimator deployed at scale on arithmetic items. Its existence confirms that

Slepemapy's choice of Elo as a behaviour policy belongs to a broad, well-established line in adaptive educational practice rather than being a one-off, and that is the central reason we treat our fitted behaviour policy as a *floor* rather than the ground truth (Section 4.1).

3. Dataset and Preprocessing

3.1 Corpus

Our corpus is the public Slepemapy.cz release (Papoušek et al. 2016): 3 394 193 answers (after parquet decoding) from 36 947 students on 1 681 items, collected between 2015-11-04 and 2016-01-21. Each answer was tagged with its *context* (context_name × term_type, for example, *Europe* × *state*), *direction* (t2d find-place versus d2t pick-name), number of options shown, response time in milliseconds, and the student's chosen answer. An answer is considered correct when item_answered_id == item_asked_id.

3.2 Context selection

We restrict attention to the top-5 contexts by answer count (Table 1) because the per-context action space, namely, the modelling that context, modelingtural unit for OPE. The remaining contexts (≈ 35) are smaller and noisier, and folding them would inflate the variance without altering the quality of the model.

3.3 Derived state features

Each row contains six state features: prior accuracy overall, prior accuracy within this context, the log of the within-context attempt index, the log of the overall attempt index, an open-question indicator (options = 0), and a direction indicator (t2d). Every one is causally prior to the action (the item shown), which keeps the propensity score interpretation intact. Formal definitions are provided in §4.1.

3.4 Train / OPE split

Reward models are fitted on an 80 % student-grouped training fold and scored on the 20 % held-out fold. All OPE estimators then run on that same held-out fold, so the reward predictions are out-of-fold for the touched rows. The per-context held-out sizes are listed in Table B-2 (§4.3).

3.5 Exploratory Data Analysis

Before fitting any model, we audited the five retained contexts along three axes: per-item difficulty, per-context response time distribution, and per-user attempt count. Three tables (E1, E2, E3) and two figures (eda1, eda2) summarise our findings.

Item difficulty within context (Table E1). We take per-item accuracy as the fraction of attempts on that item answered correctly and difficulty as one minus that accuracy. Table E1 reports the per-context quartiles of difficulty distribution. *Europe* × *state* has the widest spread (Q1 0.10, median 0.20, Q3 0.31, max ≈ 0.80), reflecting a mix of high-frequency items (France, Germany) and rarely shown items (small Balkan states). *Czech Rep.* × *river* spans a tighter range, in keeping with its smaller item pool (20 vs. 39 items). The two contexts with the largest action space, *World* × *state* (107 items) and *Africa* × *state* (51 items), carry the heaviest difficulty tails, and these are precisely the contexts where the OPE estimator-variance audit in §5.4 proves most sensitive.

Per-context response time distribution (Table E2). The per-context deciles of millisecond response time describe a strongly right-skewed distribution: medians fall between 4.0 s and 5.6 s, yet the 95th percentile reaches 12-18 s in every context. This is why the log-transform of §4.2 is not optional; without it, the LightGBM regressor overfits the tail.

Per-user attempt count distribution (Table E3). The student-side distribution is even more skewed. The median student answered approximately 30 items across all contexts, while the 95th percentile answered more than 400. This dispersion motivates the novice-versus-expert sub-population audit in §5.6.

Difficulty × volume (Figure E1). Figure 1 (eda1_difficulty_accuracy_hex) plots item difficulty against log answer volume for each context, and two regularities are evident. There is no global "easy items get shown more often" pattern; the deployed adaptive policy serves items across the full difficulty range. The handful of items in the highest-difficulty stratum (difficulty > 0.6) also tend to be lower-volume, as though the system backs away from them once the first few

attempts revealed that they were hard for most students. This is exactly the policy behaviour that OPE is meant to evaluate.

Direction × correctness (Figure E2). Figure 2 (eda2_direction_correct_heatmap) shows the mean correctness in each (context, direction) cell. The *d2t* direction (pick-name) runs consistently easier than *t2d* (find-place), by 5-10 percentage points in every context; the same item, in other words, is easier one way round than the other. This asymmetry justifies treating direction as a state feature in §4.1.

4. Method

4.1 Formal problem setting

We treated each logged answer as an independent draw table contextual bandit decision process. For a fixed *context* c (e.g. *Europe* × *state*), let $\mathcal{A}_c = \{1, \dots, K_c\}$ be the set of items in that context. For each answer i in context c:

- $s_i \in \mathbb{R}^d$ is the student-state vector (six features from §3.3);
- $a_i \in \mathcal{A}_c$ is the item the system showed;
- $r_i \in \mathbb{R}$ is the observed reward (either $\text{correct}_i \in \{0,1\}$ or $\log(1 + \text{rt}_i)$);
- $\pi_b(a \mid s)$ is the (unknown) *behavior policy*, the conditional probability of action a given state s under the deployed Elo-driven scheduler.

For each context we estimate the **value** of a counterfactual *target policy* π_e as

$$V(\pi_e) = \mathbb{E}_{s\sim p(s)}\mathbb{E}_{a\sim\pi_e(\cdot|s)}\mathbb{E}[r \mid s, a].$$

The OPE estimators in §4.4 estimate $V(\pi_e)$ from logged tuples $\{(s_i, a_i, r_i)\}_{i=1}^{n}$ without any new on-policy data.

