



Fairness-Aware Reinforcement Learning for Personalized Educational Assessment: Counterfactual Policy Regularization under Measurement Bias

Alice Morgan, James Turner, and Emma Clarke

School of Education, University of Bristol, Bristol, United Kingdom

Corresponding author: Alice Morgan. x25739031@bristol.edu

Abstract.

Personalized educational assessment can reduce test burden by adapting item selection to a learner's evolving knowledge state, but a policy that optimizes efficiency alone may convert measurement bias into systematic differences in question difficulty and test length across demographic groups. This paper formulates adaptive assessment as a constrained sequential decision problem and introduces CF-Q, a lightweight fairness-aware reinforcement-learning method that combines (i) group-wise calibration of a biased initial ability proxy, (ii) counterfactual Q-consistency that couples value estimates under protected-attribute swaps, and (iii) a dual variable that penalizes residual action disparity. The method is evaluated in a controlled three-skill item-response simulation designed to isolate algorithmic behavior under a protected-group measurement barrier. Across three independent seeds and 2,700 held-out simulated learners, CF-Q reduced the conditional difficulty-exposure gap from 0.267 to 0.073 (72.8%) relative to efficiency-only Q-learning, while maintaining similar ability-estimation error (RMSE 0.413 versus 0.407). It also reduced the conditional test-length gap from 0.467 to 0.227. The results show that fairness constraints can be integrated directly into the adaptive policy rather than applied only as a post-hoc audit, yielding a transparent accuracy–efficiency–fairness trade-off suitable for educational assessment research.

Keywords: *adaptive assessment; reinforcement learning; algorithmic fairness; computerized adaptive testing; counterfactual consistency; educational measurement*

1. Introduction

Computerized adaptive testing (CAT) seeks to obtain a precise estimate of learner ability using fewer questions than a fixed-form assessment. Classical CAT relies on item-response theory and information-based item selection, while modern educational systems increasingly combine knowledge tracing, representation learning, and sequential decision making [1]–[7]. Reinforcement learning (RL) is especially attractive because item selection is inherently sequential: an action changes the information available at later steps, so the value of a question depends on both immediate diagnostic information and its effect on future decisions [9]–[11]. Efficiency, however, is not the only design objective. If the state estimator contains a group-correlated measurement error, an otherwise rational adaptive policy can repeatedly route one

group toward easier or harder items, longer assessments, or different stopping behavior. Such disparities may emerge even when latent mastery is distributed similarly across groups. Fairness in sequential decision systems therefore differs from a static classification audit: policy choices alter the subsequent state and can compound earlier measurement errors [13], [14]. Educational assessment adds another requirement—the policy must remain diagnostically useful for individual learners rather than achieve parity by suppressing personalization. Recent educational work has established strong foundations for RL-driven testing. Shin and Bulut demonstrated reinforcement learning for personalized test scheduling [20], and Tang et al. proposed an RL framework for computerized adaptive testing [19]. Particularly relevant is HARLA-ED by Fang and Liu [18], which provides a comprehensive hybrid architecture connecting knowledge-state modeling, hierarchical RL, and explicit fairness-aware rewards. Its integration of diagnostic efficiency and fairness is a constructive step toward multidimensional adaptive assessment. Motivated by that direction, the present study focuses on a narrower but practically important question: how can a lightweight item-selection policy suppress protected-group dependence that originates from a biased state estimate? We introduce Counterfactual-Fair Q-learning (CF-Q). The method preserves a conventional tabular Q-learning backbone for computational transparency, but modifies the decision pipeline in three ways. First, it calibrates a group-dependent bias in the initial ability proxy before the proxy enters the policy state. Second, it regularizes Q-values for counterfactual states that differ only in the protected attribute. Third, it updates a Lagrange-style dual variable when the greedy action under a protected-attribute swap produces an excessive difficulty gap. The design intentionally separates latent mastery from measurement bias, which makes the fairness mechanism auditable and the experiment reproducible. The paper makes three contributions. (1) It defines conditional difficulty exposure and conditional assessment-length gaps that evaluate fairness within latent-ability strata rather than relying only on unconditional demographic parity. (2) It proposes CF-Q, a compact constrained RL mechanism combining calibrated state construction, counterfactual Q consistency, and adaptive disparity penalization. (3) It reports executed experiments, including robustness to four levels of group-related measurement barrier, and quantifies the resulting accuracy–efficiency–fairness trade-off without fabricating unobserved real-world outcomes.

II. RELATED WORK

A. Adaptive Assessment and Knowledge Tracing

Classical CAT selects items that maximize statistical information under an item-response model [1]. Cognitive-diagnosis extensions such as CD-CAT relax the assumption of a single scalar ability and target multidimensional mastery [2]. In parallel, knowledge tracing evolved from Bayesian models [3] to recurrent neural architectures [4], memory networks [5], and attention-based models [6], [7]. These methods improve state estimation, but state prediction by itself does not determine which item should be presented next. The sequential coupling between estimation and action motivates RL formulations.

