Psychometrika (2026), **91**, 1127–1148
doi:10.1017/psy.2026.10106

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Deep Computerized Adaptive Testing

Jiguang Li[1], Robert Gibbons[2] and Veronika Ročková[1]

[1]Econometrics and Statistics, The University of Chicago Booth School of Business, USA; [2]Department of Statistics, The University of Chicago, USA

Corresponding author: Jiguang Li; Email: jiguang@chicagobooth.edu

(Received 11 July 2025; revised 17 March 2026; accepted 19 March 2026)

Abstract

Computerized adaptive tests (CATs) play a crucial role in educational assessment and diagnostic screening in behavioral health. Unlike traditional linear tests that administer a fixed set of pre-assembled items, CATs adaptively tailor the test to an examinee's latent trait level based on their previous responses. We introduce a novel CAT system that builds on recent advances in Bayesian multivariate IRT. Our approach leverages direct sampling from the latent factor posterior distributions, significantly accelerating existing information-theoretic item-selection methods by eliminating the need for computationally intensive Markov chain Monte Carlo simulations. To address the potential suboptimality of one-step-ahead item-selection rules, we also develop a double deep Q-learning algorithm that efficiently learns an optimal item-selection policy offline using a calibrated item bank. Through simulation and real-data studies, we demonstrate that our approach not only accelerates existing item-selection methods but also highlights the potential of reinforcement learning (RL) in CATs. Notably, our Q-learning-based strategy consistently achieves the fastest posterior variance reduction, leading to earlier test termination. These results demonstrate the promise of combining exact posterior sampling with RL to deliver scalable, high-precision CATs.

Keywords: adaptive data collection; Bayesian methods; deep Q-learning; multidimensional item response theory

1. Introduction

Multidimensional computerized adaptive testing (MCAT) has revolutionized the field of educational and psychological assessments by dynamically selecting tailored items from a large test pool, thereby enhancing the efficiency and precision of latent ability estimates (van der Linden & Glas, 2010). Powered by multidimensional item response theory (MIRT; Bock & Gibbons, 2021), MCAT leverages multidimensional statistical inference to evaluate respondents' multidimensional latent traits and allows for more comprehensive and efficient assessments compared to unidimensional approaches. MCAT's adaptability and accuracy are also particularly crucial in high-stakes diagnostic assessments, such as in clinical psychology and psychiatry, where it can be substituted for in-person assessments by clinical professionals, especially in areas with limited medical resources (Gibbons et al., 2016).

A substantial body of research in MCAT has focused on item-selection strategies derived from experimental design principles. One prominent strategy is to select items that maximize the determinant of the Fisher information matrix evaluated at the current estimates of the latent traits (Segall, 1996, 2000), known as the D-optimality criterion. An alternative, the A-optimality criterion, aims to minimize

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

the trace of the asymptotic covariance matrix, thereby reducing overall estimation variance (van der Linden, 1999). In the absence of nuisance latent abilities, both A-optimality and D-optimality have demonstrated superior accuracy relative to other common experimental design criteria (Mulder & van der Linden, 2009).

Another prominent line of MCAT item-selection algorithms leverages Kullback–Leibler (KL) information, often within a Bayesian framework. A common approach is to select items that produce response distributions at the true latent trait value, $\boldsymbol{\theta}_0$, that differs maximally from the response distributions generated at the other value of θ (Chang & Ying, 1996; Veldkamp & van der Linden, 2002). Moving beyond response distributions alone, some researchers propose maximizing the KL divergence between the current posterior distribution and the posterior distribution at the next selection step, thereby enhancing adaptation through updated trait estimates (Mulder & van der Linden, 2010). Another promising strategy is the mutual information (MI) criterion, which aims to maximize entropy reduction of the current posterior distribution, encouraging more and more accurate posterior estimates (Weissman, 2007). In particular, Wang and Chang (2011) demonstrate both theoretical and empirical advantages of the Bayesian MI item-selection rule over common experimental criteria such as D-optimality. More detailed theoretical comparison of KL information and Fisher information criteria is presented by Wang et al. (2011).

Although numerous effective item-selection rules have been proposed in the MCAT literature, they all rely on one-step lookahead optimization of an information-theoretic criterion. Despite their ease of implementation and attractive theoretical properties, these selection rules are inherently myopic: they select items based solely on immediate information gain, ignoring how current choices influence future decisions, which can lead to suboptimal policies. For example, existing methods tend to favor items with high loading parameters (Chang, 2015). However, CAT researchers also recommend reserving such items for the later stages of testing to improve efficiency (Chang & Ying, 1999). Integrating heuristic guidance into existing selection rules remains challenging.

Addressing these limitations, we propose a novel deep CAT system that integrates a flexible Bayesian MIRT model with a non-myopic online item-selection policy, guided by reinforcement learning (RL) principles (Sutton & Barto, 2018). Leveraging recent advancements in Bayesian sparse MIRT (Li et al., 2025), our framework seamlessly accommodates multiple latent factors with complex loading structures, while maintaining scalability in both the number of items and factors. To learn the optimal item-selection policy that prioritizes the assessment of target factors, we draw on contemporary RL methodologies and introduce a general double deep Q-learning algorithm (Mnih et al., 2015; van Hasselt et al., 2016). This algorithm efficiently trains a Q-network offline using only item parameter estimates; the learned network can then be deployed online to select optimal items based on the current multivariate latent factor posterior distribution. When the test terminates, our framework robustly characterizes the full latent factor posterior distributions rather than providing only a point estimate.

A primary contribution of our work is to show how the identification of the latent factor posterior distribution leads to substantial computational gains in online item selection. Given that such posterior distribution is deemed to be non-Gaussian and unknown, traditional Bayesian methods typically rely on expensive Markov chain Monte Carlo (MCMC) simulations (Béguin & Glas, 2001) and combined with additional data augmentation to handle categorical likelihood (Albert & Chib, 1993; Polson et al., 2013). Our approach achieves substantial acceleration by directly sampling latent factor posterior distributions (Botev, 2016; Durante, 2019; Li et al., 2025). Notably, this improvement not only increases the efficiency of existing Bayesian item-selection procedures but also provides a computational foundation for training our proposed Q-learning algorithm through rapid, large-scale simulations of testing sessions.

Another critical advancement in our work is the integration of CAT within an RL framework. This approach addresses the practical need to prioritize accurate estimation and to overcome known limitations of greedy item-selection methods in sequential decision-making (Bertsekas & Tsitsiklis, 1996). Building on the Bayesian MIRT foundation, our neural-network architecture incorporates the identified posterior parameters as state variables, allowing the model to learn optimal item-selection policies through a large amount of testing simulations. This formulation bridges the two paradigms: the

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Bayesian component provides statistically grounded representations of examinee uncertainty, while the RL component leverages these representations to optimize sequential decisions. The trained neural network is deployable on standard laptops without GPU acceleration, making it suitable for online adaptive testing applications. The sequential nature of CAT aligns naturally with deep Q-learning methods, which have demonstrated remarkable success across diverse application domains (Kalashnikov et al., 2018; Silver et al., 2016). Notably, RL has been successfully employed in educational measurement settings to design personalized learning plans (Li et al., 2023; Tan et al., 2020).

Finally, our work aligns with the emerging research trend of framing traditional statistical sequential decision-making problems as optimal policy learning tasks (Rainforth et al., 2024). This perspective has appeared in Bayesian adaptive design (Chaloner & Verdinelli, 1995; Foster et al., 2021; Sebastiani & Wynn, 2000) and Bayesian optimization (Lam et al., 2016; Srinivas et al., 2010), where recent methods aim to move beyond one-step criteria toward policy learning that account for long-term consequences. Our contribution fits within this broader trend by providing a principled approach to cognitive and behavioral assessment.

The remainder of the article is organized as follows. Section 2 motivates our deep CAT framework using a cognitively complex item bank from a recent clinical study (Gibbons et al., 2024). Section 3 reviews existing information-theoretic item-selection methods and outlines the necessity to reformulate CAT as an RL task. Section 4 introduces a general Bayesian framework that accelerates existing CAT item-selection rules and serves as the foundation for our Q-learning algorithm. Section 5 details our RL approach, including the neural network architecture and double Q-learning algorithm. Finally, Sections 6 and 7 evaluate our method through both simulations and real data experiments.

