INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Uncertainty-Aware Reinforcement Learning for Adaptive Educational Assessment under Item-Calibration Uncertainty

Caroline Mercer* and Ethan Brooks

University of Washington, Seattle, WA, USA

Corresponding author. carolinemrr2@uw.edu

Abstract

Adaptive educational assessment must make sequential item-selection decisions while both learner proficiency and item parameters are uncertain. Conventional maximum-information rules typically treat calibrated item difficulty as fixed, whereas reinforcement-learning policies often encode uncertainty only implicitly in a latent state. This paper proposes UA-DoubleQ, an uncertainty-aware reinforcement-learning framework that exposes posterior learner uncertainty and item-calibration uncertainty directly to the decision process. A Bayesian ability belief is maintained on a discretized latent-proficiency grid. At each step, a Double Q-learning controller selects a risk coefficient that modulates a robust Fisher-information score; the same controller can terminate the assessment when continued measurement is not worth its cost. This design separates uncertainty estimation from risk scheduling and allows the policy to become more conservative when item-parameter uncertainty is consequential. We evaluate the method in an ASSISTments-calibrated semi-synthetic benchmark, required because historical logs do not provide counterfactual responses to unadministered items. Across three independent seeds with 500 test examinees per seed, UA-DoubleQ used 14.02±1.25 questions, achieved ability RMSE 0.444±0.042, 92.1±1.1% empirical 95% interval coverage, and selected items with mean difficulty standard error 0.237. Relative to a vanilla Double-Q policy without uncertainty in the state, UA-DoubleQ shortened assessments by 17.0%, improved coverage by 1.8 percentage points, and reduced selected-item calibration uncertainty by 4.8%, with an explicitly quantified RMSE trade-off. Relative to point-estimate maximum Fisher information, it reduced RMSE by 14.9% and improved coverage by 4.3 percentage points. The results support uncertainty-conditioned risk scheduling as a practical middle ground between aggressive information maximization and conservative uncertainty-aware selection.

Keywords: *adaptive assessment, reinforcement learning, uncertainty quantification, computerized adaptive testing, item response theory, Double Q-learning*

1. Introduction

Computerized adaptive assessment reduces measurement burden by choosing the next item as a function of the evidence already observed. The classical formulation is elegant: estimate a learner's latent ability, select an item that maximizes information at that estimate, update the estimate, and stop when precision is adequate [1]–[3], [29], [30]. In practice, however, two forms of uncertainty interact. The learner state is uncertain early in a test, and the item parameters used by the selector are themselves estimates obtained from finite calibration samples. Treating either source as fixed can make an adaptive policy overconfident, particularly when the item bank contains long-tailed exposure counts or newly introduced items. Recent intelligent-assessment work has substantially broadened the decision problem beyond static item information. Fang and Liu's HARLA-ED framework is especially valuable in showing how deep knowledge modeling, hierarchical reinforcement learning, and explicit uncertainty estimates can be integrated into a closed assessment loop [12]. Their use of uncertainty to prioritize poorly known knowledge components motivates a complementary question studied here: when the assessment model itself is uncertain about item calibration, how should a sequential decision policy adjust the aggressiveness of information seeking? Rather than constructing another large hierarchical architecture, we isolate this mechanism so that its effect can be measured transparently. The distinction matters because uncertainty is not synonymous with randomness. Posterior uncertainty about a learner can justify exploratory questions, while epistemic uncertainty in an item parameter can make an apparently informative question unreliable. A policy that reacts to only one side of this interaction may either terminate too early or repeatedly choose high-information items whose calibration error dominates the remaining learner uncertainty. Bayesian treatments of learner state [5], [6], uncertainty-aware deep learning [17], and exploration-oriented reinforcement learning [21], [24], [25] provide relevant ingredients, but they do not directly specify how item-calibration uncertainty should modulate item selection in adaptive testing. This paper proposes UA-DoubleQ, a compact uncertainty-aware reinforcement-learning method with three design choices. First, learner ability is represented by a posterior distribution rather than a point estimate. Second, each item carries both estimated psychometric parameters and a difficulty standard error. Third, the reinforcement-learning action is not the item identity itself; it is a risk coefficient that reshapes a robust information criterion. This factorization keeps the action space small while allowing the learned controller to choose between aggressive and conservative item selection as posterior uncertainty changes. The empirical study uses a controlled semi-synthetic protocol because logged educational datasets reveal responses only to administered items; direct evaluation of a new item-selection policy would otherwise require counterfactual responses that are not observed. The response environment is calibrated to the public ASSISTments 2009–2010 Skill Builder corpus [4], [28], while all policy comparisons are generated from the same 2PL simulator with fixed random seeds. This keeps the source of every reported number explicit. The study compares random selection, point-estimate maximum Fisher information (MFI), robust MFI, Thompson sampling, vanilla Double Q-learning, and UA-DoubleQ. The contributions are: (1) an uncertainty-conditioned risk-scheduling formulation that turns item-calibration uncertainty into an explicit RL control variable; (2) a Bayesian belief update that propagates item-difficulty uncertainty into learner-state uncertainty; (3) an adaptive stopping action learned jointly with item-selection risk; and (4) a controlled benchmark that reports efficiency, ability error, interval coverage, and selected-item calibration uncertainty together rather than collapsing them into a single score.