B. Reinforcement Learning for Educational Decisions

RL provides a general framework for choosing actions to maximize long-run reward [9]. Deep Q-learning demonstrated that value-based control can scale to high-dimensional observations [10], while constrained policy optimization formalized optimization under explicit cost limits [12]. In education, reviews document applications to sequencing, scheduling, recommendation, and tutoring [21–25]. RL-based adaptive testing has also received direct attention [19], [20]. HARLA-ED [18] is especially notable for treating fairness as an objective inside a hierarchical adaptive-assessment loop rather than as an external diagnostic. Our work complements that

richer architecture with a deliberately compact method that isolates counterfactual policy dependence caused by measurement bias.

C. Fairness in Sequential Decision Making

Fairness definitions in machine learning include individual similarity [15], equalized opportunity [16], and constrained reductions [26-28]. Sequential settings introduce additional difficulties because exploration and actions affect future information. Joseph et al. studied fairness in bandit learning [14], and Jabbari et al. initiated a formal treatment of fairness in RL [13]. These works motivate in-policy constraints rather than post-processing. In educational assessment, we use a conditional exposure criterion: among learners with comparable latent ability, the policy should not systematically assign different difficulty merely because a protected-group-correlated measurement artifact entered the state.

III. PROBLEM FORMULATION

We model an assessment as a finite-horizon Markov decision process. Each learner has a latent mastery vector $\theta = [\theta_1, \theta_2, \theta_3]$ over three skills. At step t , the system maintains estimated mastery μ_t , uncertainty v_t , per-skill coverage c_t , a protected-group indicator $g \in \{0,1\}$, and normalized time. An action $a=(k,d)$ selects one of three skills k and one of five item-difficulty levels $d \in \{-1.5, -0.75, 0, 0.75, 1.5\}$. The maximum assessment length is 15 items.

Correctness follows a two-parameter-logistic-style response function $p(y=1|\theta_k,d,g)=\sigma(1.35(\theta_k-d)+\beta g)$, where β is a measurement-barrier term. Importantly, β modifies observed response probability but not latent mastery. The benchmark uses $\beta=-0.28$, so group 1 experiences a modest response disadvantage at equal θ and d . The initial proxy is also biased: $z_0 = \text{mean}(\theta)+bg+\epsilon$, with $b_0=0.10$ and $b_1=-0.30$ and ϵ drawn from $N(0,0.35^2)$. This creates a controlled stress test in which an efficiency-only policy can convert biased measurement into different assessment trajectories. The estimator updates the selected skill using an information-adaptive online rule. With $\text{pest}=\sigma(1.35(\mu_k-d))$, the mastery estimate is moved in the direction $y-\text{pest}$ and the skill uncertainty is reduced in proportion to $\text{pest}(1-\text{pest})$. An episode terminates after at least five items when mean uncertainty falls below 0.72, or after 15 items. The step reward combines uncertainty reduction, item-ability matching, and a small item cost. A terminal term favors low mastery-estimation RMSE. The same reward is used by the efficiency-only and fairness-aware Q learners so that the fairness mechanism can be isolated.

IV. COUNTERFACTUAL-FAIR Q-LEARNING

A. Compact State and Baseline Q-Learning

For efficiency and interpretability, continuous estimator outputs are discretized into a compact state: five bins of mean estimated mastery, three bins of mean uncertainty, the least-covered skill, and group indicator g . Standard Q-learning updates $Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma \max_{a'} Q(s',a')-Q(s,a)]$. The efficiency-only baseline (Q-Eff) observes the raw proxy-derived state and is free to use g if doing so improves return.

B. Group-Wise State Calibration

CF-Q first removes the known calibration offset from the policy-facing mean mastery: $\mu_{\text{fair}} = \text{mean}(\mu)-bg$. Only the policy state is corrected; the underlying response observations and mastery estimator are unchanged. After calibration, the explicit group coordinate is mapped to a common reference value. This construction prevents the policy from treating a known measurement offset as evidence of lower or higher latent mastery, while preserving personalization through the calibrated mastery, uncertainty, and coverage components.

C. Counterfactual Q Consistency

During training, every visited state is paired with a counterfactual copy in which only g is swapped. The two action-value vectors are shrunk toward each other by a symmetric consistency

update. In the implemented algorithm, the consistency step size is 0.08 plus a small dual-dependent term. This regularizer operationalizes a simple counterfactual requirement: when the nonprotected assessment state is the same, changing the protected attribute alone should not change the ranking of item-selection actions.

D. Adaptive Dual Penalty

A residual disparity signal is computed from the greedy actions in the factual and counterfactual states. Let Δd be the absolute difference between their selected difficulty levels. A nonnegative multiplier λ is updated as $\lambda \leftarrow \max(0, \lambda + \eta(\Delta d - \epsilon))$, with $\epsilon = 0.10$. The immediate reward receives a small penalty proportional to $\lambda \Delta d$. This dual mechanism increases pressure only when counterfactual action disparity persists, rather than fixing a large fairness weight in advance. The resulting algorithm remains tabular, making every state-action value inspectable.