2. Adaptive high-dimensional cognitive assessment

Our proposed deep CAT system is motivated by the growing need for adaptive cognitive assessment in high-dimensional latent spaces. Cognitive impairment, particularly Alzheimer's dementia (AD), is a major public health challenge, affecting 6.9 million individuals in the United States in 2024 (Alzheimer's Association, 2024). Early detection of mild cognitive impairment (MCI), a precursor to AD, is crucial for slowing disease progression and improving patient outcomes (Huang et al., 2018). However, traditional neuropsychological assessments are costly, time-consuming, and impractical for frequent use, underscoring the need for more efficient and scalable assessment methods.

Recently, pCAT-COG, the first computerized adaptive test (CAT) item bank based on MIRT for cognitive assessment, demonstrated its potential as an alternative to clinician-administered evaluations (Gibbons et al., 2024). The data were collected from 730 participants. After careful item calibration and model selection, the final item bank consisted of $J = 57$ items covering five cognitive subdomains: episodic memory, working memory, executive function, semantic memory, and processing speed. Since each item comprised three related tasks, we used a binary score, where 1 indicates correct answers on all tasks for our analysis.

Following Gibbons et al. (2024), we fit a six-factor bifactor model (Gibbons & Hedeker, 1992) to the 730 by 57 binary response matrix, with one primary factor representing the global cognitive ability and five secondary factors. This yields a 57 by 6 factor loading matrix, visualized in Figure 1, where rows correspond to pCAT-COG items and columns represent distinct factors.

Since items in pCAT-COG are primarily designed to measure the general cognitive factor and only partially capture subdomain information, our goal is to develop an item-selection strategy that efficiently estimates the primary factor (first column) using as few items as possible while maintaining robustness to subdomain influences. Additionally, the selection algorithm must be computationally efficient to navigate a six-dimensional latent space in real time for seamless interactive testing. Online item selection is critical, as learning an optimal sequence from 57 items offline presents an intractable combinatorial problem, even in a binary response setting.

Adaptive cognitive assessment is essential for cognitive assessment, as designing items is costly and administering all items is time-consuming. Our deep CAT system requires on average 11.2 items to

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

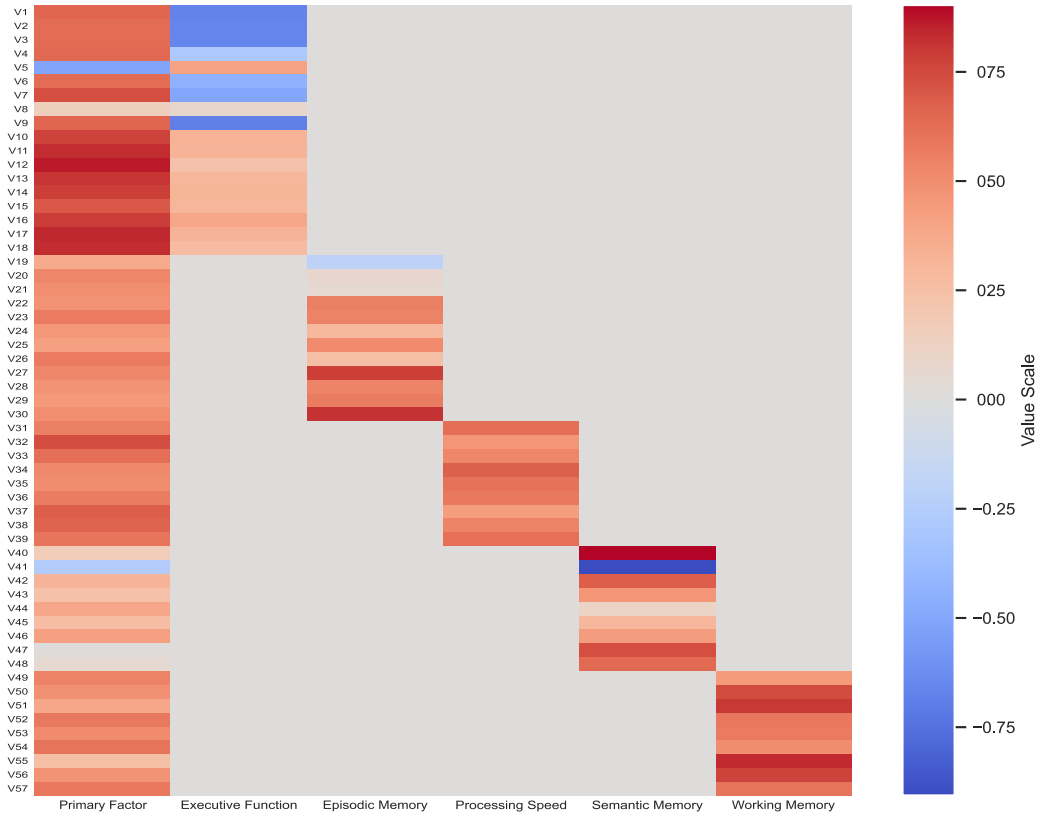


Figure 1. Estimated bifactor factor loading matrix for pCAT-COG.

reduce the posterior variance of the primary factor from 1 to 0.16 (posterior s.d. = 0.4), whereas the next-best MI method requires 12.6 items to achieve the same precision. The item bank is currently expanding to 500 items with a more nationally representative sample. Success on this prototype dataset paves the way for broader deployment in clinical research.

3. From one-step optimization to policy learning

We formally define the problem of designing a CAT system from an RL perspective. Typical CAT systems consist of two components:

- *Offline calibration*: An MIRT model is fitted to a calibrated dataset $\boldsymbol{Y} \in \mathbb{R}^{N\times J}$ to estimate the item characteristic parameters, where N represents the number of examinees, and J represents the number of items in the item bank.
- *Online deployment*: Given the estimated item parameters, an item-selection algorithm is deployed online to adaptively select items for future examinees.

The performance of CAT can be measured by the number of items required to estimate an examinee's latent traits with sufficient precision.

3.1. Notation and problem formulation

For the entirety of the article, we assume that the calibration dataset $\boldsymbol{Y}$ is binary, where each element y_{ij} represents whether subject i answered item j correctly. We consider a general two-parameter MIRT framework with K latent factors (Bock & Gibbons, 2021). Let $\boldsymbol{B} \in \mathbb{R}^{J\times K}$ denote the factor loading matrix,

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

and $\boldsymbol{D} \in \mathbb{R}^J$ denote the intercept vector. Throughout the article, boldface notation is used to denote vectors and matrices (e.g., $\boldsymbol{\theta}$,$\boldsymbol{B}$, and $\boldsymbol{D}$), while scalar quantity, such as D_j, is written in standard font. For each examinee i with multivariate latent trait $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^K$, the data-generating process for y_{ij} is given by

$$y_{ij} \mid \boldsymbol{\theta}^{(i)}, \boldsymbol{B}_j, D_j \sim \text{Bernoulli}(\Phi(\boldsymbol{B}_j' \boldsymbol{\theta}^{(i)} + D_j)), \tag{3.1}$$

where $\boldsymbol{B}_j$ is the jth row of $\boldsymbol{B}$, D_j is the jth entry of $\boldsymbol{D}$, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. The item parameters can be compactly expressed as $\{\boldsymbol{\xi}_j\}_{j=1}^J := \{(\boldsymbol{B}_j, D_j)\}_{j=1}^J$. This two-parameter MIRT framework is highly general, imposing no structural constraints on the loading matrix and requiring no specific estimation algorithms for item parameter calibration.

Given the estimated item parameters, we need to design an item-selection algorithm that efficiently tests a future examinee with unobserved latent ability $\boldsymbol{\theta} \in \mathbb{R}^K$. The sequential nature of the CAT problem makes Bayesian approaches particularly appealing. Without loss of generality, we assume a standard multivariate Gaussian prior on $\boldsymbol{\theta} \sim \mathcal{N}(0, \mathbb{I}_K)$ and introduce the following notation:

- For any positive integer n, let $[n]$ denote all the positive integers no greater than n. Let $j \in [J]$ be the index for the items in the item bank.
- Let j_t denote the index of the item selected at step t based on an arbitrary item-selection rule. Define $\mathcal{I}_t$ as the set of the first t administered items, and let $R_t := [J] \backslash \mathcal{I}_{t-1}$, the set of available items before the tth item is picked.
- We use the shorthand notation $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:T})$ to represent the latent factor posterior distributions $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:T}, \boldsymbol{\xi}_{1:T})$ after T items have been selected. Here, the response history is denoted by $\boldsymbol{Y}_{1:T} := [y_{j_1}, \ldots, y_{j_T}]'$, and the item parameters are given by $\boldsymbol{\xi}_{1:T} := (\boldsymbol{B}_{1:T}, \boldsymbol{D}_{1:T})$, where $\boldsymbol{B}_{1:T} := [\boldsymbol{B}_{j_1}, \ldots, \boldsymbol{B}_{j_T}]'$ and $\boldsymbol{D}_{1:T} := [D_{j_1}, \ldots, D_{j_T}]'$.