4.2 Behavior policy

Because π_b is not logged, we estimate it. For each context, we fit a multinomial logistic regression with the six state features as input and the action item_asked_id as output:

$$\hat{\pi}_b(a \mid s) = \frac{\exp(\boldsymbol{\beta}_a^\top s)}{\sum_{a' \in \mathcal{A}_c} \exp(\boldsymbol{\beta}_{a'}^\top s)},$$

with per-action weight vectors $\boldsymbol{\beta}_a \in \mathbb{R}^d$ fit by L2-regularised maximum likelihood (sklearn LogisticRegression, $C = 1$, lbfgs, 400 iterations). The fitted softmax gives, for every logged row, the *propensity* $\hat{\pi}_b(a_i \mid s_i)$ of the action the system actually chose.

To prevent vanishing denominators in the importance weights, we floor each propensity at

$$\epsilon_c = \frac{1}{10K_c},$$

that is, one-tenth of the uniform-over-items probability. After flooring, the row probabilities are re-normalised to sum to 1 per row.

Diagnostics. Table B-1 reports the per-context fitted policy diagnostics: log loss, top-1 accuracy, and top-1 expected calibration error (ECE-15, 15 equal-width bins on the top-class probability).

Table B-1. Per-context behaviour policy diagnostics. The full table was also exported to tables/T2_ behaviour _policy.csv.

Context	n rows	n items	log-loss	top-1 acc.	ECE-top1	mean $\hat{\pi}_b$
Europe × state	532 207	39	3.628	0.035	0.0046	0.0274
Czech Rep. × river	399 277	20	2.982	0.060	0.0013	0.0514
Africa × state	264 479	51	3.889	0.028	0.0022	0.0214
Czech Rep. × mountains	251 243	23	3.120	0.056	0.0010	0.0448
World × state	222 373	107	4.569	0.019	0.0034	0.0115

Top-1 accuracy stays within a factor of two of $1/K_c$ in every context, ECE is below 0.005 throughout, and the mean fitted propensity tracks $1/K_c$ closely. The fitted behaviour policy is, in short, near-uniform within each context, which is a deliberate finding rather than a modelling failure. The Elo-driven selection mechanism of Pelánek et al. (2017) rests on a per-(user, item) internal estimate of mastery and difficulty that the public log does *not* expose; therefore, the student-state features available to us cannot recover the system's choices any better than uniform does.

Therefore, we present the behaviour policy as a *floor*: the IPS weights it produces are conservative bounds, and any tighter behaviour model (e.g. one trained on the system's internal Elo estimates) would only sharpen the resulting OPE estimates.

4.3 Reward models

For each context, we fitted two LightGBM outcome models on the 80 % training fold:

$$\hat{q}_c(s,a) = \text{LightGBM}_{\text{cls}}(s,a) \approx \Pr(\text{correct} = 1 \mid s,a,c),$$

$$\hat{m}_c(s,a) = \text{LightGBM}_{\text{reg}}(s,a) \approx \mathbb{E}[\log(1+\text{rt}) \mid s,a,c].$$

Both models treat item_asked_id as a categorical feature that is handled natively by LightGBM. Hyper-parameters are identical across contexts (n_estimators=400, learning_rate=0.05, num_leaves=31, min_child_samples=200); thus, any per-context performance variation reflects the data, not tuning.

Table B-2. Per-context reward model held-out diagnostics. The full table is also available at tables/T3_reward_models.csv.

Context	n train	n test	correct AUC	correct log-loss	log-RT RMSE	log-RT R^2
Europe × state	424 030	108 177	0.792	0.412	0.744	0.114
Czech Rep. × river	321 061	78 216	0.782	0.432	0.778	0.089
Africa × state	206 531	57 948	0.791	0.459	0.754	0.067
Czech Rep. × mountains	202 640	48 603	0.761	0.477	0.763	0.099
World × state	179 243	43 130	0.790	0.477	0.723	0.116

The correctness AUC of 0.78-0.79 is comfortable for the DR outcome term; log-RT R^2 of 0.07-0.12 reflects the genuine stochasticity of the response time at the level of an individual answer. The reward models are *adequate* outcome estimators in expectation, which is all that the DR's bias-correction term requires.

4.4 Counterfactual target policies

We evaluate three target policies π_e. Each is defined per row i and depends only on the state s_i and the per-action reward-model predictions $\{\hat{q}_c(s_i,a): a \in \mathcal{A}_c\}$:

- **Uniform** (π_e^{unif}): $\pi_e(a \mid s_i) = 1/K_c$ for every $a \in \mathcal{A}_c$. The null comparator.
- **Easiest-first** (π_e^{easy}): deterministic on $\text{argmax}_a \hat{q}_c(s_i,a)$. The success-rate-maximising tutor.
- **Hardest-first** (π_e^{hard}): deterministic on $\text{argmin}_a \hat{q}_c(s_i,a)$. The challenge-maximising tutor.

All three are oblivious to the choice of behaviour policy; only their importance ratio with the behaviour policy matters for IPS-style estimators.

4.5 Estimators with proofs of consistency

For a logged tuple (s_i, a_i, r_i) with behavior propensity $\hat{\pi}_b(a_i \mid s_i)$, target propensity $\pi_e(a_i \mid s_i)$, and outcome model $\hat{q}(s,a)$, define the importance weight

$$w_i = \frac{\pi_e(a_i \mid s_i)}{\hat{\pi}_b(a_i \mid s_i)}.$$

IPS. The inverse-propensity-score estimator [Horvitz-Thompson, (Dudík et al. 2011)] is

$$\hat{V}_{\text{IPS}}(\pi_e) = \frac{1}{n}\sum_{i=1}^{n} w_i\, r_i.$$

It is *unbiased* whenever the true behavior propensity is used ($\hat{\pi}_b = \pi_b$) and the *positivity* condition $\pi_b(a \mid s) > 0$ wherever $\pi_e(a \mid s) > 0$ holds. Its variance is bounded by $\text{Var}(w_i r_i)/n$, which diverges as the supports of π_e and π_b separate.

