Academic Journal of Education
ISSN: 3080-9940 (Print), ISSN: 3080-9959 (Online) | Vol. 3, No. 1, 2026

Uncertainty-Aware Reinforcement Learning for Efficient Adaptive Educational Assessment

Anand Iyer *, Clara Park, Jade Xu

College of Education, University of Maryland, College Park, USA

*** Corresponding author:** Anand Iyer (Email: anandlyer021@umd.edu)

Abstract: Adaptive assessment is efficient only if the system knows how uncertain it is about the learner, because that uncertainty decides which question is worth asking and when to stop. Reinforcement-learning (RL) assessment agents often derive it from Monte Carlo (MC) dropout over a neural knowledge encoder and reward its reduction. We study uncertainty-aware RL for efficient assessment by combining a calibration-aware Bayesian knowledge-state filter (assumed-density filtering on a probit item-response model with knowledge-component deviations and item-calibration noise), a target-aware information criterion with a closed-form expected covariance contraction, and a PPO-trained termination policy with an uncertainty-based state, a per-question cost, a policy-invariant shaping reward and a cross-fitted training environment. On ASSISTments 2009 and Junyi Academy, the selector significantly outperforms random selection and, from ten questions on, a re-implemented HARLA-ED operational agent under both a Bayesian and the agent's own neural estimator (NLL 0.552 vs. 0.558 after 20 questions on ASSISTments) and is statistically tied with mutual-information selection. In our re-implementation, the agent's MC-dropout uncertainty grows as responses accumulate (0.10 to 0.18), its uncertainty-reduction reward did not improve validation likelihood in any run, and stopping on it needs 1.7–2.5 times as many questions as a fixed-length test of equal accuracy. The learned termination policy matches fixed-length efficiency (5.1–5.4 vs. 5 questions) without threshold tuning, and a value-of-information analysis explains why larger learner-specific savings are hard to obtain in replay protocols.

Keywords: Computerized Adaptive Testing; Reinforcement Learning; Uncertainty Calibration; Assumed-Density Filtering; Optimal Stopping; Reward Shaping; Knowledge Tracing.

1. Introduction

Adaptive assessment shortens a test by choosing every question in light of the answers already given. Classical computerized adaptive testing (CAT) couples an item response theory (IRT) model with an information-based selection rule and a stopping rule, and its design principles are well established [1]. Over the last few years, the field has moved toward selection policies that are learned from large response logs, through bilevel meta-learning [2], reinforcement learning (RL) [3], and multi-objective RL [4]; a recent survey documents this shift [5]. The same trend is visible in educational RL more broadly, where the choice of state and reward has repeatedly decided whether learned policies beat simple baselines [6].

An efficient test has to make two decisions well: which question to ask next, and when the evidence is sufficient to stop. Both decisions depend on how uncertain the system is about the learner. Learned assessment agents increasingly obtain this uncertainty from neural knowledge encoders through Monte Carlo (MC) dropout [7] and then reward the agent for reducing it. Whether such estimates behave like epistemic uncertainty, shrinking as evidence accumulates and pointing to where another question would help, is seldom checked, and even the calibration of deep predictions is known to be fragile [8]. A signal that does not contract cannot reward informative questions, and a threshold on it cannot tell when to stop. Psychometric stopping rules, in turn, are hand-tuned thresholds on a single statistic such as the posterior standard deviation [9] or the information left in the pool [10], and they ignore the per-question cost that defines efficiency.

This paper studies uncertainty-aware RL for efficient assessment with calibration as the organizing principle. We combine three components. First, a calibration-aware Bayesian knowledge state: a probit IRT model augmented with knowledge-component (KC) deviations is filtered by closed-form assumed-density filtering (ADF), and its predictive distribution absorbs the error of the calibrated item parameters. Second, a target-aware information criterion that asks the question expected to reduce predictive uncertainty on the items the learner is still likely to face. Third, a termination policy trained with proximal policy optimization (PPO) whose state consists of calibrated uncertainty signals and whose reward trades realized accuracy gains against a per-question cost, augmented with a potential-based shaping term that provably leaves the optimal policy unchanged [11].