At time $(t+1)$, the algorithm takes the current posterior $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t})$ as input and outputs the next item selection j_{t+1}. This process continues until at time T', either when the posterior variance of $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:T'})$ falls below a predefined threshold τ^2, or when $T' = H$, where $H \leq J$ is the maximum number of items that can be administered. Hence, the goal of CAT is to minimize T' given the estimated item parameters of the item bank.

3.2. Reviews of KL information item-selection rules

Given that our proposed deep CAT system is built on a general Bayesian MIRT framework (Li et al., 2025), we briefly revisit common Bayesian item-selection rules and show in Section 4.1 how our framework can be used to accelerate these baseline methods. The popular KL expected a priori (EAP) rule selects the tth item based on the average KL information between response distributions on the candidate item at the EAP estimate $\hat{\boldsymbol{\theta}}_{t-1} = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) d\boldsymbol{\theta}$, and random factor $\boldsymbol{\theta}$ sampled from the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$ (Veldkamp & van der Linden, 2002):

$$\arg\max_{j_t \in R_t} \int_{\boldsymbol{\theta}} \left\{ \sum_{l=0}^{1} P(y_{j_t} = l|\hat{\boldsymbol{\theta}}_{t-1}) \log \frac{P(y_{j_t} = l|\hat{\boldsymbol{\theta}}_{t-1})}{P(y_{j_t} = l|\boldsymbol{\theta})} \right\} f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) d\boldsymbol{\theta}. \tag{3.2}$$

Rather than focusing on the KL information based on the response distributions, Mulder and van der Linden (2010) propose the MAX Pos approach by maximizing the KL information between two subsequent latent factor posteriors $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$ and $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t})$. Intuitively, this approach prioritizes items that induce the largest shift in the posterior, formalized as

$$\arg\max_{j_t \in R_t} \sum_{y_{j_t}=0}^{1} f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)}) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) \log \frac{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})}{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t})} d\boldsymbol{\theta}, \tag{3.3}$$

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

where $f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)})$ represents the posterior predictive probability:

$$f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)}) = \int_{\boldsymbol{\theta}} f(y_{j_t}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) d\boldsymbol{\theta}. \tag{3.4}$$

A third Bayesian strategy, the MI approach, maximizes the MI between the current posterior distribution and the response distribution of new item y_{j_t} (Weissman, 2007). Rooted in experimental design theory (Rényi, 1961), we can also interpret MI as the entropy reduction of the current posterior distribution of $\boldsymbol{\theta}$ after observing new response y_{j_t}. More formally,

$$\arg\max_{j_t \in R_t} I_M(\boldsymbol{\theta}, y_{j_t}) = \arg\max_{j_t \in R_t} \sum_{y_{j_t}=0}^{1} \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, y_{j_t}|\boldsymbol{Y}_{1:(t-1)}) \log \frac{f(\boldsymbol{\theta}, y_{j_t}|\boldsymbol{Y}_{1:(t-1)})}{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)})} d\boldsymbol{\theta}. \tag{3.5}$$

Given the impressive empirical success of the MI method (Wang & Chang, 2011), we also introduce a competitive heuristic item-selection rule that chooses items with the highest predictive variance under the current posterior estimates, serving as an additional benchmark method. Formally, write $c_{j_t} := f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)})$ as defined in (3.4), and consider the following criterion:

$$\arg\max_{j_t \in R_t} \int_{\boldsymbol{\theta}} (\Phi(\boldsymbol{B}'_{j_t}\boldsymbol{\theta} + D_{j_t}) - c_{j_t})^2 f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) d\boldsymbol{\theta}. \tag{3.6}$$

This rule selects the item j_t that maximizes the variance of the predictive means, weighted by the current posterior $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$. In Appendix A of the Supplementary Material, we establish a connection between this selection rule and the MI method, showing that both favor items with high prediction uncertainty.

3.3. Reinforcement learning formulation

Existing item-selection rules face several limitations: first, these methods rely on one-step-lookahead selection, choosing items that provide the most immediate information without considering their impact on future selections, which can result in suboptimal policies (Sutton & Barto, 2018). Second, they are heuristically designed to balance information across all latent factors, rather than emphasizing the most essential factors of interest. Finally, these rules do not directly minimize test length, potentially increasing test duration without proportional gains in accuracy.

A more principled approach is to formulate item selection as an RL problem, where the optimal policy is learned using Bellman optimality principles rather than relying on heuristics that ignore long-term planning. Beyond addressing myopia, an RL-based formulation provides a direct mechanism to minimize the number of items required to reach a predefined posterior variance reduction threshold, explicitly guiding item selection toward accurately measuring the primary factors of interests.

More generally, consider a general finite horizon setting where each examinee can answer at most $H \leq J$ items, and the CAT algorithm terminates whenever the posterior variance of all factors of interest is smaller than a certain threshold τ^2, or when H is reached. Since H can be set sufficiently large, it serves as a practical secondary stopping criterion to prevent excessively long tests. Formally:

- **State space** $\mathcal{S}$: The space of all possible latent factor posteriors $\mathcal{S} := \{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t}) : t \in \{1, \ldots, H\}\}$. The state variable in CAT represents the Bayesian estimate of the examinee's latent traits as a multivariate distribution at time t. We may write $s_t := f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t})$, with $s_0 := \mathcal{N}(0, \mathbb{I}_K)$. The state space is discrete and can be exponentially large. Since there are J items in the item bank and the responses are binary, the total number of possible states is $\sum_{t=1}^{H} \binom{J}{t} \cdot 2^t$.
- **Action space** $\mathcal{A}$: The item bank $[J] := \{1, \ldots, J\}$. However, at the tth selection time, the action space is the remaining items in the test bank that have not yet been selected $R_t := [J] \backslash \mathcal{I}_{t-1}$, as we do not select the same item twice.
- **Transition kernel** $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the space of probability distributions over $\mathcal{S}$. The transition kernel $\mathcal{P}$ specifies the stochastic rule that governs how the state evolves

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

after an action is taken. In the CAT context, the kernel maps the current latent-trait estimate s_t and the selected item a_t to the next posterior state s_{t+1}. Conditional on (s_t, a_t), the next estimate s_{t+1} depends on the examinee's binary response $y_t \in \{0,1\}$ to item a_t, which yields only two possible posterior updates. Because the true ability of the examinee and thus the response probability $p(y_t = 1 \mid s_t, a_t)$ are unknown, the transition kernel in CAT is also unknown and must be approximated.

- **Reward function**: To directly minimize the test length, we assign a simple 0–1 reward structure, where we assign negative rewards whenever more items are needed to reduce posterior variance to a given threshold. In a K-factor MIRT model, we often prioritize a subset of factors $\mathcal{K} \subset [K]$ (e.g., $\mathcal{K} = \{1\}$ for pCAT-COG), with the test terminating once the posterior variances of all factors in $\mathcal{K}$ fall below the predefined threshold τ^2:

$$R^{(t)}(s_t, a, s_{t+1}) = \begin{cases} -1 & \text{if } V_{t+1} > \tau^2, \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

where $V_{t+1} = \max_{k \in \mathcal{K}} \mathrm{Var}(\boldsymbol{\theta}_k \mid \boldsymbol{Y}_{1:(t+1)})$ is the maximum marginal posterior variance among the prioritized factors $\mathcal{K}$. This also simplifies the learning of the value function, as the rewards are always bounded integers within $[-H, -1]$. We further illustrate the advantages of adopting this reward structure in Appendix E of the Supplementary Material.