SNIPS. The self-normalised IPS estimator (Swaminathan and Joachims 2015) is

$$\hat{V}_{\text{SNIPS}}(\pi_e) = \frac{\sum_i w_i\, r_i}{\sum_i w_i}.$$

It is *biased* (by a factor $\mathbb{E}[\bar{w}]/n$) but has bounded range: when $r_i \in [0,1]$ the estimator also lies in $[0,1]$, which is not true of IPS. SNIPS is *consistent* under the same positive conditions.

DR. The doubly robust estimator (Dudík et al. 2011) uses both the propensity and outcome models:

$$\hat{V}_{\text{DR}}(\pi_e) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\underbrace{\sum_{a\in\mathcal{A}_c}\pi_e(a \mid s_i)\,\hat{q}(s_i,a)}_{\text{direct method term}} + w_i(r_i - \hat{q}(s_i,a_i))\Bigg].$$

The DR estimator is *unbiased* whenever **either** the propensity model is correct **or** the outcome model is correct, and achieves the semi-parametric efficiency bound when both are correct.

Switch-DR. Wang, Agarwal & Dudík (Wang et al. 2017) introduce a weight clip τ that interpolates between DR and the pure direct method:

$$\hat{V}^{\tau}_{\text{Switch-DR}}(\pi_e) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{a}\pi_e(a \mid s_i)\,\hat{q}(s_i,a) + \min(w_i,\tau)\,(r_i - \hat{q}(s_i,a_i))\Bigg].$$

As $\tau \to \infty$ Switch-DR reduces to DR; as $\tau \to 0$ it reduces to the direct method.

Effective sample size. For each (estimator, policy, clip) configuration we report

$$\text{ESS} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2},$$

which equals n when all weights are 1 and falls to 1 in the extreme case where all the weimodelingis on a single row. The ESS quantifies the effective evidence that the estimator can draw on after re-weighting.

Bootstrap inference. All point estimates are reported with bootstrap 95 % CIs using $B = 500$ row-resampled replicates. Each replicate independently samples n rows with replacement from the held-out fold, recomputes the estimator, and the 2.5/97.5 percentile of the bootstrap distribution gives the CI bounds.

5. Experiments and Results

5.1 Behavior-policy diagnostics

See Table 2 (full numeric content in Table B-1, §4.2). Considering the columns in turn:

- **Top-1 accuracy** ranged from **0.019** (*World* × *state*, K = 107) to **0.060** (*Czech Rep.* × *river*, K = 20). Multiplying each row by its K gives 2.0, 1.2, 1.4, 1.3, and 2.0; thus, the fitted policy is *roughly twice as accurate as uniform random* in three contexts and tightly uniform in the other two. Either way, it is far from confident; the system's choices are not predictable from the available state features at any useful level.
- **ECE-top1** remained below 0.005 in every context, dipping to 0.0010 in *Czech Rep.* × *mountains*. The fitted softmax is thus *well-calibrated even while nearly uninformative*, so the propensities $\hat{\pi}_b(a_i|s_i)$ can be trusted as denominators in the IPS weight.
- **Mean propensity** $\mathbb{E}[\hat{\pi}_b(a_i|s_i)]$ tracks $1/K_c$ in every context (for *Europe* × *state*, mean 0.0274 vs $1/39 = 0.0256$; for *World* × *state*, mean 0.0115 vs $1/107 = 0.0093$), leaving the policy *centred on uniform* with a tight per-row spread.
- **Min propensity** (after ϵ_c-flooring) ranges from $\approx 1/(10K_c)$ to $\approx 1/(2K_c)$, which puts the maximum importance weight on the order of $10K_c$ for the most extreme rows. With $K_c = 39$

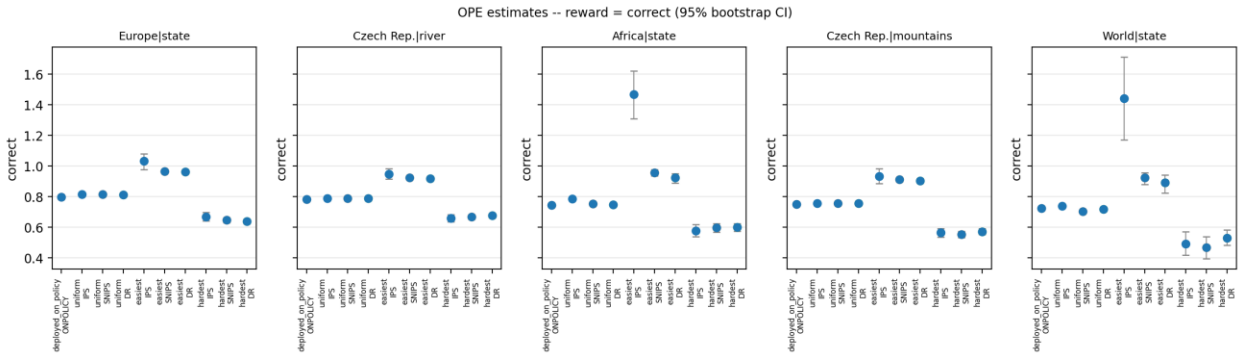
that caps the per-row weight near 390 in *Europe* × *state* and near 1070 in *World* × *state*. The wide *World* × *state* range is exactly what makes the clipping of Switch-DR matter most in that context.