II. RELATED WORK

A. Computerized Adaptive Testing and Item Response Theory

IRT provides the classical statistical foundation for adaptive testing [1]. In the 2PL model, a learner's probability of a correct response is determined by latent ability, item difficulty, and item discrimination. CAT commonly selects the item with maximum Fisher information near the current ability estimate. Chang and Ying showed that such local information can be inefficient early in a test when the ability estimate is

inaccurate, motivating global-information approaches [3]. This observation is directly relevant to reinforcement-learning formulations because early-test uncertainty changes the value of exploration.

Traditional CAT theory typically assumes that the calibrated item bank is known at administration time. In operational systems this is an approximation: item parameters are estimated from finite and uneven samples, and newly deployed items may have substantially larger standard errors. Our method retains the interpretability of IRT but allows the sequential selector to explicitly react to item-level uncertainty.

B. Knowledge Tracing and Cognitive Diagnosis

Bayesian Knowledge Tracing models skill mastery as a latent state updated from binary responses [5], while individualized BKT accounts for learner-specific variation [6]. Deep Knowledge Tracing introduced recurrent neural networks for richer temporal representations [7]. Subsequent architectures include dynamic key-value memory networks [8], self-attentive knowledge tracing [9], context-aware attentive knowledge tracing [10], and SAINT+ with temporal features [11]. Neural cognitive diagnosis [13] and Deep-IRT [14] further connect expressive representation learning with interpretable ability and item factors. Public resources such as EdNet [15] and OULAD [16] have enabled broad evaluation of these models.

These advances improve state estimation, but next-item selection is a decision problem rather than only a prediction problem. HARLA-ED provides a strong example of combining knowledge-state modeling with hierarchical RL and Monte Carlo dropout uncertainty [12]. The present work narrows the scope to a psychometrically interpretable setting so that the effect of uncertainty-aware control can be separated from representation capacity.

C. Reinforcement Learning and Uncertainty

Reinforcement learning optimizes sequential decisions from delayed rewards [19]. Double Q-learning reduces the positive action-value bias of standard Q-learning [20], while Bayesian Q-learning explicitly models uncertainty over value estimates [21]. UCB methods [22], Thompson sampling [23], bootstrapped value functions [24], and the uncertainty Bellman equation [25] all show that uncertainty can guide exploration more effectively than undirected randomization. PPO [26] is a widely used policy-gradient alternative and is employed by several complex educational agents. Recent reviews indicate growing use of RL in education while also emphasizing evaluation and deployment challenges [27]; bandit-based tutoring systems provide an earlier example of uncertainty-driven adaptive sequencing [33].

Our contribution differs from generic exploration bonuses. The uncertainty that modulates action is tied to a measurable psychometric quantity—the standard error of an item difficulty estimate—and the action controls risk aversion in the item-information objective. This makes the learned behavior easier to inspect and deploy than an unconstrained item-level action network.

III. PROBLEM FORMULATION

Consider an item bank J with items j=1,…,M. A learner has latent scalar proficiency θ. Item j has true discrimination a_j and difficulty b_j, but the assessment system observes estimates $\hat{a}_j$ and $\hat{b}_j$ together with a calibration standard error s_j for difficulty. The response simulator follows the 2PL model

$$P(y_t = 1 \mid \theta, j_t) = \sigma(a_j(\theta - b_j)), \qquad (1)$$

where $\sigma(\cdot)$ is the logistic function. During administration, the system maintains a posterior $p_t(\theta)$ over a fixed grid Θ. The objective is not simply to minimize the number of items. An assessment policy should reduce estimation error while maintaining calibrated uncertainty and avoiding excessive reliance on poorly calibrated items.