Rather than following a heuristic rule, we learn a policy π that maps the current state (latent factor estimates) to a distribution over potential items:

$$\pi : \mathcal{S} \to \Delta(\mathcal{A}), \qquad a_t \sim \pi(\cdot \mid s_t).$$

Given an initial state $S_0 = s_0$ and the discount factor $\gamma \in (0,1]$, the value function is

$$v_\pi(s_0) := \mathbb{E}_\pi \left[\sum_{t=0}^{H-1} \gamma^t R^{(t)}(s_t, \pi(s_t), s_{t+1}) | S_0 = s_0 \right]. \tag{3.8}$$

The expectation $\mathbb{E}_\pi[\cdot]$ is taken over trajectories generated by drawing a_t from $\pi(\cdot \mid s_t)$ and then s_{t+1} from the transition kernel $\mathcal{P}(\cdot \mid s_t, a_t)$. Because direct evaluation of (3.8) is challenging, it is more convenient to consider the action-value function

$$Q_\pi(s_0, a) := \mathbb{E}_\pi \left[\sum_{t=0}^{H-1} \gamma^t R^{(t)}(s_t, \pi(s_t), s_{t+1}) | S_0 = s_0, A_0 = a \right], \tag{3.9}$$

with $v_\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} Q_\pi(s,a)$. The optimal policy $\pi^\star$ satisfies the Bellman equation (Bertsekas & Tsitsiklis, 1996):

$$Q_{\pi^\star}(s,a) = \mathbb{E}_{S' \sim \mathcal{P}(\cdot|s,a)} \left[R(s,a,s') + \gamma \max_{a'} Q_{\pi^\star}(s', a') \right], \tag{3.10}$$

where s' is the next posterior estimate after applying action a to the current posterior s, and $v_{\pi^\star}(s) = \max_a Q_{\pi^\star}(s,a)$. Since the state space grows exponentially and the transition kernel is unknown, solving for $\pi^\star$ using traditional dynamic programming approaches becomes intractable (Sutton & Barto, 2018).

In practice, our deep Q-learning approach (Mnih et al., 2013) does not evaluate these expectations analytically. Let the transition (s, a, r, s') represent one testing step, where item a is selected under state s, the reward r is observed, and the posterior is updated to s' through the Bayesian MIRT model. We approximate the Bellman fixed point in (3.10) by fitting a parametric function Q_w that minimizes the squared temporal-difference loss:

$$\mathcal{L}(w) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\left(Q_w(s,a) - y(s,a,r,s') \right)^2 \right], \tag{3.11}$$

where

$$y(s,a,r,s') = r + \gamma \max_{a'} Q_{\bar{w}}(s', a'). \tag{3.12}$$

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Here, $\mathcal{D}$ (replay buffer) denotes a collection of previously observed testing steps (s,a,r,s') generated during simulation, which is used to approximate the expectation above by empirical averaging. The parameter $\bar{w}$ corresponds to a delayed copy of the model parameters w, updated less frequently to stabilize the numerical optimization. We describe the full Q-learning algorithm in Section 5.

4. Accelerating item selection via posterior identification

The central insight of our deep CAT framework is that, by iteratively applying the E-step of the PXL-EM algorithm (Li et al., 2025), the latent factor posteriors admit tractable posterior updates. This result is critical for RL, as the posterior distribution is deemed to be non-Gaussian and is analytically intractable under the traditional MIRT literature. By obtaining a tractable representation of this posterior, we can parameterize the examinee's evolving ability and uncertainty as a well-defined state variable, which is an essential prerequisite for applying Q-learning to adaptive testing. The Bayesian MIRT formulation thus not only replaces costly MCMC procedures with efficient posterior updates under a probit link for existing Bayesian item-selection rules discussed in Section 3.2, but also supplies the statistical foundation that makes the subsequent RL framework feasible.

Specifically, we show that the latent factor posterior updates during CAT belong to an instance of the unified skew-normal distribution (Arellano-Valle & Azzalini, 2006), defined as follows.

Definition 4.1. Let $\Phi_T\{\boldsymbol{V};\boldsymbol{\Sigma}\}$ represent the cumulative distribution function of a T-dimensional multivariate Gaussian distribution $N_T(0_T,\boldsymbol{\Sigma})$ evaluated at vector $\boldsymbol{V}$. A K-dimensional random vector $\boldsymbol{\theta} \sim \mathrm{SUN}_{K,T}(\boldsymbol{\mu},\boldsymbol{\Omega},\boldsymbol{\Delta},\boldsymbol{\gamma},\boldsymbol{\Gamma})$ has the *unified skew-normal* distribution if it has the probability density function:

$$\phi_K(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\Omega})\frac{\Phi_T\left\{\boldsymbol{\gamma}+\boldsymbol{\Delta}'\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu});\boldsymbol{\Gamma}-\boldsymbol{\Delta}'\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta}\right\}}{\Phi_T(\boldsymbol{\gamma};\boldsymbol{\Gamma})}.$$

Here, $\phi_K(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\Omega})$ is the density of a K-dimensional multivariate Gaussian with expectation $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_K)'$, and a K by K covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$, where $\bar{\boldsymbol{\Omega}}$ is the correlation matrix and $\boldsymbol{\omega}$ is a diagonal matrix with the square roots of the diagonal elements of $\boldsymbol{\Omega}$ in its diagonal. $\boldsymbol{\Delta}$ is a K by T matrix that determines the skewness of the distribution, and $\boldsymbol{\gamma} \in \mathbb{R}^T$ controls the flexibility in departures from normality.

In addition, the $(K+T)\times(K+T)$ matrix $\boldsymbol{\Omega}^*$, having blocks $\boldsymbol{\Omega}^*_{[11]} = \boldsymbol{\Gamma}, \boldsymbol{\Omega}^*_{[22]} = \bar{\boldsymbol{\Omega}}$, and $\boldsymbol{\Omega}^*_{[21]} = \boldsymbol{\Omega}^{*'}_{[12]} = \boldsymbol{\Delta}$, needs to be a full-rank correlation matrix.

Suppose an arbitrary CAT item-selection algorithm has already selected T items with item parameters $\boldsymbol{\xi}_{1:T} = (\boldsymbol{B}_{1:T},\boldsymbol{D}_{1:T})$, where $\boldsymbol{B}_{1:T} := [\boldsymbol{B}_{j_1},\ldots,\boldsymbol{B}_{j_T}]'$ and $\boldsymbol{D}_{1:T} := [D_{j_1},\ldots,D_{j_T}]'$. Then it is possible to show the following result.

Theorem 4.1. *Consider a K-factor CAT item-selection procedure after selecting T items, with $\mathcal{N}(0_K,\mathbb{I}_K)$ prior placed on the test taker's latent trait $\boldsymbol{\theta}$. If $\boldsymbol{Y}_{1:T} = (y_{j_1},\ldots,y_{j_T})'$ is conditionally independent binary response data from the two-parameter probit MIRT model defined in (3.1), then*

$$(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:T},\boldsymbol{B}_{1:T},\boldsymbol{D}_{1:T}) \sim \mathrm{SUN}_{K,T}\left(\boldsymbol{\mu}_{\text{post}},\boldsymbol{\Omega}_{post},\boldsymbol{\Delta}_{post},\boldsymbol{\gamma}_{post},\boldsymbol{\Gamma}_{post}\right),$$

with posterior parameters

$$\boldsymbol{\mu}_{\text{post}} = 0_K, \quad \boldsymbol{\Omega}_{post} = \mathbb{I}_K, \quad \boldsymbol{\Delta}_{post} = \boldsymbol{C}_1'\boldsymbol{C}_3^{-1},$$
$$\boldsymbol{\gamma}_{post} = \boldsymbol{C}_3^{-1}\boldsymbol{C}_2, \quad \boldsymbol{\Gamma}_{post} = \boldsymbol{C}_3^{-1}\left(\boldsymbol{C}_1\boldsymbol{C}_1' + \mathbb{I}_T\right)\boldsymbol{C}_3^{-1},$$

where $\boldsymbol{C}_1 = diag(2y_{j_1}-1,\ldots,2y_{j_T}-1)\boldsymbol{B}_{1:T}$ and $\boldsymbol{C}_2 = diag(2y_{j_1}-1,\ldots,2y_{j_T}-1)\boldsymbol{D}_{1:T}$. The matrix $\boldsymbol{C}_3$ is a T by T diagonal matrix, where the (t,t)th entry is $(\|\boldsymbol{B}_{t,T}\|_2^2+1)^{\frac{1}{2}}$, with $\boldsymbol{B}_{t,T}$ representing the tth row of $\boldsymbol{B}_{1:T}$.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Theorem 4.1 provides an exact finite-sample Bayesian characterization of the latent factor posterior. With a multivariate normal prior on $\boldsymbol{\theta}$ and a probit MIRT likelihood, the posterior $f(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:T})$ belongs to the unified skew-normal (SUN) family. This does not conflict with the empirical Bayes result of Chang and Stout (1993), which establishes asymptotic posterior normality as the number of items J grows; related discussion of Bayesian latent-trait estimation and asymptotic covariance properties in MIRT is given in Wang (2015). The finite-sample form is particularly relevant for CAT, where only a small number of items have been administered and large-sample normal approximations may be unreliable. This representation enables exact posterior calculations, improving uncertainty quantification and item selection in short tests. For ordinal responses under a probit link, closely related SUN-based approximations can also be developed as discussed in Li et al. (2025). The latent factor posterior can often be well approximated within the SUN family, although the representation is generally not exact.