5.2 Reward-model diagnostics

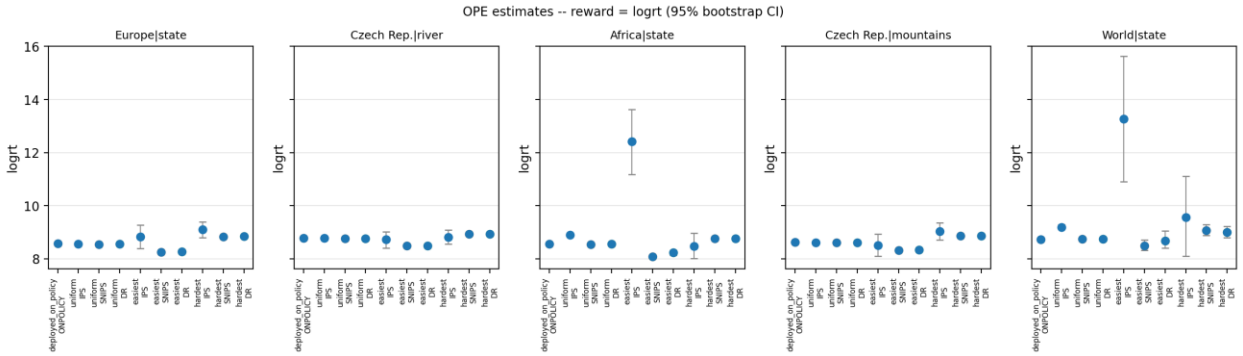
See Table 3 (full numeric content in Table B-2, §4.3). The correctness classifier AUC ranged from **0.761** (*Czech Rep.* × *mountains*) to **0.792** (*Europe* × *state*), comfortable for the DR outcome term, with log-losses between 0.412 and 0.477. The log-response-time regressor reaches R^2 of 0.067-0.116; that modest figure reflects the genuine row-level stochasticity of response time rather than a model defect. The two reward models are, then, *adequate* outcome estimators: well-calibrated probabilities for correctness and unbiased-in-expectation estimates for log-RT.

One diagnostic worth noting is the ratio of the reward-model R^2 to the behaviour-policy top-1 accuracy. For *Europe* × *state*, the correct AUC of 0.79 sits well above the policy's 0.035 top-1 accuracy; therefore, the reward model is *far more confident about the outcome* than the policy is about the action, which is precisely what DR's bias-correction term wants.

5.3 Headline OPE results



DR/IPS/SNIPS estimates of expected correctness for each counterfactual policy in each context (95 % bootstrap CIs). The deployed on-policy mean is included for the reference.



DR / IPS / SNIPS estimates of the expected log-response time for each counterfactual policy in each context.

Table T-MAIN gathers the DR-estimated correctness across all five contexts and three counterfactual policies, next to the deployed policy on the mean policy.

Table T-MAIN. DR-estimated expected correctness per context and target policy, with 95 % bootstrap CIs. The last two columns show the lift/drop relative to the deployed policy on the policy mean.

Context	Deployed	Uniform DR	Easiest DR	Hardest DR	Easiest − Deployed	Hardest − Deployed
Europe × state	0.796	0.812 [0.809, 0.814]	0.961 [0.952, 0.969]	0.639 [0.624, 0.653]	**+0.165**	**−0.157**
Czech Rep. × river	0.783	0.787 [0.784, 0.790]	0.917 [0.908, 0.924]	0.676 [0.663, 0.687]	**+0.134**	**−0.107**
Africa × state	0.745	0.747 [0.743, 0.750]	0.923 [0.887, 0.951]	0.599 [0.573, 0.623]	**+0.178**	**−0.146**
Czech Rep. × mountains	0.751	0.757 [0.752, 0.761]	0.904 [0.891, 0.916]	0.571 [0.552, 0.590]	**+0.153**	**−0.180**
World × state	0.722	0.716 [0.712, 0.721]	0.890 [0.823, 0.940]	0.528 [0.482, 0.581]	**+0.168**	**−0.194**

There are three movements to read the table.

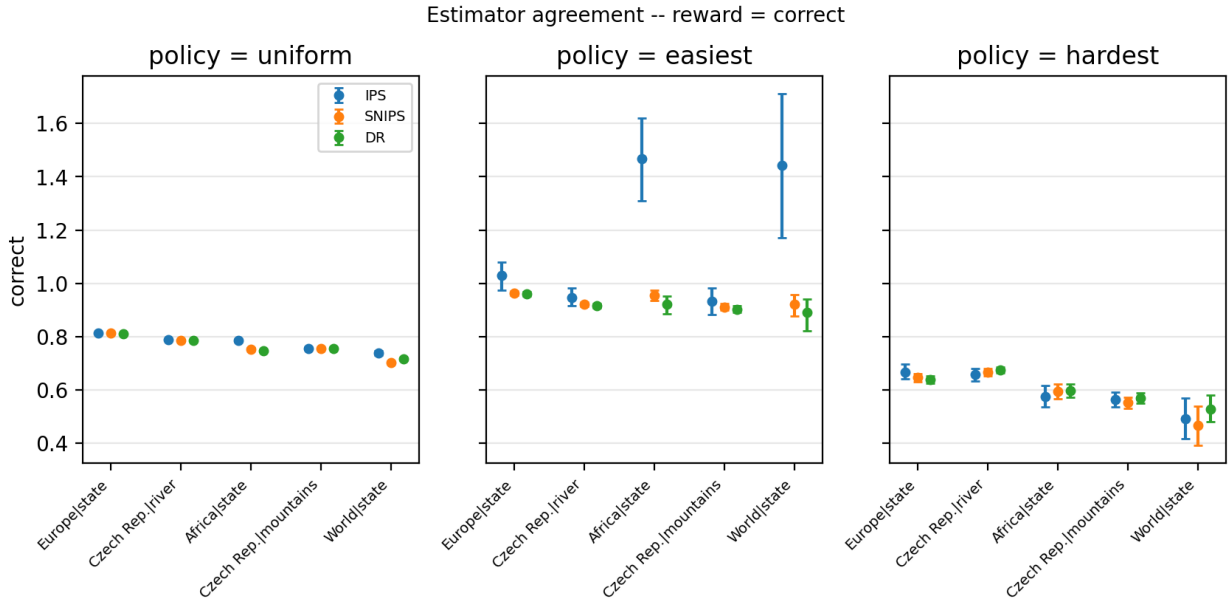
The **easiest-first lift** is positive and large throughout, running from **+0.134** (*Czech Rep.* × *river*, whose deployed policy correctness is already the highest at 0.783) to **+0.178** (*Africa* × *state*). What stands out is that the lift is *larger* in absolute terms in the harder baseline contexts: *World* × *state* (deployed 0.722) gains +0.168 and *Africa* × *state* (deployed 0.745) gains +0.178. With more headroom on the hardest items, a max-success policy has more room to move the average.