The contributions are as follows. (i) We formulate efficient assessment as joint selection and optimal stopping, and give a policy-invariant shaping of the stopping reward built from the calibrated predictive distribution, together with a cross-fitted training environment that removes the optimism of in-sample rewards. (ii) We derive a closed-form, calibration-aware knowledge-state filter and a target-aware information criterion whose expected covariance contraction is exact under ADF. (iii) We re-implement the operational selector of a recent hierarchical RL assessment framework and diagnose its uncertainty signal: MC-dropout uncertainty grows rather than shrinks as responses accumulate, even where the encoder's predictive mean is accurate and well calibrated, so that its uncertainty-reduction reward is counterproductive and its uncertainty is an unreliable stopping signal. (iv) On two public datasets with three random splits, the target-aware selector reaches the 20-question NLL of random selection after 11.3 (ASSIST09) and 9.2 (Junyi) questions and, evaluated with the HARLA-style neural estimator, reaches that agent's 20-question NLL on ASSIST09 after 10.4

questions; MC-dropout stopping needs up to 2.5 times the questions of a fixed-length test, while the learned termination policy matches fixed length without tuning. A value-of-information analysis shows that the benefit of further questions is nearly unpredictable for individual learners in this protocol, which bounds what variable-length testing can save.

2. Related Work

2.1. Item Selection and Stopping in Adaptive Testing

Maximum Fisher information selection is the classical default [1]. Because Fisher information is evaluated at a noisy point estimate early in the test, global criteria based on Kullback–Leibler information [12] and Bayesian criteria that integrate over the posterior [13] were proposed, and the multidimensional case was treated with Bayesian D-optimality [14]. Item parameters are themselves estimated, and selection rules that maximize information capitalize on calibration error [15]; our predictive distribution addresses this directly. Variable-length tests stop when the posterior standard deviation of the expected a posteriori estimate falls below a target [9], when the remaining items carry little information, or when the predicted reduction of the standard error is small [10]. These rules are myopic thresholds and are only as good as the uncertainty they monitor.

2.2. Learning-Based Adaptive Testing and RL in Education

BOBCAT casts CAT as bilevel optimization and learns a data-driven selection algorithm [2]; NCAT formulates selection as RL with an attentive policy [3]; MAAT balances quality and diversity in a model-agnostic way [16]; GMOCAT adds graph-enhanced representations and multiple objectives [4]. These systems rely on learned response models such as deep knowledge tracing [17] or neural cognitive diagnosis [18], and they fix the test length. A review of RL for instructional sequencing stresses that reward design and evaluation protocol often matter more than policy capacity [6]. We follow that observation by keeping selection closed-form and spending the learning effort on termination, a decision that fixed-length learned selectors leave untouched.

2.3. Uncertainty for Sequential Decisions

MC dropout interprets dropout as approximate Bayesian inference [7], and calibration can be measured with the expected calibration error (ECE) [8]. In Bayesian active learning, BALD selects queries by mutual information with the parameters [19], whereas EPIG targets information about predictions on the input distribution of interest [20]; our target-aware criterion is a closed-form analogue of the latter for adaptive testing. ADF, the one-pass special case of expectation propagation [21], gives the Gaussian posterior updates we use. Potential-based shaping preserves optimal policies under any potential [11], which lets us inject calibrated uncertainty into the reward without biasing the termination policy. Graph-based attention decoding for closed-loop neuroregulation demonstrates the methodological value of adapting decisions to evolving behavioral signals in real time [29].

3. Method

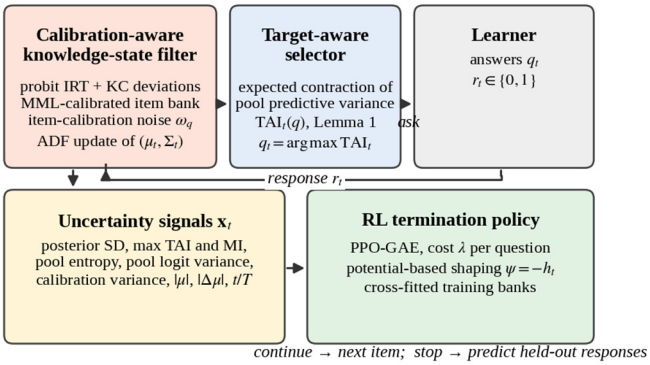


Fig 1. Overview of the framework. The calibration-aware filter maintains a Gaussian posterior over general ability and KC deviations; the target-aware selector asks the question with the largest expected contraction of the pool's predictive uncertainty; the RL termination policy decides from calibrated uncertainty signals whether another question is worth its cost.

3.1. Problem Formulation

A learner answers items from a bank ; each item is linked to a set of KCs . In the data-driven protocol that we adopt from learned CAT [2], [3], the items a learner actually answered are split into a candidate pool and a held-out set . At step the system picks , observes , and updates its state ; after stopping at it predicts the held-out responses. With the held-out negative log-likelihood (NLL) of the predictions made from state and the cost of one question, an efficient assessment solves

We treat selection with a calibrated, closed-form criterion (Sections III-B and III-C) and learn the termination policy with RL (Section III-D).