According to Arellano-Valle and Azzalini (2006), an arbitrary unified skew-normal distribution $\boldsymbol{\theta} \sim \mathrm{SUN}_{K,T}(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ has a stochastic representation as a linear combination of a K-dimensional multivariate normal random variable $\boldsymbol{V}_0$, and a T-dimensional truncated multivariate normal random variable $\boldsymbol{V}_{1,-\gamma}$ as follows:

$$\boldsymbol{\theta} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{\omega}(\boldsymbol{V}_0 + \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{V}_{1,-\gamma}), \tag{4.1}$$

where $\boldsymbol{V}_0 \sim \mathcal{N}(0, \bar{\boldsymbol{\Omega}} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}') \in \mathbb{R}^K$, and $\boldsymbol{V}_{1,-\gamma}$ is a multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Gamma})$ truncated to $\{\boldsymbol{V}_1 \in \mathbb{R}^T : V_{1i} \geq -\gamma_i \quad, \forall i\}$.

Based on Theorem 4.1, we have a closed-form expression for the posterior parameters $(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ for $\boldsymbol{\theta}$. Recall that $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$ is the standard covariance correlation matrix decomposition from Definition 4.1, and hence the only unrealized stochastic terms in Equation (4.1) are $\boldsymbol{V}_0$ and $\boldsymbol{V}_{1,-\gamma}$. This suggests that sampling from the latent factor posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:T})$ requires two independent steps: sampling from a K-dimensional multivariate normal distribution $\boldsymbol{V}_0$ and a T-dimensional multivariate truncated normal distribution $\boldsymbol{V}_1$. As a result, the direct sampling approach scales efficiently with the number of factors K, since generating samples from K-dimensional multivariate normal distributions is trivial. Moreover, in CAT settings, the test is typically terminated early, meaning T remains relatively small. When T is moderate (e.g., $T < 1{,}000$), sampling from the truncated multivariate normal distribution remains computationally efficient using the minimax tilting method (Botev, 2016). The exact proof of Theorem 4.1 and the sampling details can be found in Appendix B of the Supplementary Material.

The direct sampling approach provides substantial gains in both computational efficiency and numerical precision compared with traditional MCMC methods commonly used in the MIRT literature (Béguin & Glas, 2001; Jiang & Templin, 2019). In standard MIRT settings, the posterior distribution of the latent factors is typically regarded as intractable and non-Gaussian, requiring a Markov chain to be constructed via data-augmentation techniques (Albert & Chib, 1993; Polson et al., 2013) so that its stationary distribution approximates the posterior. This procedure entails repeated simulation of augmented data and is inherently sequential, which limits opportunities for parallelization. Moreover, it demands additional tuning, burn-in, and convergence diagnostics to ensure that the chain adequately converges to the posterior distribution. In contrast, Theorem 4.1 establishes that the posterior distribution can be expressed exactly as a unified skew-normal distribution, allowing direct and parallel sampling without the need for iterative convergence procedures.

Theorem 4.1 also plays a central role in our proposed deep Q-learning algorithm, as it fully characterizes the latent factor posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:T})$, enabling the parameterization of the state variable s_T, which serves as an input to the Q-network illustrated in Section 5.1. We illustrate how Theorem 4.1 can be used to accelerate existing item-selection rules, and the same idea can be extended to fully Bayesian item-selection settings as considered by van der Linden and Ren (2020).

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

4.1. Accelerating existing rules

While Bayesian CAT criteria require multidimensional integration, these quantities are typically evaluated in practice via Monte Carlo approximation rather than analytic quadrature in moderate to high dimensions. In such Bayesian CAT implementations, the dominant computational cost lies in obtaining valid samples from the latent factor posterior. By providing exact posterior samples without Markov chain construction, our approach removes this primary bottleneck and renders the remaining Monte Carlo integration computationally efficient and easily parallelizable. For example, in computing the KL-EAP item-selection rule (3.2), we directly sample from $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$ instead of resorting to MCMC and evaluate the integral via Monte Carlo integration. Since the EAP estimate $\hat{\boldsymbol{\theta}}_{t-1}$ remains fixed at time step t, evaluating the KL information term is straightforward. For the Max Pos item-selection rule in Equation (3.3), we again obtain i.i.d. samples from $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$ and compute the posterior predictive probabilities in (3.4). Although the density ratio $\frac{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})}{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t})}$ is difficult to evaluate, we can leverage the conditional independence assumption of the MIRT model. In particular, we can express the joint distribution of $(\boldsymbol{\theta}, y_{j_t})$ as

$$f(\boldsymbol{\theta}, y_{j_t}|\boldsymbol{Y}_{1:(t-1)}) = f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t}) f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)}) = f(y_{j_t}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}). \tag{4.2}$$

Using Equation (4.2), we rewrite the KL information term in Equation (3.3) as

$$\begin{aligned}\int f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) \log \frac{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})}{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t})} d\boldsymbol{\theta} &= \int f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) \log \frac{f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)})}{f(\boldsymbol{\theta}, y_{j_t}|\boldsymbol{Y}_{1:(t-1)})} d\boldsymbol{\theta} \\ &= \int f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) \log \frac{f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)})}{f(y_{j_t}|\boldsymbol{\theta})} d\boldsymbol{\theta}.\end{aligned}$$

Since $f(y_{j_t}|\boldsymbol{\theta})$ can be easily computed for each $\boldsymbol{\theta} \sim f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$, the online computation of Max Pos remains efficient.

Although the MI selection rule in Equation (3.5) has demonstrated strong empirical performance (Wang & Chang, 2011), its computational complexity remains a significant challenge. By applying Equation (4.2), we can rewrite MI as

$$\arg\max_{j_t \in R_t} \sum_{y_{j_t}=0}^{1} f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)}) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t}) \log \frac{f(y_{j_t}|\boldsymbol{\theta})}{f(y_{j_t}|\boldsymbol{Y}_{1:(t-1)})} d\boldsymbol{\theta}. \tag{4.3}$$

This formulation reveals that maximizing MI is structurally similar to Max Pos but requires much more computational effort. Unlike Max Pos, where sampling is performed from the current posterior $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$, MI requires sampling from future posteriors $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t})$. Since each candidate item $j_t \in R_t$ has two possible outcomes ($y_{j_t} = 0$ or $y_{j_t} = 1$), evaluating Equation (4.3) requires obtaining samples from $|R_t| \times 2$ distinct posterior distributions. Even if sampling each individual posterior is computationally efficient, this approach becomes impractical for large item banks.

We hence propose a new approach to dramatically accelerate the computation of the MI quantity using the idea of importance sampling and resampling and bootstrap filter (Gordon et al., 1993; Smith & Gelfand, 1992). Rather than explicitly sampling the future posterior $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t})$ for each $j_t \in R_t$ and $y_{j_t} \in \{0, 1\}$, we can simply sample from the current posterior $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$ once, and then perform proper posterior reweighting. Under Equation (3.1), let $p(\boldsymbol{\theta})$ be the prior on the latent factors $\boldsymbol{\theta}$, $l_1(\boldsymbol{\theta})$ be the current data likelihood, and $l_2(\boldsymbol{\theta})$ denote the future data likelihood after observing y_{j_t}. We have the future posterior density as follows:

$$\begin{aligned}f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t}) &\propto l_2(\boldsymbol{\theta}) p(\boldsymbol{\theta}) \propto \frac{l_2(\boldsymbol{\theta}) p(\boldsymbol{\theta})}{l_1(\boldsymbol{\theta}) p(\boldsymbol{\theta})} f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}) \\ &= \Phi(\boldsymbol{B}_{j_t}' \boldsymbol{\theta} + D_{j_t})^{y_{j_t}} (1 - \Phi(\boldsymbol{B}_{j_t}' \boldsymbol{\theta} + D_{j_t}))^{1-y_{j_t}} f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}).\end{aligned} \tag{4.4}$$

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Equation (4.4) suggests that we can generate samples from $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t})$ via reweighting and resampling from the current posterior samples $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$. Specifically, given a sufficiently large set of posterior samples $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) \sim f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)})$, we assign distinct weights for each sample $m \in \{1, \ldots, M\}$ and future item $j_t \in R_t$:

$$q_m = \frac{w_m}{\sum_{i=1}^{M} w_i}, \quad \text{where } w_i = \Phi(\boldsymbol{B}'_{j_t}\boldsymbol{\theta}_i + D_{j_t})^{y_{j_t}} \left(1 - \Phi(\boldsymbol{B}'_{j_t}\boldsymbol{\theta}_i + D_{j_t})\right)^{1-y_{j_t}}.$$

To sample from $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:(t-1)}, y_{j_t})$, we then draw from the discrete distribution over $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$, placing weight q_m on $\boldsymbol{\theta}_m$. This approach eliminates the need to sample from $|R_t| \times 2$ distinct posteriors directly, further accelerating MI computation and making it scalable for large item banks.