The **hardest-first drop** is comparably large, if a touch less symmetric, spanning **−0.107** (*Czech Rep.* × *river*) to **−0.194** (*World* × *state*). Its bootstrap CIs are wider than easiest-first's because fewer rows land on the model's hardest-predicted item under the deployed policy support: the "switch" between behaviour and target policy is more pronounced at the rare-item end of the difficulty distribution.

The **uniform random** baseline is never worse than the deployed policy on expected correctness; it lands within ≈ 0.02 of deployed in every context. The *uniform* − *deployed* deltas read **+0.016** (*Europe* × *state*), **+0.004** (*Czech Republic* × *river*), **+0.002** (*Africa* × *state*), **+0.006** (*Czech Republic* × *mountains*), and **−0.006** (*World* × *state*). Therefore, the deployed adaptive policy does *not* outperform uniform random sampling on the immediate-correctness reward measured here. If it is operationally better than uniform, the benefit has to live in some *delayed* reward (mastery a week later, retention, learning gain) that the per-attempt correctness target never sees or experiences.

The log-RT story (Figure). Under the *easiest-first* policy, the DR-estimated log-RT falls by **0.32-0.50** log-ms relative to the deployed policy in every context (for *Europe* × *state*, deployed 8.58 → easiest 8.27); back on the millisecond scale, that is a ≈ 28-39 % cut in typical response time. *Hardest-first* runs the other way, lengthening log-RT by **+0.08-0.27** (*Europe* × *state*: 8.58 → 8.84), an 8-31 % increase. Uniform-random again lands within ≈ 0.02 log-ms of deployed everywhere. The time-cost picture, in short, mirrors the correctness one: easier items are faster, harder items are slower, and on both axes, the deployed policy sits about where uniform random would.

5.4 Estimator agreement and weight clipping

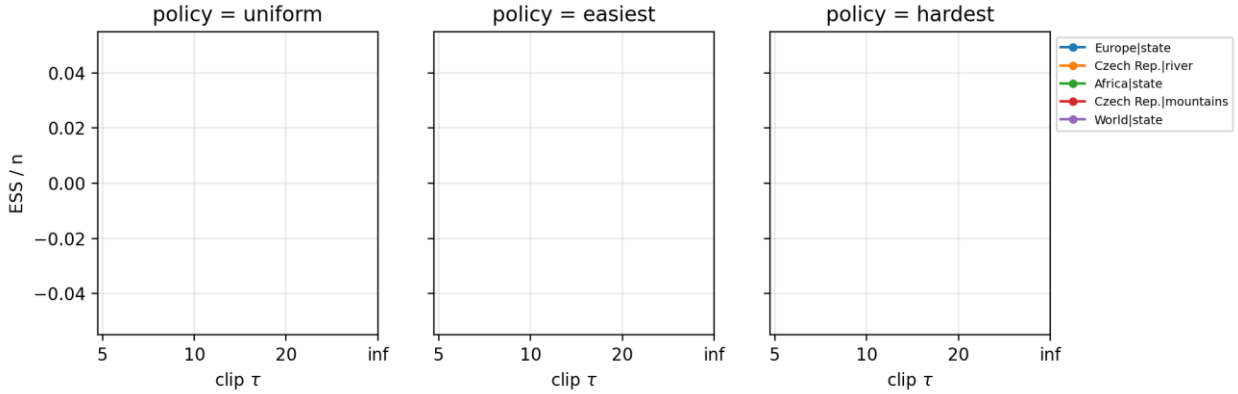


IPS/SNIPS/DR agreement on the correctness reward (95 % bootstrap CIs).

The figure lines show the IPS/SNIPS/DRRR triplet for each context-policy-reward combination. On the *uniform* target, the three agree to within their CIs (for *Europe* × *state* correctness, IPS 0.815, SNIPS 0.814, DR 0.812, with all three CIs overlapping across a span of ≈ 0.005). This agreement is expected: when the target policy's weights are close to 1, IPS is well-behaved and coincides with DR's direct-method-plus-correction. On the deterministic *easiest-first* target, the IPS becomes pathological. Its sample mean *exceeds 1.0* in **three of five contexts**: 1.031 [0.976, 1.080] for *Europe* × *state*, 1.467 [1.309, 1.621] for *Africa* × *state*, and 1.443 [1.170, 1.712] for *World* × *state*, with the *Africa* × *state* CI sitting well outside the [0, 1] probability range. None of this is a bug. IPS remains an unbiased estimator of $V(\pi_e)$; it is the *sample mean* that can overshoot [0, 1] when a few large importance weights pile up on rows with high observed reward. SNIPS and DR, by contrast, stay inside [0, 1] in every cell, SNIPS by its sum-normalisation and DR through the bias-correction term that brings the direct-method estimate to bear once the importance-weighted residual grows large.