3.2. Calibration-Aware Knowledge-State Filter

The latent state contains a general ability and KC-specific deviations . Item with discrimination and difficulty follows the probit model

with loading vector and prior . Setting recovers the two-parameter normal-ogive model. Item parameters are calibrated on training learners by marginal maximum likelihood with Gauss–Hermite quadrature [22], which avoids the bias of joint estimation. From the complete-data information of the calibration sample we obtain variances and for and . Integrating this calibration error into the likelihood inflates the probit noise by

where are the posterior moments. Writing , and, the posterior predictive is. After observing , let and . Moment matching [21] gives the exact first two moments of the one-step posterior:

Each update costs with, needs no sampling, and keeps positive definite. Because enters both the update and the predictive, items with poorly estimated parameters move the state less and are predicted less confidently.

3.3. Target-Aware Information Selection

Mutual information with the latent state, the BALD criterion [19], rewards uncertainty in any latent direction, including KC deviations that matter for none of the items the learner will face. We instead score a candidate by the expected reduction of predictive uncertainty on the remaining pool , in the spirit of EPIG [20]. The required expectation is available in closed form under ADF.

Lemma 1. Under the update (5), with , where the

expectation is over the posterior predictive of . *Proof:* averaging over with weights gives .

The reduction of the logit variance of a pool item is therefore . Weighting it by the sensitivity of the predicted probability gives the target-aware information and the selector asks. With the pool-dependent sum is common to all candidates and the rule reduces to minimum expected posterior variance, one of the Bayesian criteria in [13]; with KC deviations it favors items whose KCs are shared by the learner's remaining pool.

3.4. RL Termination Policy

After every response the termination policy chooses ; the test stops at at the latest. The state collects calibrated uncertainty signals: the posterior standard deviation , the largest remaining TAI and mutual information (log scale), the mean predictive entropy and mean epistemic logit variance of the remaining pool, the mean calibration variance , and its last change, the elapsed fraction , and the proportion of correct answers. Let be the realized held-out NLL after responses and the expected NLL under the calibrated predictive, i.e., the mean binary entropy of the held-out predictions. With potential and , the reward is

Without the shaping terms the undiscounted return equals , so maximizing it solves (1) for the given selector. The shaping terms replace the noisy realized improvement of each step by its calibrated expectation without changing what is optimal.

Proposition 1. For any policy , the shaped and unshaped value functions satisfy ; hence both rewards induce the same optimal termination policies. This follows from the telescoping argument of [11], since depends only on the state and vanishes at termination. By contrast, a reward equal to the realized reduction of uncertainty averaged over the KCs of the selected item depends on the action through , is not of potential form, and carries no such guarantee; if it is the only reward, the agent is optimized for contraction of its own uncertainty estimate rather than for predictive accuracy.

We optimize with PPO [23] and generalized advantage estimation [24]. Rewards computed on learners whose responses were used to calibrate the bank are optimistic, because in-sample predictions are too good. We therefore cross-fit the environment: the training learners are split into four folds, each fold is assessed with a bank calibrated on the other three, and only the resulting out-of-sample trajectories are used for policy optimization. Model selection uses the objective (1) on validation learners.

3.5. Relation to Hierarchical RL Assessment

HARLA-ED [25] couples a hierarchical cognitive-graph knowledge encoder with a three-level PPO agent. Its strategic level sets diagnostic goals and termination criteria, its tactical level chooses question types, and its operational level selects items from a state that concatenates the KC-level knowledge estimate with its MC-dropout uncertainty; the information reward of that level is the realized reduction of the uncertainty over the KCs of the chosen item. Our design builds on the operational formulation of HARLA-ED, an item-level decision conditioned on an uncertainty-augmented state and optimized by PPO with generalized advantage estimation, and we reuse its optimization settings (clipping 0.2, discount 0.99, GAE parameter 0.95, ten policy epochs per iteration). We change the two ingredients that the diagnosis in Section V-C singles out: the uncertainty, which comes from the calibration-aware posterior of Section III-B instead of MC dropout, and the reward, which is the policy-invariant shaping (7) instead of the realized uncertainty reduction. The termination decision, which HARLA-ED assigns to a goal-level criterion, becomes the learned policy of Section III-D.

4. Experimental Setup

4.1. Datasets and Protocol

We use two public datasets in the preprocessed form released with RCD [26]: ASSISTments 2009–2010 skill builder (ASSIST09) [27] and Junyi Academy [28]. Table I summarizes them. The two differ in a way that matters for uncertainty: ASSIST09 has 17,671 items that share 123 KCs, and the median item has only 7 calibration responses, so item parameters are uncertain; in Junyi each of the 706 exercises is its own KC and the median item has 96 calibration responses. Learners are split 70/10/20 into training, validation and test learners, and item parameters are calibrated on all responses of the training learners. For adaptive testing we keep learners with at least 40 responses to calibrated items and use at most 100 responses per learner, so every pool holds at least 32 items. Each learner's items are split 80/20 into candidate pool and held-out set, and a test lasts at most questions. Results are averaged over three random pool/held-out splits of the test learners, and every learned policy is trained with the matching seed. We report accuracy (ACC), area under the ROC curve (AUC) and NLL pooled over all held-out responses.