5. Learning optimal item-selection policy

Building on the Bayesian MIRT foundation established in the previous section, we now formulate the problem of CAT as an RL task. The Bayesian framework provides two key advantages that make this integration both principled and computationally feasible. First, the identified latent factor posterior distributions offer a well-defined representation of examinee knowledge, allowing their corresponding posterior parameters to be parameterized directly as state variables. Without such identification, encoding an unknown and analytically intractable posterior would be ambiguous. Second, because RL typically requires extensive simulations to learn an optimal policy, the acceleration achieved in online item selection enables rapid simulation of testing sessions, thereby substantially improving the efficiency of policy training.

As illustrated in Section 3.3, an RL approach addresses the myopic nature of traditional CAT selection rules, enables a more flexible reward structure, and directly minimizes the number of items required for performing online adaptive testing. Specifically, we propose a novel double Q-learning algorithm (Mnih et al., 2015; van Hasselt et al., 2016) for online item selection in CAT. The algorithm trains a deep neural network offline using only the item parameters estimated from an arbitrary two-parameter MIRT model. The offline training phase is completed before any live CAT administration. After training, the network weights are fixed, and online CAT only requires sequentially updating the posterior state and applying the learned network to select the next item. During online item selection, the neural network takes the current posterior distribution $s_t := f(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:t})$ as input and outputs the next item selection for step $(t+1)$. The neural network architecture is described in Section 5.1, while the double deep Q-learning algorithm is detailed in Section 5.2.

5.1. Deep Q-learning neural network design

A potential reason that a deep RL approach has not been proposed in the CAT literature is the ambiguity arising from unidentified latent factor distributions. However, by Theorem 4.1, it is straightforward to compactly parameterize the posterior distribution at each time step t using the parameters $\boldsymbol{C}_1 \in \mathbb{R}^{t \times K}$, $\boldsymbol{C}_2 \in \mathbb{R}^t$, and $\text{diag}(\boldsymbol{C}_3) \in \mathbb{R}^t$, where $\text{diag}(\boldsymbol{C}_3)$ represents the diagonal vector of $\boldsymbol{C}_3$. This structured representation of the posterior enables item-selection policy learning via deep neural networks, which takes posterior parameters as input and outputs item selection.

Neural networks can be regarded as flexible function approximators that learn nonlinear mappings between inputs and outputs through a composition of simple transformations (Goodfellow et al., 2016). In our framework, the deep neural network consists of two key components: an encoder and a classifier. At each time step t, the encoder maps the collection of posterior parameters to a latent representation in $\mathbb{R}^L$, where L is a hyperparameter and represents the dimension of the latent feature space. The classifier then outputs a J-dimensional vector of estimated Q-values and selects the next item from the item bank that is expected to yield the highest Q-value.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Define the collection of the posterior parameters $\tilde{\boldsymbol{\xi}}_t := \{(\boldsymbol{C}_{1h}, C_{2h}, C_{3h})\}_{h=1}^{t}$, where $\boldsymbol{C}_{1h} \in \mathbb{R}^K$ is the hth row of $\boldsymbol{C}_1$, $C_{2h} \in \mathbb{R}$ is the hth element of $\boldsymbol{C}_2$, and $C_{3h} \in \mathbb{R}$ is the hth elements in the diagonal of $\boldsymbol{C}_3$. Since the size of the posterior parameters $\tilde{\boldsymbol{\xi}}_t$ grows over time, and permuting the tuples within $\tilde{\boldsymbol{\xi}}_t$ still describes the same posterior, we have to design a neural network that can take inputs of growing size, and can provide output that is permutation invariant of the inputs. One solution is to consider the weight sharing idea from the Bayesian experimental design literature (Foster et al., 2021). Let $\phi_1(.) : \mathbb{R}^{K+2} \to \mathbb{R}^{L_1}$ denote an encoder component that maps each tuple $(\boldsymbol{C}_{1h}, C_{2h}, C_{3h}) \in \mathbb{R}^{K+2}$ to an L_1-dimensional latent space, and consider the operation $g_1(.)$ as follows:

$$g_1(\tilde{\boldsymbol{\xi}}_t) := \sum_{h=1}^{t} \phi_1\{(\boldsymbol{C}_{1h}, C_{2h}, C_{3h})\}. \tag{5.1}$$

Observe that $g_1(.)$ is capable of handling a growing number of inputs through summations of $\phi_1(.)$ functions over t. More importantly, permuting the order of the tuples in $\tilde{\boldsymbol{\xi}}_t$ does not change the value of $g_1(.)$, since summation is permutation invariant. Although the form of g_1 may look restrictive, any function $f(.)$ operating on a countable set can be decomposed into the form $\rho \circ g_1(.)$, where $\rho(.)$ is a suitable transformation that can be learned from another neural network (see Theorem 2 of Zaheer et al., 2017).

In principle, the network design $\rho \circ g_1(.)$, coupled with the Q-learning algorithm, is sufficient to learn the optimal policy. However, to enhance learning efficiency, we further enrich the state representation $\tilde{\boldsymbol{\xi}}_t$ with a matrix of prediction quartiles $\boldsymbol{\Psi}_t \in \mathbb{R}^{J \times Q}$, where each row contains a vector of quantiles of the predictive distribution for item j. We form these quantiles by first drawing samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M \sim f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t})$ as in Section 4, and then computing the quantiles of the prediction samples $\{\Phi(\boldsymbol{B}_j'\boldsymbol{\theta}_i + D_j)\}_{i=1}^{M}$ for each item j. Since sampling from $f(\boldsymbol{\theta}|\boldsymbol{Y}_{1:t})$ only needs to be done once, computing the matrix $\boldsymbol{\Psi}_t$ online is computationally efficient.

This practice of augmenting the raw state variable with additional contextual features of the states ($\boldsymbol{\Psi}_t$) echoes the common strategy in the RL literature (Mnih et al., 2015; Schaul et al., 2015): learning tends to be more stable and efficient when the state representation captures not only the current ability estimate but also characteristics of the item bank. In our setting, $\boldsymbol{\Psi}_t$ offers a richer description of the predictive distributions across items given the current estimates, effectively serving as additional covariates for item selection. Empirically, we find that incorporating $\boldsymbol{\Psi}_t$ substantially accelerates convergence toward the optimal item-selection policy. Including $\boldsymbol{\Psi}_t$ in the state representation is harmless, as it is a deterministic function of the posterior distribution; identical posteriors will always produce identical $\boldsymbol{\Psi}_t$, ensuring that the policy remains consistent across equivalent states.