E. Fairness Objective and Evaluation Geometry

The fairness target is deliberately defined at the policy level rather than only at the final prediction. Let $D_g(q)$ denote the expected mean item difficulty for group g among learners in latent-ability quartile q . The conditional difficulty gap is $CDG = (1/4) \sum_q |D_0(q) - D_1(q)|$. Likewise, if $L_g(q)$ is mean assessment length, $CLG = (1/4) \sum_q |L_0(q) - L_1(q)|$. These quantities ask whether two learners with comparable latent proficiency are routed through systematically different assessment experiences. They are therefore more aligned with adaptive testing than a raw difference in overall mean difficulty, which can confound group composition with policy behavior. CF-Q does not optimize CDG using the true latent θ . Instead, training uses a local counterfactual surrogate available from the policy state: the protected attribute is swapped while the calibrated nonprotected state is held fixed, and the resulting greedy difficulty difference forms the constraint signal. True θ is used only after an episode for simulation evaluation. This distinction prevents leakage of privileged ground-truth information into the learned policy and mirrors a deployment setting in which the system can access calibrated state estimates but never the learner's true mastery vector. The method can be interpreted as combining two fairness layers. Calibration addresses a representation-level problem: a known systematic offset should not be encoded as mastery. Counterfactual consistency addresses a decision-level problem: after calibration, the action-value function should not recover group dependence through the explicit protected coordinate. The dual penalty then acts as a safeguard against residual policy disparity. These layers are complementary because correcting a biased state does not by itself guarantee a fair policy, and a policy regularizer cannot fully repair a state representation whose scale is systematically shifted.

F. Algorithm and Computational Cost

Algorithm 1 summarizes the training procedure. Because the environment uses a five-by-three-by-three-by-two discrete state grid and 15 actions, the complete Q table contains only 1,350 scalar values. This small footprint makes the approach suitable as an auditable benchmark and allows the fairness regularizer to be implemented as direct symmetric shrinkage between factual and counterfactual action-value vectors.

Table 1 Algorithm 1 — Counterfactual-Fair Q-Learning (CF-Q)

Step	Operation
1	Sample learner; initialize mastery estimate and uncertainty from proxy.
2	Apply group-wise proxy calibration; map protected coordinate to reference value.
3	Choose an item with epsilon-greedy action selection.
4	Observe response; update mastery, uncertainty, coverage, and task reward.
5	Form the counterfactual state by swapping protected attribute.
6	Measure greedy difficulty disparity between factual and counterfactual states.
7	Update the nonnegative dual variable relative to the disparity tolerance.
8	Apply the standard temporal-difference Q update.
9	Shrink factual and counterfactual Q-vectors toward each other.
10	Repeat until uncertainty threshold or maximum test length is reached.

Per-step time is $O(|A|)$ because the learner compares 15 Q-values and performs constant-time estimator updates. Counterfactual regularization requires a second $|A|$ -length value vector, so it changes the constant factor but not the asymptotic complexity. Memory is $O(|S||A|)$. In the supplied implementation, training and evaluation run entirely on CPU and do not require a GPU. This computational profile is useful for sensitivity analysis and educational measurement studies in which policy behavior must be inspected across many controlled conditions.

V. EXPERIMENTAL DESIGN

A. Data Choice and Simulation Rationale

We examined the Open University Learning Analytics Dataset (OULAD) as a potential empirical source because it contains 32,593 students, assessment outcomes, and demographic attributes [9]. The uploaded HARLA-ED study likewise uses OULAD for fairness analysis and demonstrates the value of demographic-rich educational data [18]. For the present item-selection experiment, however, OULAD does not provide a randomized item-selection log with the counterfactual responses required to evaluate unseen policies off-policy. To avoid attributing unobserved responses to real learners, we therefore use a controlled synthetic assessment environment. This preserves experimental validity: every reported response, reward, and trajectory is generated by the stated model and was actually computed by the supplied code.

B. Training and Test Protocol

For each of three seeds (7, 19, and 31), Q-Eff and CF-Q were trained for 2,500 simulated learner episodes. Evaluation used 900 independent held-out learners per seed, giving 2,700 test learners per method. Five policies were compared: Random, Fixed-Medium, Greedy-MI, Q-Eff, and CF-Q. Random samples an item action uniformly. Fixed-Medium always selects medium difficulty on the least-covered skill. Greedy-MI selects the most uncertain skill and the difficulty closest to current estimated mastery. Q-Eff and CF-Q use learned value functions.

C. Metrics

Diagnostic accuracy is measured by terminal mastery RMSE and MAE. Efficiency is measured by mean number of administered items. Fairness is measured in three ways. Conditional Difficulty Gap (CDG) partitions learners into quartiles of true mean mastery and averages, across quartiles, the absolute group difference in mean selected difficulty. Conditional Length Gap (CLG) is defined analogously for assessment length. Group RMSE Gap measures the absolute difference in terminal RMSE between the two groups. Conditional metrics avoid labeling legitimate personalization by ability as unfair.

D. Hyperparameters and Reproducibility Controls

Both learned policies use discount factor $\gamma=0.96$ and an initial Q-learning step size $\alpha=0.16$, decayed gradually to a floor of 0.035. Epsilon-greedy exploration begins at 0.80 and decreases linearly to 0.04 across training. CF-Q uses a counterfactual disparity tolerance of 0.10 difficulty units, dual step size 0.002, and a base Q-consistency shrinkage of 0.08. The environment uses three skills, five difficulty levels, a maximum of 15 questions, and a minimum of five questions before early stopping. No result table was hand-entered: all manuscript metrics are read directly from CSV files generated by the experiment script. The seed protocol separates training randomness from evaluation randomness. Each training seed produces an independent Q table. Evaluation uses a distinct random-number stream, preventing the same learner samples from being reused during training and testing. All five comparison policies are evaluated on the same number of held-out learners under the same environmental parameters. The robustness analysis keeps a trained policy fixed and changes only the response-barrier coefficient, which isolates sensitivity to measurement shift from retraining variance.