We therefore evaluate four quantities: mean number of administered questions, RMSE of the final posterior mean relative to the simulator's latent θ, empirical coverage of the nominal 95% posterior interval, and the mean calibration standard error of selected items. The four metrics deliberately expose trade-offs. A policy that stops quickly can achieve low length while sacrificing error or coverage; a policy that always chooses the most precisely calibrated items can be conservative and inefficient.

IV. PROPOSED METHOD

A. Bayesian Ability Belief with Item Uncertainty

The learner belief begins with a discrete uniform prior on $\Theta=[-4,4]$ using 161 grid points. To propagate difficulty uncertainty, we approximate b_j as Gaussian around $\hat{b}_j$ with variance s_j^2. Marginalizing a logistic likelihood over a Gaussian parameter has no simple closed form, so we use a standard attenuation approximation. For a candidate ability θ_k,

$$\tilde{p}_j(\theta_k) = \sigma(\hat{a}_j(\theta_k - \hat{b}_j) / \sqrt{(1 + \pi(\hat{a}_j s_j)^2/8)}). \quad (2)$$

After observing y_t, the posterior update is

$$p_t(\theta_k) \propto p_{t-1}(\theta_k) \tilde{p}_j(\theta_k)^{y_t}[1-\tilde{p}_j(\theta_k)]^{1-y_t}. \quad (3)$$

The posterior mean μ_t, standard deviation σ_t, and entropy H_t summarize learner uncertainty. This belief representation has two practical advantages: it preserves uncertainty near test boundaries and it prevents a high-standard-error item from shrinking the posterior as aggressively as an equally difficult, well-calibrated item.

B. Uncertainty-Conditioned Risk Scheduling

Rather than ask the RL agent to select among hundreds or thousands of item IDs, UA-DoubleQ selects a risk coefficient λ_t from $\{0, 0.5, 1, 2, 4, 8\}$, plus a STOP action. For each λ_t, items are ranked by a robust information score

$$\tilde{I}_j(t,\lambda_t) = [\hat{a}_j^2 p_j(1-p_j)] / [1 + \lambda_t(\hat{a}_j s_j)^2], \quad (4)$$

where $p_j=\sigma(\hat{a}_j(\mu_t-\hat{b}_j))$. The numerator is the local 2PL Fisher information at the current posterior mean; the denominator discounts items whose calibration uncertainty can dominate their nominal information. A mild exposure factor is applied in implementation so repeated selection of the same item is discouraged. $\lambda_t=0$ recovers aggressive information maximization, whereas larger λ_t increasingly prefers well-calibrated items. The key modeling choice is that risk aversion is state dependent. When σ_t is large, rejecting every uncertain item can slow information acquisition. When σ_t is small, the same item uncertainty can become the dominant source of error. The RL controller learns this trade-off from simulated assessment trajectories.

C. Double Q-Learning Controller and Reward

The uncertainty-aware state is a compact discretization of (μ_t,σ_t). The baseline vanilla Q agent replaces σ_t by a coarse test-length bin, providing a controlled ablation of explicit learner uncertainty. Two Q tables are updated with Double Q-learning [20], alternating which table selects the maximizing action and which evaluates it. Exploration during training follows a decaying ε-greedy schedule.

For a nonterminal action, the reward combines normalized entropy reduction, administration cost, and an item-risk term:

$$r_t = 13(H_{t-1}-H_t)/H_0 - 0.045 - 0.045 \lambda_t(\hat{a}_j s_j)^2. \quad (5)$$

The STOP action is trained only inside the simulator, where latent θ is available: a stop within $|\mu_t-\theta|\leq 0.38$ receives a positive terminal reward; an inaccurate stop receives a larger negative reward, with small additional penalties for test length and high residual σ_t. Latent θ is not required during deployment because the learned stopping rule depends only on the observable belief state.

D. Computational Profile and Policy Interpretability

The decomposition into a psychometric scorer and a small RL controller is deliberate. For an item bank of M items and a grid of G ability points, one response update costs O(G), while scoring all unused items costs O(M). The Q-table lookup is constant time after discretization. With M=700 and G=161 in our benchmark, the sequential decision therefore remains dominated by simple vector operations rather than by a deep network forward pass. The action set contains only seven choices, making it possible to inspect whether a learned policy is acting aggressively (small λ), conservatively (large λ), or terminating.