Table 1. Dataset Statistics

	ASSIST09	Junyi
Learners / responses	2,493 / 267,415	10,000 / 353,835
Items / KCs	17,671 / 123	706 / 706
Calibration responses	194,386	247,345
Median responses per item	7	96
Eligible learners (train/val/test)	977 / 131 / 250	2,013 / 259 / 603
Mean responses per eligible learner	82.2	62.3

4.2. Compared Methods

Selection. Random; MFI, maximum Fisher information at the posterior mean, which for our state equals Bayesian D-optimality [14]; MI, mutual information with the latent state (BALD) [19]; and UA-Select, the target-aware rule (6). All four share the calibration-aware estimator of Section III-B, so their differences reflect only which items are asked. *HARLA-Op* is our re-implementation of the operational level of HARLA-ED [25]. Its encoder follows Eqs. (4)–(9) of that paper: interaction embeddings with positional encoding, an LSTM whose decay factor is set to one because replayed tests carry no timestamps, four-head attention with logarithmic recency bias, LayerNorm, a KC-level state, and MC-dropout uncertainty from passes with dropout 0.3. The policy scores each candidate from the KC-level estimates and uncertainties, is rewarded by its Eq. (16), and is trained with the PPO-GAE settings listed in Section III-E for 200 iterations of 20 trajectories; the checkpoint with the best validation NLL is kept. We omit its graph branch and its time, engagement and fairness reward terms, which need prerequisite graphs, response times or demographics that are unavailable or not comparable across the two datasets, so the comparison isolates the uncertainty mechanism and the information reward that drive its item choice. HARLA-Op is evaluated with its own encoder and, by replaying its item sequences, with our estimator.

Termination. All termination methods act on the same UA-Select trajectories: fixed length; the SE rule, which stops

when the posterior standard deviation falls below a threshold [9]; the minimum-information rule, which stops when the best remaining TAI falls below a threshold [10]; the MC-dropout rule, which thresholds the HARLA-Op uncertainty averaged over the KCs of the remaining pool; and RL-UA, our learned policy. Ablations of RL-UA replace its state by point features only (, proportion correct, and its change) or by the MC-dropout uncertainties, or drop the shaping term. Threshold rules are swept over 41 quantiles of their statistic and the RL policies over , which traces accuracy–length curves.

4.3. Implementation Details

Item banks use marginal maximum likelihood with 41 Gauss–Hermite nodes, Adam (learning rate 0.05, 300 steps) and L2 penalties of 1.0 on and 0.5 on . The model was selected on validation learners. Joint MAP estimation with 1, 2 or 4 latent dimensions was inferior; on Junyi, for example, the two-dimensional joint MAP model raised the NLL of random selection from 0.529 before the test to 0.566 after five responses, the signature of overconfident updates, whereas marginal maximum likelihood lowered it from 0.505 to 0.489. The KC deviation scale was selected on validation learners among , giving on ASSIST09; on Junyi because KCs coincide with items. The HARLA-Op encoder uses and is pre-trained for 3,000 steps of 128 random histories with AdamW (learning rate). The termination policy and its value function are two-layer MLPs with 32 tanh units, trained for 150 iterations of 512 sampled episodes with Adam (learning rate); the checkpoint with the best validation objective (1) is kept. Everything is implemented in JAX and runs on a single CPU core: calibrating a bank takes about 1.5 minutes and training one termination policy about 10 seconds.

5. Results

5.1. Item Selection

Table 2. Held-out AUC (higher is better) and NLL (lower is better) after 5, 10 and 20 questions