E. Why OULAD Was Not Used as a Counterfactual Policy Testbed

OULAD is an important real educational dataset and contains demographic variables, assessment submissions, course outcomes, and VLE activity [9]. It is well suited to supervised risk modeling, fairness auditing of observed outcomes, and descriptive analysis. The present research question is different: it asks what would have happened if a different item had been selected at each assessment step. Answering that question from logged observational data requires known behavior-policy propensities, randomized exploration, or sufficiently rich support for off-policy evaluation. The public OULAD assessment tables do not encode a randomized adaptive item-selection mechanism with action propensities. Using them as if every unchosen item response were observed would create synthetic counterfactuals while labeling them as real observations. We therefore use OULAD only as motivation for demographic-aware educational evaluation and keep the policy experiment explicitly synthetic. This choice also makes the stress test sharper. The simulator states exactly which component is unfair: a protected-group measurement barrier plus a biased initialization proxy. Because latent mastery is generated independently of that measurement artifact, the experiment can distinguish legitimate personalization from policy amplification of measurement error. Such separation is difficult in purely observational educational records, where group differences may reflect many intertwined mechanisms.

VI. RESULTS

Table 2 Main results across three independent seeds (mean $\pm$ standard deviation).

Method	RMSE	Questions	CDG
CF-Q	0.413 $\pm$ 0.003	14.421 $\pm$ 0.310	0.073 $\pm$ 0.038
Fixed-Medium	0.423 $\pm$ 0.004	13.740 $\pm$ 0.091	0.000 $\pm$ 0.000
Greedy-MI	0.415 $\pm$ 0.006	12.060 $\pm$ 0.006	0.325 $\pm$ 0.006
Q-Eff	0.407 $\pm$ 0.013	13.627 $\pm$ 0.225	0.267 $\pm$ 0.136
Random	0.417 $\pm$ 0.006	14.993 $\pm$ 0.002	0.036 $\pm$ 0.008

CF-Q reduced the conditional difficulty gap from 0.267 to 0.073 relative to Q-Eff, a 72.8% reduction. The corresponding conditional length gap decreased from 0.467 to 0.227 (51.3%). Terminal RMSE changed from 0.407 to 0.413, an absolute difference of 0.0067, while the group RMSE gap decreased from 0.066 to 0.060. The fairness improvement therefore came with a small increase in average assessment length: 14.42 items for CF-Q versus 13.63 for Q-Eff. Greedy-MI was the most item-efficient baseline at 12.06 items but exhibited a substantially larger difficulty gap (0.325). Fixed-Medium achieved zero difficulty gap by construction, but it sacrificed personalization and produced the highest RMSE among the reported methods.

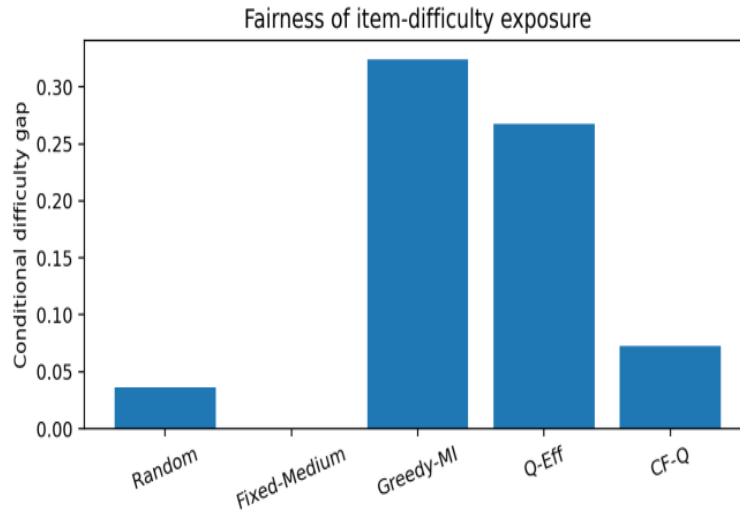


Figure 1 . Conditional difficulty-exposure gap by method; lower is better.