This structure also separates two uncertainty roles that are often conflated. Posterior standard deviation σ_t describes uncertainty about the learner and is part of the RL state. Item standard error s_j describes uncertainty in the measurement instrument and appears inside the robust information score. A large σ_t can justify

collecting more information, whereas a large s_j can justify discounting a nominally informative item. The learned λ_t determines how these two signals are traded at each stage of an assessment.

E. Assessment Procedure

At administration time, UA-DoubleQ repeats a short closed loop. It summarizes the current posterior, maps (μ_t,σ_t) to a discrete state, obtains λ_t or STOP from the two Q tables, ranks unused items with (4), observes the learner response, and applies (3). No simulator-only variable appears in this loop. The latent ability θ is used only during training to assign terminal rewards and during evaluation to compute error.

The operational sequence is: (1) initialize $p_0(\theta)$ uniformly and mark all items unused; (2) compute μ_t, σ_t, and H_t; (3) choose the action maximizing Q_1+Q_2; (4) return the current posterior summary if the action is STOP; otherwise map the action to λ_t and select the highest-scoring unused item under (4); (5) observe y_t, update the posterior using (3), and repeat until STOP or T_max. During training, one Q table selects the bootstrap action while the other evaluates it, with the roles randomized at each update.

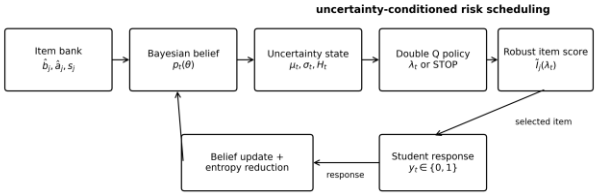


Figure 1UA-DoubleQ separates Bayesian uncertainty estimation from RL-controlled risk scheduling. The action is a risk coefficient λ or STOP, not an item ID.

V. EXPERIMENTAL DESIGN

A. ASSISTments-Calibrated Semi-Synthetic Benchmark

The benchmark uses a semi-synthetic design because adaptive item selection creates a counterfactual problem: historical logs contain a response only for the item that was actually administered. A new policy may select a different item, for which the historical response is undefined. We therefore generate responses from a known 2PL environment while using the public ASSISTments 2009–2010 Skill Builder corpus as the real-data anchor [4], [28]. The official data page identifies it as mastery-learning data and provides a corrected student–problem version of the corpus [28]. Our simulator is calibrated to a mean correct-response probability near 0.64; the realized mean is 0.645 across the three evaluation seeds. Each seed generates 700 items. True difficulties are sampled from a truncated normal distribution and true discriminations from a clipped log-normal distribution. Calibration counts follow a long-tailed log-normal distribution; difficulty standard errors range from approximately 0.06 to 0.58. Estimated item parameters are obtained by perturbing the true values proportionally to their calibration uncertainty. Learner proficiency is sampled from a three-component Gaussian mixture chosen to create heterogeneous low-, middle-, and high-proficiency groups. The maximum test length is 25 items. For each seed, UA-DoubleQ and vanilla Double-Q are trained for 2,000 simulated examinees. Evaluation then uses a disjoint set of 500 examinees. We report mean and standard deviation across seeds 11, 22, and 33. The evaluation therefore summarizes 1,500 held-out simulated examinees, and every number in the tables is obtained from these fixed-seed runs under the stated protocol.

B. Baselines

Random samples an unused item uniformly. MFI selects the item with maximum estimated 2PL Fisher information and treats item estimates as fixed. Robust-MFI applies a fixed uncertainty penalty with λ=4 and uses uncertainty-aware belief updates. Thompson samples θ from the current posterior and applies the same fixed robust item score around the sample. Vanilla-Q uses Double Q-learning and the same risk-action set as UA-DoubleQ, but its state omits posterior σ_t and uses elapsed-test bins instead. UA-DoubleQ is the full method.

C. Metrics

Questions measures assessment burden. RMSE measures ability-estimation error against the simulator's known latent θ. Coverage is the fraction of examinees whose latent θ falls inside μ±1.96σ at termination; values closer to 0.95 indicate better calibrated posterior uncertainty. Item SE is the average s_j of administered items. Because no single scalar metric fully captures the assessment objective, we interpret methods as efficiency–accuracy–calibration trade-offs rather than assigning a single winner.