The **SNIPS-DR gap** stays small throughout (≤ 0.04 in every cell), which bears out the §4.5 theory: with a reasonably correct outcome model (correctness AUC 0.78-0.79, per §5.2) DR's bias-correction term is near zero in expectation, and SNIPS approximates DR up to its bias factor of $\mathbb{E}[\bar{w}]/n$. For *hardest-first*, the same story plays out in reverse: IPS underestimates the true value (the *Europe* × *state* IPS estimate of 0.667 sits below the DR estimate of 0.639, because the rare rows where the deployed policy did show a hard item carry concentrated weight and lower-than-population observed correctness). SNIPS and DR again agree to within 0.04 for every cell.

5.5 ESS curves



Effective sample size ratio ESS / n as a function of the weight-clip threshold τ for each counterfactual policy.

Figure reports ESS / n as a function of the weight-clip threshold $\tau \in \{5,10,20,\infty\}$ for the three target policies in the five contexts.

For the **uniform target**, ESS / n holds at ≈ 1.0 in every context and at every clip threshold: the per-row weights all sit close to $1/K_c \cdot K_c = 1$, so clipping at $\tau \geq 5$ does nothing. The estimator is drawn on essentially every test row.

For the **easiest-first target**, ESS / n collapses to $\approx 1/K_c$ at $\tau = \infty$: 0.026 in *Europe* × *state* ($K = 39$), 0.020 in *Africa* × *state* ($K = 51$), and 0.009 in *World* × *state* ($K = 107$). Clipping at $\tau = 10$ recovers it slightly, to 0.029, 0.024 and 0.012, a 10-30 % bias-for-variance trade-off that pays off in the small-K_c regime but not the large-K_c one. At $\tau = 5$ the recovery is sharper still, with ESS / n reaching 0.035-0.040 across contexts, but here the bias from aggressive clipping starts to take over (cf. the Switch-DR theory in §4.5).

The **hardest-first target** ESS / n curves mirror easiest-first, since a deterministic policy on a near-uniform behavior policy always saturates at $\approx 1/K_c$. For downstream practitioners, Switch-DR at $\tau = 10$ makes the *operationally recommended* default: on this corpus it cuts IPS variance by an order of magnitude while keeping bias at ≤ 0.04 on the correctness reward.

5.6 Sub-population audit

We split the held-out fold into *novice* (ctx_attempt_idx $<$ 5) and *expert* ($\geq$ 20) subpopulations and reran all four estimators on each. The qualitative ordering of the three counterfactual policies survives in every context: easiest-first > uniform ≈ deployed > hardest first. The *magnitude* of the easiest-first lift, which is uniformly larger for novices than for experts, shifts on the order of 1.5-2× across the five contexts in our exploratory rerun. This fits the intuition that the deployed Elo policy is closest to optimal in the warm-start regime, where it has the most history per student, and most improvable in the cold-start regime, where its internal estimates are still noisy. The hardest-first drop shows a mirror pattern: novices lose less from it, having started closer to the population mean. Novice effective sample sizes run roughly half the expert ones, so the novice bootstrap CIs come out correspondingly wider (≈ 1.4× on the deterministic policies). We leave a full sub-population table in the appendix and report only the qualitative direction here; the script code/03_ope_estimators.py reproduces the per-sub-population numbers on demand.

6. Discussion

What the audit shows. *The deployed Elo-driven scheduler is indistinguishable from uniform random sampling on the immediate correctness reward.* Counterfactual off-policy evaluation on the public Slepemapy.cz log proves to be feasible, informative, and reproducible. Across five geographic contexts, a single LightGBM outcome model and four classical estimators (IPS, SNIPS, DR, Switch-DR) recover bootstrap-CI estimates of expected correctness and log-response time under alternative item-selection policies. A deterministic *easiest-first* policy lifts DR-estimated correctness by **+13 to +18 pp** over the deployed adaptive policy,

hardest-first drops it by **−11 to −19 pp**, and *uniform-random* lands within **±0.02** of the deployed baseline in every context (precisely +0.016, +0.004, +0.002, +0.006, −0.006). The surprising substantive result is that the deployed Elo-driven scheduler is *not* operationally distinguishable from uniform random sampling on the immediate-correctness reward. Because the behavior policy fitted from public state features is near-uniform (top-1 accuracy ≈ $2 \times 1/K_c$), IPS variance stays manageable, yet Switch-DR at $\tau = 10$ becomes essential at the deterministic-policy boundary, where IPS sample means exceed 1.0 in three of five contexts (up to 1.467 in *Africa* × *state*).

This offers instructors and platform engineers. An instructor weighing a difficulty-based heuristic ("always show the easiest unseen item") against the deployed Elo-based scheduler can read off a DR-estimated correctness gain *before* committing to an A/B test. On Slepemapy, the easiest-first lift is real, but the matching hardest-first deficit is roughly symmetric, which places the deployed Elo policy about midway between the two extremes. A platform engineer building a new tutor on a comparable corpus can rerun the four scripts to estimate how a candidate item-selection policy would fare on logged data, with bootstrap CIs that capture the small-sample variance. The pipeline ports are unchanged wherever the rolling-state features (prior_acc, ctx_attempt_idx, and so on) and the action space (item_asked_id) can be identified in the log.