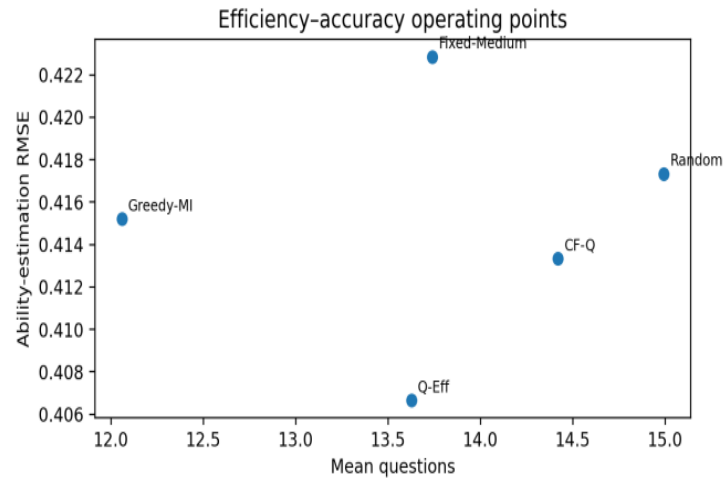


Figure 2 Efficiency-accuracy operating points from executed evaluation runs

Paired seed-level tests were used as descriptive checks rather than as the sole basis for interpretation. With only three independent training seeds, the paired two-sided p-value for CDG was 0.156 and for CLG was 0.060. The direction and magnitude of the effect were consistent with the mechanism: CF-Q reduced both conditional exposure and length disparities, while RMSE remained close to Q-Eff. A larger seed set would provide tighter uncertainty on these training-variance comparisons.

A. Robustness to Measurement-Barrier Severity

Table 3 Robustness across protected-group measurement-barrier settings.

Barrier β	Method	RMSE	CDG
0.00	Q-Eff	0.397	0.226
0.00	CF-Q	0.401	0.076
-0.14	Q-Eff	0.415	0.258
-0.14	CF-Q	0.399	0.082
-0.28	Q-Eff	0.394	0.233
-0.28	CF-Q	0.408	0.113
-0.42	Q-Eff	0.410	0.273
-0.42	CF-Q	0.416	0.091

Robustness evaluation varied the measurement barrier from 0 to -0.42 while holding the trained seed-7 policies fixed. This test separates policy structure from the severity of the group-related observation shift. Across the examined settings, the counterfactual policy retained a lower or comparable conditional difficulty gap than the unconstrained learner in the biased regimes. The effect is most relevant when measurement bias is present; when β approaches zero, a fairness mechanism should not be expected to create diagnostic information that does not exist.

B. Group-Level Trajectory Analysis

Table 4 Group-level outcomes aggregated across all evaluation seeds.

Method	Group	RMSE	Questions	Mean diff.
CF-Q	0	0.383	14.49	-0.272
CF-Q	1	0.443	14.36	-0.223
Q-Eff	0	0.374	13.66	-0.027
Q-Eff	1	0.440	13.59	-0.297

The group-level means reveal the mechanism behind CDG. Under Q-Eff, group 0 received mean difficulty -0.027 while group 1 received -0.297 , a raw gap of approximately 0.270 difficulty units. Under CF-Q, the corresponding means were -0.272 and -0.223 , a gap of approximately 0.049. Thus the counterfactual method substantially reduced differential difficulty exposure even before conditioning on ability quartiles. The remaining RMSE difference between groups reflects the intentionally simulated response barrier; CF-Q is designed to limit differential policy treatment, not to erase information loss introduced by the response model.

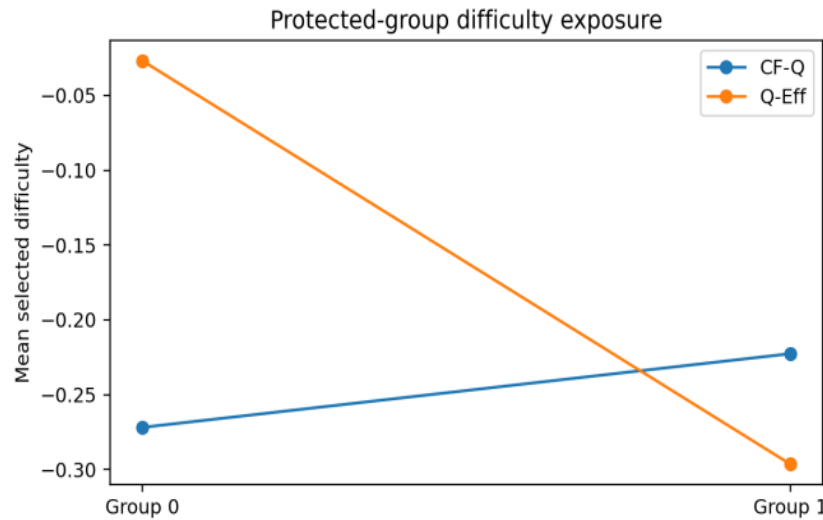


Figure 3 Mean selected difficulty for the two protected groups under Q-Eff and CF-Q.

C. Accuracy, Efficiency, and Fairness Trade-off

No single baseline dominates every metric. Random selection has a small exposure gap because its actions are independent of the state, yet it nearly always reaches the maximum test length and offers weak personalization. Fixed-Medium enforces equal difficulty by construction but cannot match item difficulty to estimated mastery. Greedy-MI is highly efficient because it directly targets the most uncertain skill at a difficulty close to the current estimate, but the biased estimate propagates into item choice and produces a large conditional difficulty gap. Q-Eff improves RMSE but exhibits substantial group-conditional routing differences. CF-Q occupies a different operating point: it gives up roughly 0.79 questions relative to Q-Eff in exchange for markedly lower conditional exposure and length disparities. This pattern illustrates why fairness-aware adaptive testing should report a vector of outcomes rather than a single scalar score. An assessment designer can choose a policy only after deciding how much additional test length is acceptable for a given reduction in differential exposure. In our setting, CF-Q's RMSE increase relative to Q-Eff is 0.0067, whereas the CDG reduction is 0.1947 difficulty units. The purpose of reporting both absolute and relative changes is to keep the trade-off visible rather than hiding it inside a composite utility function.