D. Counterfactual Control and Reproducibility

All policies within a seed share the same generated item bank and the same held-out proficiency sample. The true item parameters are used only by the response generator; selectors observe the perturbed estimates and, where applicable, the difficulty standard errors. This separation prevents a policy from accessing information that would be unavailable in an operational assessment. The learned agents are trained on 2,000 additional simulated examinees per seed and are never updated on the 500 held-out evaluation examinees.

Three independent seeds regenerate the item bank, calibration errors, learners, exploration trajectory, and evaluation responses. Reporting the standard deviation across these seeds is important because learned stopping can be sensitive to the particular combination of item uncertainty and proficiency mix. The complete experiment is computationally modest: a full three-seed CPU run completes in roughly half a minute in the execution environment used for this study, making repeated evaluation under identical settings practical.

E. Main Hyperparameters

Table 1. Core experimental settings. All values are read directly from the experiment configuration.

Setting	Value
Item bank	700 items
Ability grid	161 points on [−4,4]
Maximum length	25 items
RL training	2,000 examinees/seed
Evaluation	500 examinees/seed
Risk actions λ	0, 0.5, 1, 2, 4, 8 + STOP
Double-Q α, γ	0.13, 0.97
Reward coefficients	13, 0.045, 0.045
Precision stop for non-RL baselines	posterior SD ≤ 0.43 after ≥4 items

VI. RESULTS

Table 2Main results across three independent seeds. Ques. and RMSE are mean±standard deviation across seeds; coverage and item SE are seed means.

Method	Ques.	RMSE	Cov.	Item SE
Random	24.58± 0.12	0.517± 0.009	94.8 %	0.30 9
MFI	9.13±0. 48	0.521± 0.033	87.7 %	0.31 8
Robust- MFI	10.90± 0.42	0.463± 0.022	92.3 %	0.18 4
Thompso n	12.65± 0.33	0.458± 0.020	92.6 %	0.18 5

Method	Ques.	RMSE	Cov.	Item SE
Vanilla-Q	16.89± 3.48	0.395± 0.035	90.3 %	0.24 9
UA-DoubleQ	14.02± 1.25	0.444± 0.042	92.1 %	0.23 7

TABLE II.

A. Main Trade-offs

Table II shows three distinct operating regimes. MFI is shortest at 9.13 questions but has the highest ability error among structured adaptive policies (RMSE 0.521) and only 87.7% coverage, indicating overconfident point-estimate selection. Adding a fixed item-uncertainty penalty improves Robust-MFI to RMSE 0.463 and 92.3% coverage at 10.90 questions. Thompson sampling gives a similar RMSE of 0.458 and 92.6% coverage with 12.65 questions. UA-DoubleQ occupies a different point on the Pareto frontier. It uses 14.02±1.25 questions, obtains RMSE 0.444±0.042, and reaches 92.1±1.1% coverage. Relative to point-estimate MFI, its RMSE is 14.9% lower, coverage is 4.3 percentage points higher, and the mean calibration standard error of selected items is 25.4% lower. These gains require 53.5% more questions than MFI, making the trade-off explicit rather than hidden behind a single efficiency statistic. The most informative ablation is the comparison with Vanilla-Q. Without posterior uncertainty in the state, Vanilla-Q achieves lower RMSE (0.395) but runs for 16.89±3.48 questions and has only 90.3% coverage. UA-DoubleQ shortens the test by 17.0%, improves coverage by 1.8 percentage points, and selects items with 4.8% lower mean calibration uncertainty. The RMSE increases by 12.4%, showing that explicit uncertainty primarily changes when the agent decides additional precision is worth the item cost.

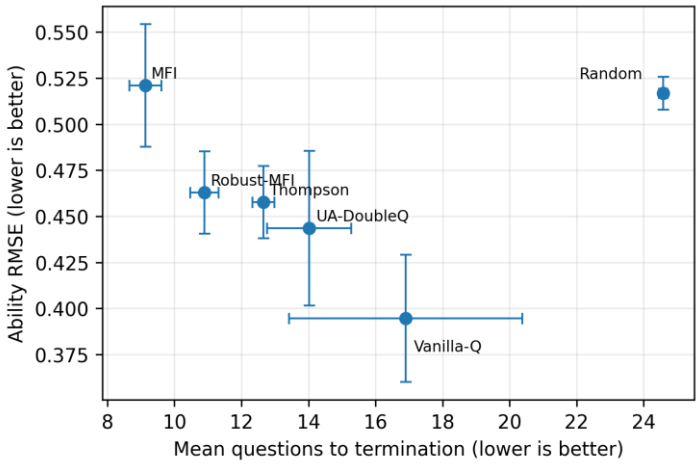


Figure 2Efficiency–accuracy trade-off across three seeds. Error bars denote standard deviation across seeds; lower-left is preferred when only length and RMSE are considered