Method	AUC5	AUC10	AUC20	NLL5	NLL10	NLL20
ASSIST09, Bayesian est.						
Random	0.695	0.716	0.733	0.606	0.589	0.574
MFI	0.715	0.729	0.741	0.589	**0.576**	0.566
MI (BALD)	0.715	0.729	**0.742**	0.589	**0.576**	**0.565**
HARLA-Op items	0.694	0.710	0.729	0.608	0.593	0.576
UA-Select (ours)	**0.718**	**0.731**	**0.742**	**0.588**	**0.576**	**0.565**
ASSIST09, Neural est.						
Random items	0.725	0.736	0.746	0.571	0.563	0.557
HARLA-Op (native)	0.723	0.733	0.743	0.573	0.564	0.558
UA-Select items	**0.730**	**0.743**	**0.751**	**0.570**	**0.559**	**0.552**
Junyi, Bayesian est.						
Random	0.805	0.815	0.821	0.504	0.493	0.485
MFI	0.815	**0.822**	**0.826**	0.492	**0.484**	**0.479**
MI (BALD)	**0.816**	**0.822**	**0.826**	0.492	**0.484**	**0.479**
HARLA-Op items	0.805	0.812	0.820	0.505	0.495	0.485
UA-Select (ours)	**0.816**	**0.822**	**0.826**	**0.491**	**0.484**	**0.479**
Junyi, Neural est.						
Random items	0.806	0.813	0.818	0.503	0.495	0.489
HARLA-Op (native)	0.804	0.811	0.817	0.505	0.497	0.490
UA-Select items	**0.813**	**0.817**	**0.821**	**0.495**	**0.490**	**0.485**

AUC/NLL: after questions. Means over three random pool/held-out splits (standard deviations ≤0.012 AUC and ≤0.007 NLL). “Bayesian est.”: all selectors share the calibration-aware filter; “Neural est.”: item sequences scored by the MC-dropout encoder of HARLA-Op. HARLA-Op rows use 3 (ASSIST09) and 3 (Junyi) seeds. Best value per block in bold.

Table II and Fig. 2 report held-out AUC and NLL after 5, 10 and 20 questions. With the common Bayesian estimator, UA-Select gives the best or tied-best AUC and NLL at every length on both datasets. Its advantage over random selection is significant for every length and dataset (paired Wilcoxon test on per-learner NLL, all), and it reaches the 20-question NLL of random selection after 11.3 questions on ASSIST09 and 9.2 on Junyi. Differences to MI and MFI are small and not significant in most comparisons, which is expected: with the target-aware rule reduces to a Bayesian variance criterion, and on ASSIST09 the KC deviations add only a modest preference for items whose KCs recur in the pool. Replaying the item sequences of HARLA-Op through the same estimator gives an NLL of 0.576 (ASSIST09) and 0.485 (Junyi) after 20 questions, worse than UA-Select (0.565 and 0.479), and on ASSIST09 even worse than random selection. The per-learner differences to HARLA-Op after 20 questions are significant (and).

The lower blocks of Table II score item sequences with the MC-dropout encoder of HARLA-Op. On ASSIST09 this neural encoder, which follows the knowledge encoder of HARLA-ED [25], is the stronger predictor: with our item sequences it reaches an NLL of 0.552 after 20 questions versus 0.565 for the Bayesian filter, so expressive knowledge models remain valuable on data with many sparsely calibrated items. The ranking of the item sequences is unchanged, however. Our sequences beat random sequences (0.557) and HARLA-Op's own sequences (0.558;) and reach HARLA-Op's 20-question NLL after 10.4 questions. On Junyi the Bayesian filter is the more accurate estimator at every length (0.479 versus 0.485 after 20 questions), and our sequences again beat HARLA-Op's under its own encoder ().

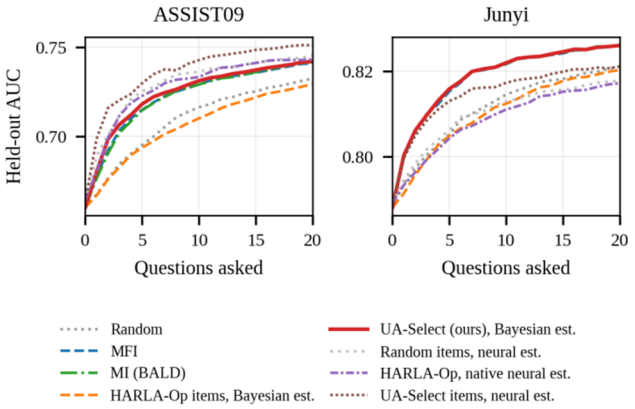


Fig 2. Held-out AUC versus the number of questions. Solid red: UA-Select with the Bayesian estimator; dotted brown: the same item sequences scored by the MC-dropout neural encoder; HARLA-Op is shown with its native encoder and replayed through the Bayesian estimator.