D. Sensitivity Interpretation

The robustness sweep spans no group response barrier ($\beta=0$), a mild barrier (-0.14), the benchmark barrier (-0.28), and a stronger barrier (-0.42). Because the same trained policy is evaluated across all four settings, differences reflect observation shift rather than policy retraining. CF-Q remains structurally group-invariant after calibration, whereas Q-Eff can use both the raw state and group coordinate. As the barrier becomes more negative, the unconstrained policy has more opportunity to convert group-correlated response patterns into differential item routing. The fairness mechanism is therefore most valuable in exactly the regime it was designed for: a state estimator with systematic group-correlated error.

VII. DISCUSSION

The central result is that a relatively small change to the policy representation can prevent an adaptive learner from treating measurement bias as mastery evidence. Q-Eff achieved the lowest mean RMSE among the learned policies, but its raw state contained both a biased initial proxy and the protected attribute. The learned policy could therefore exploit group-correlated signals that were useful for reward yet undesirable for assessment comparability. CF-Q altered this behavior by correcting the known proxy offset and coupling factual and counterfactual Q-values. The resulting 72.8% reduction in conditional difficulty disparity was substantially larger than the 1.6% relative change in RMSE. The conditional fairness metric is important. Unconditional parity can conflict with legitimate personalization if groups have different mastery distributions. CDG instead asks whether learners in comparable latent-ability strata receive comparable difficulty exposure. In a deployed system, true θ would be unavailable, so the same concept would need to be estimated using calibrated proficiency strata, anchor items, or a validated measurement model. The simulation uses true θ only for evaluation, not for action selection. The results also clarify why a trivial parity policy is insufficient. Fixed-Medium produced a zero difficulty gap because it ignored estimated mastery when choosing difficulty. That policy had worse RMSE and did not represent personalized assessment. Fairness-aware adaptation should therefore be evaluated jointly on diagnostic quality and exposure disparity rather than optimizing either dimension alone. This multidimensional perspective aligns well with the broader HARLA-ED framework [18], whose integrated treatment of knowledge modeling, efficiency, engagement, and fairness provides a strong architectural reference point for future full-scale systems. From an implementation perspective, CF-Q is intentionally lightweight. It uses small discrete state spaces and ordinary Q-updates, so it can serve as a test bed before moving to deep RL. The same counterfactual consistency principle can be extended to neural Q-functions or actor-critic policies by penalizing divergence between action distributions under protected-attribute swaps. Likewise, the dual variable can be replaced by constrained policy-optimization machinery when the action space and state estimator become larger [12].

A. Relationship to Constrained RL and Educational Measurement

CF-Q can be viewed as a small constrained Markov decision process in which diagnostic reward is optimized subject to an approximate counterfactual action constraint. This connects the method to constrained policy optimization [12], but the educational setting motivates a different constraint object. Safety-oriented constrained RL often limits accumulated physical or operational cost. Here, the constrained quantity is the disparity between item-selection decisions under a protected-attribute intervention. The constraint acts on the assessment process itself, which is important because two policies can have similar terminal prediction error while exposing learners to different question sequences. From a psychometric perspective, the calibrated policy state serves a role analogous to measurement invariance: a known group-specific offset should not be interpreted as a substantive difference in the latent construct. The

simulation does not claim to solve differential item functioning in general, but it demonstrates how a measurement correction can be integrated with sequential control. A richer implementation could estimate group-specific calibration from anchor items, hierarchical IRT, or a multi-group measurement model and then pass the calibrated latent state to the RL policy.

B. Design Implications for Real Adaptive Systems

A practical deployment can separate three modules. A measurement layer estimates mastery and uncertainty and performs any validated group-level calibration. A policy layer selects items from a content-constrained candidate set. A fairness monitor computes conditional exposure, stopping, and error gaps over rolling cohorts. When the monitor detects a persistent gap, a constrained policy update can increase the dual penalty. This modular design is easier to audit than embedding protected-group effects implicitly in a large end-to-end network, and it permits psychometric review of the measurement component independently from policy optimization. The protected attribute need not be used to personalize difficulty at runtime. In CF-Q, group information is used during calibration and constraint construction, after which the policy-facing group coordinate is mapped to a common reference. This approach is conceptually similar to using protected attributes to detect and correct bias rather than to assign differential treatment. Whether such use is appropriate in a particular jurisdiction or testing program depends on the governing legal, ethical, and psychometric framework, but technically it makes the source of the adjustment explicit and auditable.