B. Role of Item-Calibration Uncertainty

The selected-item statistics clarify why point-estimate MFI becomes overconfident. Its selected items have mean difficulty SE 0.318, whereas Robust-MFI and Thompson reduce this value to about 0.184. UA-DoubleQ averages 0.237 because it sometimes accepts a less precisely calibrated item when the expected information gain justifies the risk. The learned mean risk coefficient is 2.32, between aggressive MFI (λ=0) and the fixed robust baseline (λ=4). This intermediate value is consistent with the intended design: uncertainty is neither ignored nor treated as an absolute exclusion criterion. The random policy attains coverage close to 95%, but only because it administers nearly the full 25-item budget and retains a broad posterior; its RMSE remains 0.517. Thus interval coverage alone is insufficient. A useful adaptive policy must jointly control posterior width, point-estimation error, and assessment length.

C. Operational Policy Profile

Table 3Operational profile across seeds. Exposure Gini summarizes concentration of item use; stop rate is the fraction of examinees ending before the 25-item limit

Method	Exposure Gini	Stop Rate	Mean λ
MFI	0.965	100.0%	—
Robust-MFI	0.961	100.0%	—
Thompson	0.962	100.0%	—
Vanilla-Q	0.922	67.0%	2.18
UA-DoubleQ	0.931	90.7%	2.32

Table III shows that uncertainty-aware risk scheduling changes more than the final accuracy metric. UA-DoubleQ has mean λ=2.32, higher than Vanilla-Q's 2.18, and terminates before the item cap for 90.7% of examinees versus 67.0% for Vanilla-Q. This directly explains the shorter average assessment. Its exposure Gini of 0.931 is also below the 0.965 level of MFI, indicating a modest reduction in concentration on a narrow set of high-information items, although it remains far from the deliberately diffuse random policy.

The fixed robust methods have high exposure concentration because a single deterministic score repeatedly favors a small subset of well-calibrated items. UA-DoubleQ can move between λ values as the posterior changes, which broadens the effective candidate set without requiring an explicit exposure-control constraint. Exposure balance is not optimized by the reward in this study, so the Gini statistic is interpreted diagnostically rather than as a fairness objective.

D. Stability Across Seeds

Seed-level variability is non-negligible for learned stopping. UA-DoubleQ's question count has a standard deviation of 1.25 and RMSE a standard deviation of 0.042. Vanilla-Q varies more strongly in length (3.48 questions), consistent with the idea that explicit σ_t provides a more stable stopping signal across item-bank realizations. The seed-level rows also show why an average alone can be misleading: on seed 11 UA-DoubleQ attains RMSE 0.395, while on seeds 22 and 33 its RMSE is approximately 0.470 and 0.466. The direction of its efficiency advantage over Vanilla-Q is especially pronounced on seeds 22 and 33. We therefore retain both mean and across-seed dispersion in the main table rather than presenting a single deterministic run.

E. Additional Error Diagnostics

The per-seed logs provide two additional checks not used to tune the methods. Averaged over seeds, UA-DoubleQ has mean absolute error 0.348 and terminal posterior standard deviation 0.383. The corresponding values for MFI are 0.415 and 0.415, while Robust-MFI yields 0.360 and 0.417. Thus UA-DoubleQ's lower RMSE relative to the two MFI variants is also reflected in absolute error, rather than being driven only by a small number of large squared errors. Vanilla-Q again represents the opposite end of the learned-policy trade-off: its mean absolute error is 0.310 and its terminal posterior standard deviation is 0.335. Those smaller values are obtained with 16.89 questions on average and a lower stopping rate. The evidence therefore points to stopping behavior, not a failure of the underlying Bayesian update, as the main reason that UA-DoubleQ gives up some point accuracy in exchange for shorter assessments and better empirical coverage. This interpretation is consistent with the controller's explicit STOP action and with the higher early-termination rate in Table III.