5.2. Components of the Knowledge-State Filter

Table 3. Ablation of the knowledge-state filter (UA-Select; held-out NLL)

Variant	AS@5	AS@10	AS@20	JY@5	JY@10	JY@20
Full filter	0.588	0.576	0.565	0.491	0.484	0.479
w/o item-calibration noise	0.591	0.579	0.569	0.493	0.485	0.480
w/o KC deviations ()	0.589	0.577	0.569	–	–	–
Classical: MI, 1-D, no noise	0.595	0.584	0.576	–	–	–

AS: ASSIST09, JY: Junyi. KC deviations are inactive on Junyi, where every exercise is its own KC

Table III ablates the filter. On ASSIST09, removing the item-calibration noise raises the 20-question NLL from 0.565 to 0.569, and removing the KC deviations raises it to 0.569; the classical combination of MI selection, one latent dimension and no calibration noise reaches only 0.576. The calibration noise matters most where items are poorly calibrated: on Junyi, whose median item has 96 calibration responses, the difference is 0.480 versus 0.479. The gaps widen as the test proceeds, since both components enter every additional update. Adaptive reinforcement learning for cybersecurity incident response offers a complementary experimental example of balancing sequential decision quality against resource and operational constraints [30].

5.3. Uncertainty Signals

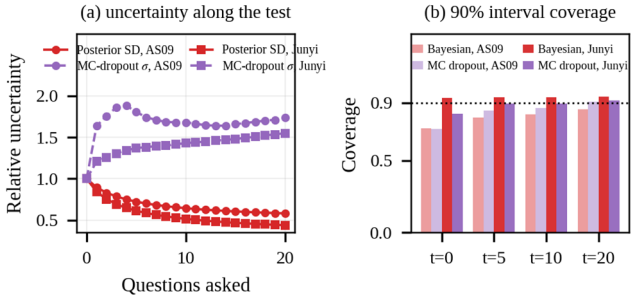


Fig 3. (a) Mean posterior SD of the general ability and mean MC-dropout uncertainty of the HARLA-style encoder along identical item sequences (normalized by their value before the first question). (b) Coverage of central 90% predictive intervals for the number of correct held-out answers (dotted line: nominal level).

Fig. 3(a) follows the two uncertainty signals along identical item sequences. The posterior SD of the general ability falls from 1 to 0.58 on ASSIST09 and to 0.44 on Junyi after 20 questions. The MC-dropout uncertainty rises instead, from 0.105 to 0.182 and from 0.146 to 0.226. A plausible reason is that every additional response passes through further stochastic units, so the spread of the sampled knowledge states grows with the length of the history; in any case the signal is smallest before any evidence has been seen. This is a property of the uncertainty signal, not of the predictions.

Table IV shows that on ASSIST09 the encoder's predictive mean is more accurate and better calibrated (ECE 0.023 versus 0.040 after 20 questions), and its 90% intervals cover 0.91 of the learners versus 0.86 for the Bayesian intervals, which are too narrow early in the test (0.80 at). On Junyi the Bayesian predictions are more accurate and their intervals are conservative (0.94). Calibration of the predictive mean and usefulness of the epistemic signal are therefore distinct properties: an encoder can predict well while its uncertainty is unusable as a measure of what remains to be learned.

Table 4. Predictive quality and calibration on identical item sequences

Data	Estimator	NLL	AUC	ECE	Cov.50	Cov.90
ASSIST09, =5	Bayesian (ours)	0.596	0.709	0.052	0.43	0.80
ASSIST09, =5	MC dropout	0.572	0.729	0.014	0.50	0.85
ASSIST09, =20	Bayesian (ours)	0.574	0.734	0.040	0.52	0.86
ASSIST09, =20	MC dropout	0.556	0.748	0.023	0.57	0.91
Junyi, =5	Bayesian (ours)	0.485	0.820	0.022	0.66	0.94
Junyi, =5	MC dropout	0.487	0.817	0.012	0.63	0.90
Junyi, =20	Bayesian (ours)	0.473	0.830	0.015	0.68	0.94
Junyi, =20	MC dropout	0.479	0.824	0.016	0.63	0.92

Cov.50/Cov.90: fraction of test learners whose number of correct held-out answers lies inside the central 50%/90% predictive interval (nominal 0.50/0.90), from 200 posterior or MC-dropout samples.

We also asked whether any signal anticipates how much a learner's held-out NLL will improve over the next five questions. Across test learners, Spearman correlations between the signal at and the realized improvement are small for the posterior SD, the largest remaining TAI and the MC-dropout uncertainty alike (). With 12 to 16 held-out items per learner on average, the realized benefit of further questions for an individual is dominated by response noise.