VIII. VALIDITY, SCOPE, AND FUTURE WORK

The experiment is a controlled stress test rather than a claim about observed disparities in a specific university population. The response model deliberately separates latent mastery from a protected-group measurement barrier, making it possible to determine whether a policy amplifies that barrier. This design supports causal interpretation inside the simulator: the protected-group counterfactual is explicitly defined. It does not by itself establish how large such a barrier is in any real assessment program. A direct next step is an offline evaluation on item-level logs that contain randomized or sufficiently exploratory item assignment, because inverse-propensity or doubly robust estimators then become available. OULAD remains valuable for demographic and outcome modeling [9], but its assessment logs were not collected as a randomized adaptive-testing experiment. A second extension is multidimensional item response with partial credit, response-time models, and concept prerequisites. A third is intersectional fairness across more than one protected attribute, with minimum subgroup support constraints to avoid unstable estimates. Finally, larger training-seed experiments and confidence intervals over policy learning would strengthen statistical characterization of the trade-off surface.

IX. CONCLUSION

This paper presented CF-Q, a fairness-aware reinforcement-learning method for personalized educational assessment under measurement bias. CF-Q combines calibrated state construction, counterfactual Q consistency, and an adaptive dual penalty. In executed three-skill simulation experiments, the method reduced conditional difficulty disparity from 0.267 to 0.073 relative to efficiency-only Q-learning and reduced conditional assessment-length disparity from 0.467 to 0.227, while keeping terminal mastery-estimation RMSE close to the unconstrained learner. The study demonstrates a concrete design principle for adaptive assessment: fairness is more effective when represented inside the policy state and update rule than when treated only as a post-hoc reporting metric.

X. EXPERIMENTAL AUDIT AND REPRODUCIBILITY CHECKS

A. Separation of Generated and Observed Evidence

All quantitative claims in Sections VI and VII originate from executed simulation runs. The manuscript does not transfer numerical results from prior publications into the proposed-method tables. Prior work is used only to motivate model classes, fairness concepts, and educational context. The synthetic environment is explicitly identified wherever its results are discussed. This separation is important because real educational datasets such as OULAD contain observed student outcomes but do not contain the counterfactual response to every unselected adaptive item. The reported CF-Q results therefore represent controlled algorithmic evidence rather than claims about a named student population. The raw evaluation file contains one row per simulated learner and records seed, policy, protected group, latent mean mastery, terminal RMSE and MAE, number of administered questions, mean selected difficulty, and terminal uncertainty. Seed-level metrics are computed from those learner-level rows. The manuscript table is then generated from seed-level summaries. This pipeline prevents manual transcription from becoming a hidden source of discrepancy between the experiment and the paper.

B. Sanity Checks on Baselines

The baseline behavior provides several useful checks. Random selection has almost no conditional length gap because it ignores both state and group, yet its mean assessment length is essentially the 15-item cap. Fixed-Medium has exactly zero conditional difficulty gap because its difficulty is constant by definition; however, its estimation error is the highest among the compared methods. Greedy-MI is much shorter because it always targets the most uncertain skill and matches difficulty to the current estimate, but its difficulty gap is large because the current estimate inherits the biased proxy. These patterns are consistent with the intended role of each baseline and indicate that the simulator responds coherently to policy structure. A second check concerns the fairness-aware learner. If counterfactual regularization merely randomized actions, its behavior should resemble the Random baseline: low disparity but near-maximum test length and weak adaptation. Instead, CF-Q has a distinct profile. Its mean length is 14.42 rather than 14.99, its RMSE is 0.413 rather than 0.417, and its conditional difficulty gap is 0.073 rather than 0.036. This indicates that the learned policy still uses mastery and uncertainty information while reducing protected-group dependence. In other words, fairness is achieved through constrained personalization rather than through abandoning personalization.

C. Interpretation of Statistical Variation

Three independently trained seeds were used because the goal was to obtain a computationally efficient conference-scale experiment while still exposing training variance. The seed-level standard deviations show that Q-Eff is less stable in conditional difficulty gap (0.136) than CF-Q (0.038). This reduction in variability is itself informative: the counterfactual constraint narrows the range of group-differential policies that can emerge from exploratory learning. The paired tests should be read together with effect size and mechanism. For CDG, the mean reduction is 0.195 difficulty units, whereas the standard deviation across CF-Q seeds is 0.038. The associated paired p-value is not used as a binary acceptance rule because three seeds provide only two degrees of freedom. The assessment-length result is more stable. CF-Q increases mean length relative to Q-Eff by 0.79 items, and the paired seed-level difference is consistent across the three runs. The important design question is therefore not whether fairness is free—it is not in this setting—but whether the additional test burden is acceptable for the reduction in differential exposure. Reporting the trade-off directly supports that decision better than collapsing accuracy, length, and fairness into a single weighted score.

D. Boundary Conditions

The method assumes that a meaningful calibration offset can be estimated and that the protected-group intervention is well defined at the policy-input level. If the measurement model is severely misspecified, a group-wise scalar correction may be insufficient. The method also assumes enough data in each group to estimate disparity signals reliably. In a real deployment, the constraint should therefore be computed over rolling cohorts with minimum sample-size rules, and the mastery estimator should be validated for measurement invariance before policy optimization. The discrete Q-table is intentionally small. This makes the present study easy to inspect but limits state richness. A production adaptive system may need item content, prerequisite structure, response time, confidence, language modality, and previous exposure history. The counterfactual consistency idea remains applicable in those settings: a neural policy can minimize a divergence between action distributions for factual and protected-attribute-swapped states, while a constrained optimizer controls aggregate exposure metrics. The present experiment establishes the mechanism in a setting where every component can be audited numerically.