In summary, Figure 2 describes the architecture of the policy network used in our approach. The same network is trained offline and then deployed during live CAT with fixed weights, so only the posterior state is updated sequentially during online administration. To select the $(t+1)$th item, our proposed neural net takes two inputs: the posterior parameters $\tilde{\boldsymbol{\xi}}_t$ and the prediction matrix $\boldsymbol{\Psi}_t$, and outputs the selected item. Let $\phi_2 : \mathbb{R}^{J \times Q} \to \mathbb{R}^{L_2}$ represent the encoder component that maps the matrix $\boldsymbol{\Psi}_t \in \mathbb{R}^{J \times Q}$ to L_2-dimensional space. Write $L = L_1 + L_2$ and the concatenated outputs of $g_1(.)$ and $\phi_2(.)$ as $[g_1(\tilde{\boldsymbol{\xi}}_t)', \phi_2(\boldsymbol{\Psi}_t)']' \in \mathbb{R}^L$, we then define the classification component of neural network as $\rho(.) : \mathbb{R}^L \to \mathbb{R}^J$, which maps the concatenated outputs of $g_1(.)$ and $\phi_2(.)$ to a J-dimensional logit vector. Denote the final policy network as π_ϕ and recall that $\rho(.)$ represents the classification layer, our proposed network selects the item at time t corresponding to the maximum value of the following function:

$$\pi_\phi(\tilde{\boldsymbol{\xi}}_t, \boldsymbol{\Psi}_t) := \rho \circ \left\{[g_1(\tilde{\boldsymbol{\xi}}_t)', \phi_2(\boldsymbol{\Psi}_t)']'\right\}.$$

We implement both primary and target Q-networks as simple feed-forward multilayer perceptron and find that their performance is largely insensitive to the exact choice of L_1 and L_2, and does not require very large depth. All architectural details, such as layer sizes, activations, and optimizer settings, are provided in Appendix D of the Supplementary Material. This robustness suggests that

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

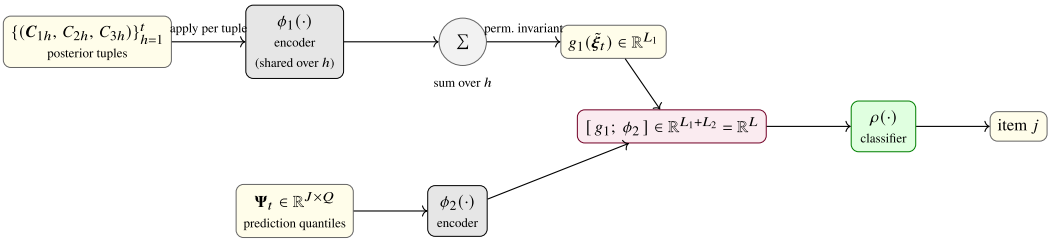


Figure 2. High-level architecture of the Q-network. The shared encoder ϕ_1 maps each tuple of posterior parameters to $\mathbb{R}^{L_1}$ and the sum yields the permutation invariant representation $g_1(\tilde{\xi}_t)$. The matrix Ψ_t is encoded by ϕ_2. The concatenated vector in $\mathbb{R}^L$ is passed to the classifier ρ to select the jth item (largest value in the J logits). This network is trained offline using Algorithm 1; during live CAT, its weights are fixed and only the posterior state is updated sequentially.

our empirical improvements stem from the CAT-specific state design and augmentation rather than network complexity.

5.2. Double deep Q-learning for CAT

In classical tabular Q-learning, the action-value function is stored as a Q-matrix whose rows correspond to states and whose columns correspond to actions. Such a representation is infeasible in CAT because the state space is combinatorially large and the state variable s_t is represented by continuous posterior summaries rather than a small finite index. We therefore replace the tabular Q-matrix by a neural-network approximator $Q_w(s,a)$. For a given state s_t, the network in Figure 2 outputs the full vector of estimated Q-values before selecting the optimal item, which plays the same role as one row of a tabular Q-matrix, but is produced through function approximation rather than table lookup.

The standard Q-learning algorithm often overestimates Q-values because the same action-value approximation is used both to select the maximizing action and to evaluate it when constructing the temporal-difference target. To address this issue, we adopt the double Q-learning approach (van Hasselt et al., 2016), which replaces the single action-value approximation by two Q-networks with distinct roles. The primary network Q_w is updated by gradient descent and is used to select the action with the largest estimated Q-value, while the target network $Q_{\tilde{w}}$ is a delayed copy used only to evaluate that selected action when forming the target. This is the neural-network analog of maintaining two Q-matrices in double Q-learning. The separation reduces overestimation bias and leads to more stable training. Specifically, compared with standard Q-learning, double Q-learning minimizes the same TD loss in (3.11) but replaces the target in (3.12) with

$$y(s,a,r,s') \;=\; r + \gamma\, Q_{\tilde{w}}\Big(s',\, \arg\max_{a'} Q_w(s',a')\Big).$$

Our proposed algorithm is presented in Algorithm 1. The algorithm trains, entirely offline, the policy network illustrated in Figure 2. Once training is complete, the learned network is deployed during live CAT with fixed weights, and only the posterior state is updated sequentially as new responses are observed. Importantly, the offline training phase in Algorithm 1 relies only on the item parameters $(\boldsymbol{B},\boldsymbol{D})$ and simulated examinees drawn from a specified latent factor distribution (e.g., $\boldsymbol{\theta}_i \sim \mathcal{N}(0,\mathbb{I}_K)$), rather than on observed item response data from the item bank. If prior information about the distribution of future online examinees is available, the simulation distribution can be adapted accordingly. In this manuscript, we consistently use a standard multivariate normal distribution to ensure comparability across methods and experiments.

Note that for the Q-learning to converge to an optimal policy, it is essential to adopt an ε-greedy policy, where the algorithm makes random item selections with probability ε and gradually decreases ε over the course of training. For simplicity, the state variable s_t in Algorithm 1 is a compact representation of $(\tilde{\boldsymbol{\xi}}_t,\boldsymbol{\Psi}_t)$ defined in Section 5.1. In practice, we can terminate the training when both

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

the rewards and the validation loss stabilize, indicating that the neural network has well approximated the optimal policy.

Algorithm 1: Double Q-learning algorithm for Bayesian CAT (offline).

Input: Item Parameters $(\boldsymbol{B}, \boldsymbol{D})$, total Episodes E, $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]$ with $\varepsilon^{(0)} = \bar{\varepsilon}$, buffer $\mathcal{H}$, target network update frequency T, batch size M.

Output: Learned Q-network Q_w for item selection.

Initialize primary Q-network Q_w and target Q-network $Q_{\tilde{w}}$, with weights $w_0 = \tilde{w}_0$;

for $i = 1$ **to** E **do**

  Draw $\boldsymbol{\theta}_i$ from $\mathcal{N}(0, \mathbb{I}_K)$;

  Set initial state $s_0 := [\mathcal{N}(0, \mathbb{I}_K), \boldsymbol{\Psi}_0]$ and available item set $R_0 = \{1, \ldots, J\}$;

  for $t = 0$ **to** H **do**

    Select action a_t using ε-greedy policy:

$$a_t = \begin{cases} \text{Random Selection} & \text{with probability } \varepsilon^{(t)} \\ \arg\max_{a \in R_t} Q_w(s_t, a) & \text{otherwise.} \end{cases}$$

    Simulate response $y_t \sim \text{Bernoulli}(\Phi(\boldsymbol{B}'_{a_t}\boldsymbol{\theta}_i + D_{a_t}))$;

    Update available item set $R_{t+1} = R_t \backslash a_t$;

    Obtain new state s_{t+1} from $(\tilde{\boldsymbol{\xi}}_{t+1}, \boldsymbol{\Psi}_{t+1})$;

    Compute reward r_t;

    if $r_t = 0$ **then**

      Break and proceed to the next episode;

    Store transition $H_{i,t} := (s_t, a_t, r_t, R_t, s_{t+1})$ into buffer $\mathcal{H}$;

    if $|\mathcal{H}| >$ *batch size* M **then**

      Randomly sample M transitions $\{H_{i,t}\}_{j=1}^{M}$ from $\mathcal{H}$;

      Compute $a_j = \arg\max_{a \in R_j} Q_w(s'_j, a)$;

      Compute target:

$$y^{(j)} = \begin{cases} -1 & r_j = 0 \\ -1 + \gamma Q_{\tilde{w}}(s'_j, a_j) & \text{otherwise.} \end{cases}$$

      Perform batch gradient descent on loss: $\sum_{j=1}^{M}(y^{(j)} - Q_w(s_j, a_j))^2$;

    Decrease $\varepsilon^{(t)}$ if $\varepsilon^{(t)} \geq \underline{\varepsilon}$;

  Update the target network by setting $\tilde{w} = w$ if i is a multiple of T;

6. Simulation

6.1. Simulation design

To evaluate our approach in a challenging and realistic simulation setting, we randomly generated a 5-factor, 150-item factor loading matrix $\boldsymbol{B} \in \mathbb{R}^{150\times 5}$. Each column of $\boldsymbol{B}$ was initialized by randomly permuting 150 equally spaced values, with the magnitude constrained to lie within $[0.3, 3]$. We then imposed a lower triangular structure to ensure identifiability. To better reflect practical datasets, where items rarely load on all five factors, each item was required to load on the first factor and on at most two additional factors beyond the first. A visualization of the loading matrix can be found in Appendix F of the Supplementary Material. Item intercepts were independently drawn from $\text{Unif}(-1.5, 1.5)$.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Table 1. Comparison of win shares (W.S), termination, and computation