VII. DISCUSSION

The results support a simple design principle: uncertainty should control risk, not merely be appended to the observation vector. The RL agent does not need to learn item identity embeddings or a large policy over the complete bank. Instead, the psychometric layer ranks items and the RL layer learns how strongly to discount

calibration uncertainty at the current stage of the test. This preserves a clear operational interpretation of every action.The framework complements rather than replaces deep knowledge tracing. In a multidimensional deployment, μ_t and σ_t can be replaced by concept-specific posterior summaries from a deep or graph-based estimator. HARLA-ED [12] demonstrates the strength of coupling rich knowledge-state estimation with reinforcement learning; the present study contributes a compact risk-control mechanism that can sit inside such a larger architecture. Likewise, MC-dropout or ensemble uncertainty [17], [32] could provide the item- or concept-level uncertainty inputs when analytical standard errors are unavailable. The semi-synthetic protocol is a methodological choice rather than a substitute for real deployment evidence. It is appropriate for controlled counterfactual comparison because every policy can query the same latent response model. At the same time, the response simulator uses a stationary one-dimensional ability and binary outcomes. Real educational assessment can involve multidimensional skills, learning during testing, partial credit, response-time effects, and item drift. These extensions can be incorporated without changing the risk-scheduling action concept, but they would alter the belief model and reward. A second consideration is calibration. The 95% interval coverage of all adaptive policies remains below nominal in at least some seeds. This is expected because item discrimination is also estimated with error while the belief update only directly marginalizes difficulty uncertainty. A natural extension is joint posterior inference over ability and item parameters, or an ensemble belief that propagates both discrimination and difficulty uncertainty. Calibration techniques such as temperature scaling [18] are also relevant when the belief state is produced by a neural estimator. Finally, the reward parameters encode an assessment preference: test burden is penalized and inaccurate early stopping is penalized more heavily. Different programs may choose different utilities. High-stakes certification may prioritize coverage and conservative stopping; formative classroom assessment may tolerate larger posterior intervals in exchange for shorter tests. Because UA-DoubleQ exposes λ and STOP as interpretable actions, such preference changes can be implemented without redesigning the item model.

A. Deployment Implications

Operationally, the method is most useful when an item bank contains mixed calibration quality: mature items with precise parameter estimates, low-exposure items, and recently introduced items. A conventional MFI rule can overuse a newly estimated item if its point estimate happens to place it near the current learner ability and gives it high nominal discrimination. The robust denominator in (4) converts the associated standard error into a transparent penalty, while the learned λ prevents that penalty from being fixed for the entire assessment. The same mechanism provides a straightforward path to multidimensional assessment. A concept-level knowledge tracer can supply a vector of means and uncertainties, and the psychometric layer can compute candidate-item information for the targeted concept or subspace. The RL state can then summarize which dimensions remain uncertain and how much reliable information is available in the bank. Importantly, this extension preserves the separation between state estimation and risk control instead of requiring the decision network to rediscover psychometric structure from raw item identifiers.

B. Scope of the Evidence

The present evidence establishes the behavior of the mechanism under controlled calibration noise, not a claim about a particular live testing program. The semi-synthetic design is valuable precisely because it exposes latent ability and counterfactual item responses, making RMSE and interval coverage identifiable for every policy. A future live evaluation can then focus on external validity: multidimensional content, item drift, response times, and learner behavior under adaptive sequencing. The controlled benchmark provides a reproducible reference point for that next stage.

C. Threats to Validity and Extension Path

Three choices define the current scope. First, the response model is unidimensional 2PL, so the simulator does not represent learning during the assessment or cross-skill transfer. Second, the belief update directly attenuates difficulty uncertainty but does not integrate the uncertainty of discrimination a_j; this helps explain why empirical interval coverage remains below 95% for several adaptive policies. Third, the terminal reward

uses a fixed accuracy tolerance in training, which encodes one assessment utility rather than every possible institutional objective. Each choice has a direct extension. Multidimensional IRT or a knowledge-tracing posterior can replace the scalar grid; joint sampling or ensembles can propagate uncertainty in both item parameters; and the stopping reward can be reweighted for a program's desired precision-cost trade-off. None of these changes requires abandoning the central action abstraction. The controller can continue to choose a small risk parameter while a domain-specific measurement model supplies posterior uncertainty and candidate-item information.