5.4. Termination Efficiency

Table V gives the average number of questions each termination method needs to match the held-out NLL of fixed-length tests with 5, 10 and 15 questions, and Fig. 4 shows the full trade-off. Stopping on MC-dropout uncertainty is the least efficient: it needs 8.6 (ASSIST09) and 12.3 (Junyi) questions to match a 5-question test and 13.5 and 16.4 to match a 10-question test, because learners whose sampled uncertainty is still small stop almost immediately. The posterior-SD rule also falls behind fixed length (11.7 and 12.9 questions for the 10-question target), since the posterior SD shrinks at a similar rate for all learners and largely reflects item discriminations. Thresholding the largest remaining TAI

is the most efficient threshold rule on Junyi (4.7 and 8.9 questions), but not on ASSIST09 (5.5 and 10.7), consistent with its weak correlation with realized gains there. The learned policy RL-UA matches fixed-length efficiency on both datasets (5.4 and 5.1 questions for the 5-question target, 10.4 and 9.5 for the 10-question target) without any threshold tuning. Its ablations behave similarly; the shaping term does not improve efficiency in this setting, and the unshaped variant is even slightly ahead on ASSIST09 (9.9 questions for the 10-question target). In line with the correlation analysis, the learned policies find little learner-specific saving to exploit; their advantage over threshold rules is robustness, because they never fall far behind a fixed-length test, whereas stopping rules built on the wrong uncertainty do.

Table 5. Average number of questions needed to match the held-out NLL of fixed-length tests

Termination	AS@5	AS@10	AS@15	JY@5	JY@10	JY@15
Fixed length	5.0	10.0	15.0	5.0	10.0	15.0
SE rule	5.6	11.7	15.9	6.2	12.9	17.1
Min-information (TAI) rule	5.5	10.7	15.1	4.7	8.9	14.9
MC-dropout rule	8.6	13.5	16.9	12.3	16.4	18.4
RL, point features	5.2	9.7	15.4†	5.4	9.6	–
RL, MC-dropout features	5.2	9.8	15.5†	6.8	10.5	–
RL-UA w/o shaping	4.8	9.9	14.8	4.7	9.5	–
RL-UA (ours)	5.4	10.4	16.4†	5.1	9.5	–

Lower is better; fixed length needs exactly 5, 10 or 15 questions. Values interpolate each method's accuracy–length frontier and are averaged over the three test splits; † marks targets reached in fewer than three splits, – a target never reached.

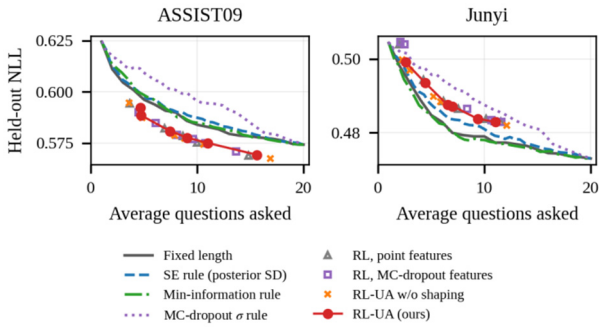


Fig 4. Accuracy–length trade-off of termination methods on the test learners (split 0 for threshold rules; RL points average three splits for each cost). Curves further down and to the left are more efficient

5.5. Why the Uncertainty-Reduction Reward Fails in Our Re-implementation

In all 6 HARLA-Op runs the validation NLL never improved on the initial policy during PPO training with its uncertainty-reduction reward, so model selection kept the initial checkpoint; for seed 0 the validation NLL drifted from 0.5641 to 0.5699 on ASSIST09 and from 0.4972 to 0.5138 on Junyi. Two effects explain this. Because the MC-dropout uncertainty grows along the test, its realized reduction is negative on average, and the reward mainly ranks items by how little they inflate the encoder's uncertainty. Because the reward is averaged over the KCs of the chosen item, it is not potential-based and carries no guarantee of preserving the predictive objective (Proposition 1). The selector of Section III-C avoids both problems by scoring the expected contraction of a posterior that does contract, and the termination reward (7) injects uncertainty only through a potential. These findings concern the operational level of HARLA-ED, on which our design builds and which we re-implement without its graph branch, meta-learning, and time, engagement and fairness rewards; the full framework is not re-evaluated here.

6. Discussion and Limitations

Three limitations qualify these findings. First, the data-driven replay protocol restricts each learner's pool to items actually answered and evaluates on small held-out sets, which keeps realized gains noisy; with large banks, live testing or simulated learners the value of further questions may be more predictable and learned termination may find savings that it could not find here. Second, the neural encoder is the better point predictor on ASSIST09, so the strongest system for such data would pair a neural predictor with a decision layer driven by an uncertainty that contracts, for example a Bayesian head on top of the encoder; our results show that MC dropout over the encoder does not provide that signal. Third, our HARLA-Op baseline re-implements the operational level only, without the graph branch, meta-learning, or the time, engagement and fairness rewards of the full framework, so our conclusions concern its uncertainty mechanism and information reward. Exposure control, content balancing and fairness across learner groups were not studied and are natural extensions of the termination reward.