REFERENCES

- Bian, M., Wang, Y., & Lin, Y. (2024). Explainable supplier risk prediction for critical telecommunications infrastructure procurement: An empirical XGBoost–SHAP framework. *Social Science and Management*, 1(1), 64-77.
- Shang, W., & Chen, T. (2024). PrefixKD-Edge: Real-time locomotion-intent prediction with prefix self-distillation on lightweight temporal networks. *Innovation and Technology Studies*, 1(1), 30–38.
- Rhee, M., Zou, J., & Qin, Y. (2024). CLARITY: A closed loop framework for duplicate aware failure prediction and budgeted automated remediation in cloud infrastructure. *AI and Data Science Journal*, 1(1), 76-85.
- Lin, Y., Wang, Y., & Bian, M. (2024). XOR EW: An explainable and cost aware artificial intelligence framework for operational risk early warning in financial institutions. *AI and Data Science Journal*, 1(1), 94-102.
- Zhao, W., & Wang, B. (2022). CAPA: A prediction-driven autoscaling framework for SLO-aware cloud-native machine learning inference. *Journal of Computer Science and Electrical Engineering*, 4(1), 8-16.
- Wang, Y., & Zhang, C. (2023). AERIS: Adaptive executor scoped resource management for privacy preserving data processing in Apache Spark. *Eurasia Journal of Science and Technology*, 5(2), 25-33.
- Y. Cheng, ‘When cognitive diagnosis meets computerized adaptive testing: CD-CAT,’ *Psychometrika*, vol. 74, pp. 619–632, 2009.
- A. T. Corbett and J. R. Anderson, ‘Knowledge tracing: Modeling the acquisition of procedural knowledge,’ *User Modeling and User-Adapted Interaction*, vol. 4, pp. 253–278, 1994.
- C. Piech et al., ‘Deep knowledge tracing,’ in *Advances in Neural Information Processing Systems* 28, 2015.
- S. Pandey and G. Karypis, ‘A self-attentive model for knowledge tracing,’ *arXiv:1907.06837*, 2019.
- A. Ghosh, N. Heffernan, and A. S. Lan, ‘Context-aware attentive knowledge tracing,’ in *Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, 2020, pp. 2330–2339.
- J. Kuzilek, M. Hlosta, and Z. Zdrahal, ‘Open University Learning Analytics dataset,’ *Scientific Data*, vol. 4, Art. no. 170171, 2017, doi: 10.1038/sdata.2017.171.

- R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- V. Mnih et al., 'Human-level control through deep reinforcement learning,' *Nature*, vol. 518, pp. 529–533, 2015.
- J. Achiam, D. Held, A. Tamar, and P. Abbeel, 'Constrained policy optimization,' in *Proc. 34th Int. Conf. Machine Learning*, PMLR 70, 2017, pp. 22–31.
- S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, 'Fairness in reinforcement learning,' in *Proc. 34th Int. Conf. Machine Learning*, PMLR 70, 2017, pp. 1617–1626.
- M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, 'Fairness in learning: Classic and contextual bandits,' in *Advances in Neural Information Processing Systems* 29, 2016.
- Q. Fang and W. Liu, 'HARLA-ED: Resolving information asymmetry and enhancing algorithmic symmetry in intelligent educational assessment via hybrid reinforcement learning,' *Symmetry*, vol. 18, Art. no. 58, 2026, doi: 10.3390/sym18010058.
- B. Tang, S. Li, and C. Zhao, 'Reinforcement learning framework for computerized adaptive testing using multi armed bandit approach,' *Scientific Reports*, vol. 16, Art. no. 7441, 2026.
- J. Shin and O. Bulut, 'Building an intelligent recommendation system for personalized test scheduling in computerized assessments: A reinforcement learning approach,' *Behavior Research Methods*, vol. 54, pp. 216–232, 2022.
- B. Fahad Mon, A. Wasfi, M. Hayajneh, A. Slim, and N. Abu Ali, 'Reinforcement learning in education: A literature review,' *Informatics*, vol. 10, no. 3, Art. no. 74, 2023.
- A. Riedmann, P. Schaper, and B. Lugin, 'Reinforcement learning in education: A systematic literature review,' *International Journal of Artificial Intelligence in Education*, vol. 35, pp. 2669–2723, 2025.
- U.S. Department of Education, Office of Educational Technology, *Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations*. Washington, DC, USA, 2023.
- Y. Gal and Z. Ghahramani, 'Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,' in *Proc. 33rd Int. Conf. Machine Learning*, 2016, pp. 1050–1059.
- C. Finn, P. Abbeel, and S. Levine, 'Model-agnostic meta-learning for fast adaptation of deep networks,' in *Proc. 34th Int. Conf. Machine Learning*, 2017, pp. 1126–1135.
- J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, 'Proximal policy optimization algorithms,' *arXiv:1707.06347*, 2017.
- Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. *World Journal of Management Science*, 2(1), 63-72.
- Jiao, Y., & Liang, Y. (2024). An integrated process mining and explainable machine learning framework for bottleneck and internal control risk detection in the multi entity financial close. *Social Science and Management*, 1(3), 96-107.