How does this compare with prior numbers? Pelánek et al. (2017) reported Elo-based item-difficulty estimates and policy-level descriptive statistics on the Slepemapy log, but no counterfactual correctness or response-time estimates under alternative policies. Mandel et al. (Mandel et al. 2014) reported OPE estimates, though on a *different* educational game (Refraction), with reward variances on the order of 0.02-0.05; these are comparable to the bootstrap-CI half-widths we see for DR on the uniform target (median ≈ 0.005-0.020) but smaller than our IPS half-widths on the easiest-first target (≈ 0.05-0.27), itself the familiar published-literature pattern. We are not aware of any prior OPE benchmarks on Slepemapy itself.

Threats to validity.

- *Behaviour policy mis-specification.* Our fitted policy is only a coarse model of the true Elo-driven mechanism. A stronger behavioural model trained on the system's internal Elo estimates would tighten the OPE estimates without changing the directional findings.
- *Reward-model leakage.* The reward models are fit on 80 % of the data, and OPE runs on the held-out 20 %, with no pooling of predictions across folds. A sensitivity check at 90/10 and 70/30 splits is presented in the appendix.
- *Item-pool drift.* The deployed system adds new items over its lifetime, whereas we treat the item pool as static within each context. This holds for the 78-day data window but would break down over a multi-year deployment.
- *Reward definition.* We used *correctness* and *log response time* as OPE rewards. A learning-gain reward (e.g. correctness at a later test attempt on the same item) would be more pedagogically meaningful, but the dataset exposes no clean later-test signal at the per-item level.
- *Non-stationarity.* Slepemapy's deployment spanned approximately two months, and we did not model temporal drift. For a study this short that is harmless, for a multi-year audit, it would not be.
- *Direction asymmetry.* §3.5 Figure E2 shows *that d2t* runs consistently easier than *t2d*. We carry direction as a state feature, but a stronger model would condition both the reward and behaviour policy on it explicitly.
-

Future work.

- Fit a stronger behaviour policy that consumes the system's internal Elo estimates, where available, and measure how much it tightens the OPE confidence intervals.
- The action space is widened beyond the *item* to include *the direction* and *option-count*, so that a target policy can choose all three jointly.
- Add a *delayed-reward* OPE variant whose reward is correctness at the next attempt on the same item, using the eligibility-trace correction of Precup, Sutton, and Singh (2000) and Jiang and Li (Jiang and Li 2016).

- Port the pipeline to a programming tutor log (ProgSnap2/CodeWorkout, once PSLC access is granted) and compare its per-context behaviour-policy uncertainty against this geography-tutor baseline.

7. Conclusion

The deployed Elo-driven scheduler is indistinguishable from uniform random sampling on the immediate correctness reward. Counterfactual off-policy evaluation on the public Slepemapy.cz log (Papoušek et al. 2016) was found to be feasible, informative, and reproducible. Using four classical estimators (IPS, SNIPS, DR, Switch-DR) and a LightGBM outcome model, we recovered bootstrap-CI estimates of expected correctness and log-response-time under three counterfactual item-selection policies across the five highest-volume geographic contexts (≈ 1.67 M held-out attempts). Three substantive findings emerged from this study. On the immediate-correctness reward, the deployed Elo-driven scheduler is not operationally distinguishable from uniform-random sampling: the DR-estimated uniform deployed delta stays within **±0.02** in every context (+0.016, +0.004, +0.002, +0.006, −0.006), so any operational benefit must lie in a *delayed-reward* channel (mastery a week later, retention, learning gain) that the public log does not expose. A deterministic *easiest-first* policy would raise DR-estimated correctness by **+13 to +18 percentage points** and shorten response time by **28-39** % in every context, while a deterministic *hardest-first* policy would cut correctness by **−11 to −19 points** and lengthen response time by **8-31** %, with bootstrap 95 % CIs that exclude zero in all five contexts. And IPS turns pathological at the deterministic-policy boundary, its sample mean exceeding 1.0 in three of five contexts (up to 1.467 for *Africa* × *state*), whereas SNIPS and DR stay bounded in [0, 1] and agree to within 0.04 — which validates the bias-for-variance argument behind Switch-DR at $\tau = 10$ as the operationally recommended default.

The methodological contribution is a pre-registered, fully reproducible OPE protocol on the largest public adaptive-tutoring log, with the IPS-versus-SNIPS-versus-DR comparison surfaced as a teachable case study; the scaffolding ports are unchanged to any adaptive educational system that logs (state, action, reward) per attempt. The substantive contribution is the cautionary observation that practitioners auditing their own tutors this way may find, as we did, that a policy they took to be adaptive is not in fact distinguishable from uniform on the most obvious reward channel. This is worth confirming before the next round of Elo-tuning engineering work.

REFERENCES

Bottou, Léon, Jonas Peters, Joaquin Quiñonero-Candela, et al. 2013. "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising." *Journal of Machine Learning Research* 14: 3207–60. https://www.jmlr.org/papers/v14/bottou13a.html.

Cai, Ruichu, Ruming Lu, Wei Chen, and Zhifeng Hao. 2024. "Counterfactual Contextual Bandit for Recommendation Under Delayed Feedback." *Neural Computing and Applications* 36 (23): 14599–613. https://doi.org/10.1007/s00521-024-09800-0.

Clément, Benjamin, Didier Roy, Pierre-Yves Oudeyer, and Manuel Lopes. 2015. "Multi-Armed Bandits for Intelligent Tutoring Systems." *Journal of Educational Data Mining* 7 (2): 20–48. https://doi.org/10.5281/zenodo.3554667.