7. Conclusion

We studied uncertainty-aware RL for efficient adaptive assessment from the viewpoint of the uncertainty signal itself. A calibration-aware Bayesian filter and a target-aware information criterion give item sequences that are significantly better than random sequences and, from ten questions on, than learned RL sequences under both a Bayesian and a neural estimator, and a PPO termination policy with a policy-invariant shaping reward matches fixed-length efficiency without tuning. In our re-implementation of the HARLA-ED operational agent, MC-dropout uncertainty instead grows with evidence, makes the uncertainty-reduction reward counterproductive, and makes uncertainty-based stopping 1.7 to 2.5 times less efficient than fixed-length testing. Future work will attach contracting uncertainty heads to neural knowledge encoders and evaluate learned termination in live and simulated testing with larger item banks.

References

[1] van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). Elements of adaptive testing. Springer. https://doi.org/10.1007/978-0-387-85461-8.

[2] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[3] Zhuang, Y., Liu, Q., Huang, Z., Li, Z., Shen, S., & Ma, H. (2022). Fully adaptive framework: Neural computerized adaptive testing for online education. Proceedings of the AAAI Conference on Artificial Intelligence, 36(4), 4734–4742. https://doi.org/10.1609/aaai.v36i4.20399.

[4] Wang, H., Long, T., Yin, L., Zhang, W., Xia, W., Hong, Q., Xia, D., Tang, R., & Yu, Y. (2023). GMOCAT: A graph-enhanced multi-objective method for computerized adaptive testing. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

[5] Liu, Q. (2024). Survey of computerized adaptive testing: A machine learning perspective. arXiv:2404.00712.

[6] Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward? A review of reinforcement learning for instructional sequencing. International Journal of Artificial Intelligence in Education, 29(4), 568–620. https://doi.org/10.1007/s40593-019-00187-x.

[7] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059). PMLR.

[8] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330). PMLR.

[9] Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6(4), 431–444. https://doi.org/10.1177/014662168200600405.

[10] Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. Educational and Psychological Measurement, 71(1), 37–53.

[11] Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In Proceedings of the 16th International Conference on Machine Learning (pp. 278–287).

[12] Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20(3), 213–229. https://doi.org/10.1177/014662169602000303.

[13] van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. Psychometrika, 63(2), 201–216.

[14] Segall, D. O. (1996). Multidimensional adaptive testing. Psychometrika, 61(2), 331–354.

[15] van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. Applied Measurement in Education, 13(1), 35–53.

[16] Bi, H., Ma, H., Huang, Z., Yin, Y., Liu, Q., Chen, E., Su, Y., & Wang, S. (2020). Quality meets diversity: A model-agnostic framework for computerized adaptive testing. Proceedings of the IEEE International Conference on Data Mining, 42–51.

[17] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. Advances in Neural Information Processing Systems, 28, 505–513.

[18] Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., Huang, Z., & Wang, S. (2020). Neural cognitive diagnosis for intelligent education systems. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), 6153–6161. https://doi.org/10.1609/aaai.v34i04.6080.

[19] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[20] Bickford Smith, F., Kirsch, A., Farquhar, S., Gal, Y., Foster, A., & Rainforth, T. (2023). Prediction-oriented Bayesian active learning. In Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (pp. 7331–7348). PMLR.

[21] Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (pp. 362–369).

[22] Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443–459. https://doi.org/10.1007/BF02293801.

[23] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.

[24] Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015). High-dimensional continuous control using generalized advantage estimation. arXiv:1506.02438.

[25] Fang, Q., & Liu, W. (2026). HARLA-ED: Resolving information asymmetry and enhancing algorithmic symmetry in intelligent educational assessment via hybrid reinforcement learning. Symmetry, 18(1), 58. https://doi.org/10.3390/sym18010058.

[26] Teng, D. (2025). RSRBench: Reproducibility and performance benchmarking for regional seismic risk workflows. Academic Journal of Architecture and Civil Engineering, 3(1), 40–52.

[27] Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. User Modeling and User-Adapted Interaction, 19(3), 243–266.

[28] Chang, H.-S., Hsu, H.-J., & Chen, K.-T. (2015). Modeling exercise relationships in e-learning: A unified approach. Proceedings of the 8th International Conference on Educational Data Mining.

[29] Guo, Z. (2026). Closed-Loop Audio Neuroregulation for ADHD: Graph-Based Attention Decoding and Rhythm Entrainment. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3731552.

[30] Ren, S. (2025). ARCS: Adaptive Reinforcement Learning Framework for Automated Cybersecurity Incident Response Strategy Optimization. Applied Sciences, 15, 951. https://doi.org/10.3390/app15020951.