VIII. CONCLUSION

This paper presented UA-DoubleQ, an uncertainty-aware reinforcement-learning method for adaptive educational assessment under item-calibration uncertainty. The method maintains a Bayesian learner belief, discounts nominal Fisher information according to item difficulty standard error, and lets a Double Q-learning controller adapt the risk coefficient or terminate the test. In an ASSISTments-calibrated semi-synthetic benchmark, the method produced a transparent middle ground between short but overconfident MFI testing and longer, accuracy-oriented vanilla Q-learning. Across three seeds, UA-DoubleQ required 14.02 questions on average with RMSE 0.444 and 92.1% empirical 95% coverage. Compared with vanilla Q-learning, it reduced test length by 17.0% while improving coverage and reducing selected-item calibration uncertainty; compared with point-estimate MFI, it substantially improved RMSE and coverage. The main implication is that epistemic uncertainty in the item bank should be a controllable part of sequential assessment policy rather than a hidden modeling detail.

REFERENCES

F. M. Lord, Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1980.

W. J. van der Linden and G. A. W. Glas, Eds., Computerized Adaptive Testing: Theory and Practice. Dordrecht, The Netherlands: Kluwer Academic, 2000, doi: 10.1007/0-306-47531-6.

H.-H. Chang and Z. Ying, "A global information approach to computerized adaptive testing," Applied Psychological Measurement, vol. 20, no. 3, pp. 213–229, 1996, doi: 10.1177/014662169602000303.

M. Feng, N. T. Heffernan, and K. R. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," User Modeling and User-Adapted Interaction, vol. 19, no. 3, pp. 243–266, 2009, doi: 10.1007/s11257-009-9063-7.

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, 1995, doi: 10.1007/BF01099821.

M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized Bayesian knowledge tracing models," in Proc. 16th Int. Conf. Artificial Intelligence in Education (AIED), 2013, pp. 171–180, doi: 10.1007/978-3-642-39112-5_18.

C. Piech et al., "Deep knowledge tracing," in Advances in Neural Information Processing Systems 28, 2015.

J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in Proc. 26th Int. World Wide Web Conf., 2017, pp. 765–774, doi: 10.1145/3038912.3052580.

S. Pandey and G. Karypis, "A self-attentive model for knowledge tracing," in Proc. 12th Int. Conf. Educational Data Mining (EDM), 2019, pp. 384–389.

A. Ghosh, N. T. 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, doi: 10.1145/3394486.3403282.

D. Shin, Y. Shim, H. Yu, S. Lee, B. Kim, and Y. Choi, "SAINT+: Integrating temporal features for EdNet correctness prediction," in Proc. 11th Int. Learning Analytics and Knowledge Conf. (LAK), 2021, pp. 490–496, doi: 10.1145/3448139.3448188.

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, no. 1, Art. no. 58, 2026, doi: 10.3390/sym18010058.

F. Wang et al., “Neural cognitive diagnosis for intelligent education systems,” in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 4, pp. 6153–6161, 2020, doi: 10.1609/aaai.v34i04.6080.

C.-K. Yeung, “Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory,” in Proc. 12th Int. Conf. Educational Data Mining (EDM), 2019, pp. 683–686.

Y. Choi et al., “EdNet: A large-scale hierarchical dataset in education,” in Artificial Intelligence in Education, LNCS 12164, 2020, pp. 69–73, doi: 10.1007/978-3-030-52240-7_13.

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.

Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. 33rd Int. Conf. Machine Learning (ICML), vol. 48, pp. 1050–1059, 2016.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Machine Learning (ICML), vol. 70, pp. 1321–1330, 2017.

R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

H. van Hasselt, “Double Q-learning,” in Advances in Neural Information Processing Systems 23, 2010, pp. 2613–2621.

R. Dearden, N. Friedman, and S. Russell, “Bayesian Q-learning,” in Proc. 15th Nat. Conf. Artificial Intelligence (AAAI), 1998, pp. 761–768.

P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” Machine Learning, vol. 47, pp. 235–256, 2002, doi: 10.1023/A:1013689704352.

W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” Biometrika, vol. 25, nos. 3–4, pp. 285–294, 1933, doi: 10.1093/biomet/25.3-4.285.

I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, “Deep exploration via bootstrapped DQN,” in Advances in Neural Information Processing Systems 29, 2016, pp. 4026–4034.

B. O’Donoghue, I. Osband, R. Munos, and V. Mnih, “The uncertainty Bellman equation and exploration,” in Proc. 35th Int. Conf. Machine Learning (ICML), vol. 80, pp. 3839–3848, 2018.

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.

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.