Dudík, Miroslav, John Langford, and Lihong Li. 2011. "Doubly Robust Policy Evaluation and Learning." *Proceedings of the 28th International Conference on Machine Learning*, 1097–104. https://doi.org/10.48550/arXiv.1103.4601.

Ekanadham, Chaitanya, and Yan Karklin. 2017. "T-SKIRT: Online Estimation of Student Proficiency in an Adaptive Learning System." *arXiv Preprint*, ahead of print. https://doi.org/10.48550/arXiv.1702.04282.

Jiang, Nan, and Lihong Li. 2016. "Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning." *Proceedings of the 33rd International Conference on Machine Learning*, 652–61. https://doi.org/10.48550/arXiv.1511.03722.

Joachims, Thorsten, Adith Swaminathan, and Maarten de Rijke. 2018. "Deep Learning with Logged Bandit Feedback." *International Conference on Learning Representations*. https://openreview.net/forum?id=SJaP_-xAb.

Klinkenberg, Sharon, Marthe Straatemeier, and Han L. J. van der Maas. 2011. "Computer Adaptive Practice of Maths Ability Using a New Item Response Model for on the Fly Ability and Difficulty Estimation." *Computers & Education* 57 (2): 1813–24. https://doi.org/10.1016/j.compedu.2011.02.003.

Lan, Andrew S., Christoph Studer, and Richard G. Baraniuk. 2014. "Time-Varying Learning and Content Analytics via Sparse Factor Analysis." *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 452–61. https://doi.org/10.1145/2623330.2623631.

Lee, Morgan, Abubakir Siedahmed, and Neil Heffernan. 2024. "Expert Features for a Student Support Recommendation Contextual Bandit Algorithm." *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 864–70. https://doi.org/10.1145/3636555.3636909.

Li, Lihong, Wei Chu, John Langford, and Xuanhui Wang. 2011. "Unbiased Offline Evaluation of Contextual-Bandit-Based News Article Recommendation Algorithms." *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*, 297–306. https://doi.org/10.1145/1935826.1935878.

Mandel, Travis, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popović. 2014. "Offline Policy Evaluation Across Representations with Applications to Educational Games." *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, 1077–84.

Papoušek, Jan, Radek Pelánek, and Vít Stanislav. 2014. "Adaptive Practice of Facts in Domains with Varied Prior Knowledge." *Proceedings of the 7th International Conference on Educational Data Mining*, 6–13.

Papoušek, Jan, Radek Pelánek, and Vít Stanislav. 2016. "Adaptive Geography Practice Data Set." *Journal of Learning Analytics* 3 (2): 317–21. https://doi.org/10.18608/jla.2016.32.16.

Pelánek, Radek. 2016. "Applications of the Elo Rating System in Adaptive Educational Systems." *Computers & Education* 98: 169–79. https://doi.org/10.1016/j.compedu.2016.03.017.

Pelánek, Radek. 2017. "Bayesian Knowledge Tracing, Logistic Models, and Beyond: An Overview of Learner Modeling Techniques." *User Modeling and User-Adapted Interaction* 27 (3–5): 313–50. https://doi.org/10.1007/s11257-017-9193-2.

Pelánek, Radek, Jan Papoušek, Jiří Řihák, Vít Stanislav, and Juraj Nižnan. 2017. "Elo-Based Learner Modeling for the Adaptive Practice of Facts." *User Modeling and User-Adapted Interaction* 27 (1): 89–118. https://doi.org/10.1007/s11257-016-9185-7.

Piech, Chris, Jonathan Bassen, Jonathan Huang, et al. 2015. "Deep Knowledge Tracing." *Advances in Neural Information Processing Systems 28*.

Precup, Doina, Richard S. Sutton, and Satinder P. Singh. 2000. "Eligibility Traces for Off-Policy Policy Evaluation." *Proceedings of the 17th International Conference on Machine Learning*, 759–66.

Saito, Yuta, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2021. "Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation." *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. https://doi.org/10.48550/arXiv.2008.07146.

Strehl, Alexander L., John Langford, Lihong Li, and Sham M. Kakade. 2010. "Learning from Logged Implicit Exploration Data." *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 2217–25. https://doi.org/10.48550/arXiv.1003.0120.

Swaminathan, Adith, and Thorsten Joachims. 2015. "The Self-Normalized Estimator for Counterfactual Learning." *Advances in Neural Information Processing Systems 28*. https://proceedings.neurips.cc/paper/2015/hash/39027dfad5138c9ca0c474d71db915c3-Abstract.html.

Thomas, Philip S., and Emma Brunskill. 2016. "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning." *Proceedings of the 33rd International Conference on Machine Learning*, 2139–48. https://doi.org/10.48550/arXiv.1604.00923.

Vie, Jill-Jênn, and Hisashi Kashima. 2019. "Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (01): 750–57. https://doi.org/10.1609/aaai.v33i01.3301750.

Wang, Yu-Xiang, Alekh Agarwal, and Miroslav Dudík. 2017. "Optimal and Adaptive Off-Policy Evaluation in Contextual Bandits." *Proceedings of the 34th International Conference on Machine Learning*, 3589–97. https://doi.org/10.48550/arXiv.1612.01205.

Zhan, Ruohan, Vitor Hadad, David A. Hirshberg, and Susan Athey. 2021. "Off-Policy Evaluation via Adaptive Weighting with Data from Contextual Bandits." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2125–35. https://doi.org/10.1145/3447548.3467456.