Algorithm	Avg termination (items)	W.S dim1	W.S dim2	W.S dim3	Avg time (s/item)
EAP	31.7	17.0%	16.4%	13.6%	**0.037**
Max Pos	34.0	16.8%	16.2%	18.6%	0.055
MI	23.2	19.0%	21.4%	20.2%	0.082
Max Var	25.7	23.6%	22.6%	23.6%	0.038
Q-learning	**21.5**	**23.6**%	**23.4%**	**24.0%**	0.064

The goal of the simulation is to accurately estimate the first three factors ($\mathcal{K} = \{1,2,3\}$), while accounting for the presence of factors 4 and 5. Leveraging our proposed direct sampling approach outlined in Section 4, we implemented our double Q-learning algorithm alongside all existing methods discussed in Section 3.2. These include the EAP approach (Equation (3.2)), the Max Pos approach (Equation (3.3)), the MI approach (Equation (3.5)), and the Max Var approach (Equation (3.6)). We further modified all baseline information-based criteria so that they also target only the prioritized subset of factors $\mathcal{K} = \{1,2,3\}$. Specifically, the EAP, Max Pos, and MI rules are now written as integrals over $\boldsymbol{\theta}_{\mathcal{K}}$ rather than all five factors. These adjustments ensure that all competing CAT algorithms are tuned to the same estimation target, making the simulation comparisons fair and directly comparable.

For performance evaluation, we generated $N = 500$ online examinees with latent traits $\boldsymbol{\theta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbb{I}_5)$. For each examinee, we administered a 50-item test using each of the CAT algorithms. We then compared performance in terms of posterior variance reduction, mean-squared error (MSE), termination efficiency, and item exposure rates based on these 500 adaptive testing sessions.

For the Q-learning algorithm, we trained the Q-network for 80,000 episodes, with exploration parameter ε decreasing linearly from 0.99 to 0.01 over 700,000 steps. We set a sufficiently large $H = 60$ items, with discount factor $\gamma = 0.95$. To verify that the double Q-learning algorithm indeed converges to the optimal policy π^*, we provide further details on the training dynamics of our deep Q-learning algorithm, illustrating the increase in rewards until convergence in Appendix F.2 of the Supplementary Material.

6.2. Simulation results

To evaluate termination efficiency, we start with a standard multivariate Gaussian prior with unit marginal variances and terminate the test once the maximum posterior variance across *all three* target factors in $\mathcal{K}$ falls from 1 to below $\tau^2 < 0.16$. In practice, practitioners may select the threshold τ^2 based on the specific requirements of their application. Based on 500 simulated adaptive testing sessions, Figure 3 illustrates how the percentage of completed test sessions increases as more items are administered. A faster growth rate of the completion percentages indicates faster posterior variance reduction. Notably, the purple Q-learning curve consistently rises more quickly than the others, demonstrating significant testing efficiency gains. As summarized in the second column of Table 1, Q-learning requires an average of only 21.5 items to reach the termination criterion for all three targeted factors, outperforming all other methods.

To assess early-stage performance, we compute the MSEs after 20 items for each test session and summarize results using win shares, defined as the percentage of test takers for whom each method achieves the lowest MSE after 20 items. These win shares are reported across all three latent dimensions in Table 1. Notably, Q-learning achieves the best win shares across all three factors. This advantage at early stages is consistent with the training objective of Q-learning, which explicitly rewards rapid posterior variance reduction and early test termination, with the average stopping time occurring at around 20 items.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

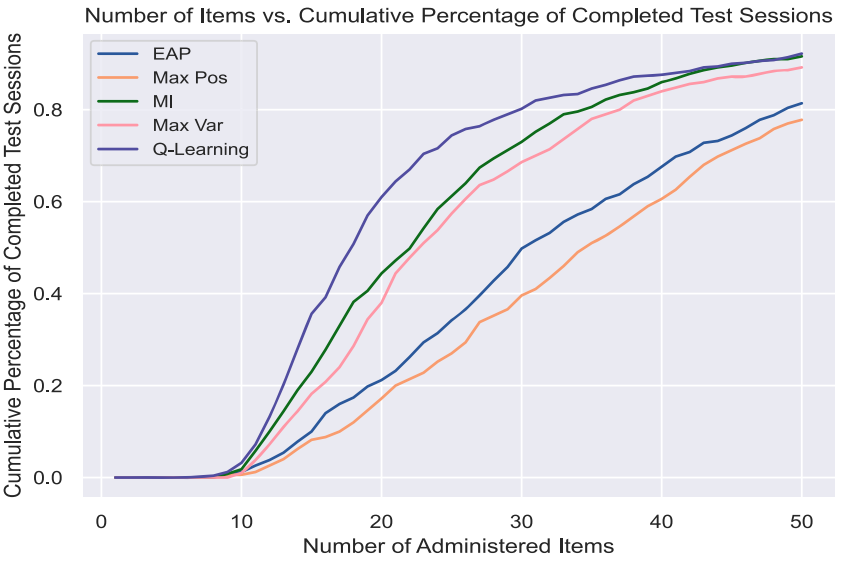


Figure 3. Number of items versus cumulative percentage of completed tests.

Table 2 reports the evolution of MSEs as a function of test length up to 50 administered items. Consistent with the win share analysis, Q-learning exhibits competitive performance during the early stages of testing. In particular, at $T = 20$ and $T = 30$, Q-learning attains the smallest MSEs for Factors 2 and 3. As the test length increases, performance differences across methods diminish, and Q-learning achieves MSEs comparable to those of MI and Max Var. This pattern is expected, as the Q-learning policy is trained primarily on short-horizon testing trajectories: since most simulated tests terminate before 30 items, the algorithm is exposed less frequently to long-horizon simulations during training. Additional visualizations of the MSE trajectories are provided in Appendix F.3 of the Supplementary Material.

Figure 4 summarizes the distribution of item exposure rates across the 150 item bank for each CAT algorithm, where the exposure rate of an item is defined as the proportion of examinees whose test includes that item at least once. With test length $H = 50$ and bank size $J = 150$, the expected average exposure is $H/J \approx 0.33$, and all five algorithms yield mean and median exposure rates close to this benchmark. Differences across methods primarily arise in the tails of the exposure distribution. In particular, the EAP and Q-learning policies exhibit slightly higher upper quartile and maximum exposure rates, indicating a more selective use of highly informative items. Importantly, no method concentrates exposure excessively on a small subset of items, suggesting that all algorithms maintain reasonable item utilization in this simulation setting. For the Q-learning approach, incorporating explicit exposure control mechanisms into the learning objective, such as penalizing repeated use of highly exposed items, is a natural extension but is beyond the scope of the present study.

The last column of Table 1 reports the average online item-selection time, highlighting the significant computational advantages of our direct sampling approach. Even the most computationally intensive MI approach requires only 0.082 seconds per item selection, via our proposed posterior reweighting strategy. Q-learning adds only a single feed-forward pass per selection, keeping online latency under a few milliseconds on standard hardware without GPU acceleration. While training the Q-network offline using Algorithm 1 took approximately 30 hours for this exercise, this is a one-time investment that can be accelerated if GPUs are available; thereafter, the virtually instantaneous online policy makes the approach highly practical for real-time CAT deployment.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Table 2. MSEs between posterior mean and ground truth for the first three latent factors as a function of test length

		Number of items				
Factor	Algorithm	10	20	30	40	50
1	EAP	0.175	0.123	0.090	0.073	0.062
	Max Pos	0.194	0.128	0.104	0.083	0.062
	MI	**0.180**	0.100	**0.066**	**0.055**	**0.050**
	Max Var	0.186	**0.080**	0.073	0.059	0.056
	Q-learning	0.221	0.096	0.071	0.060	0.054
2	EAP	0.362	0.214	0.164	0.125	0.105
	Max Pos	0.354	0.221	0.166	0.129	0.103
	MI	0.223	0.150	0.120	0.102	**0.086**
	Max Var	**0.214**	0.141	0.118	**0.092**	0.087
	Q-learning	0.282	**0.138**	**0.111**	0.096	0.087
3	EAP	0.317	0.226	0.152	0.112	0.085
	Max Pos	0.324	0.207	0.139	0.115	0.092
	MI	**0.213**	0.127	0.103	0.096	0.089
	Max Var	0.234	0.136	0.110	0.099	0.096
	Q-learning	0.233	**0.114**	**0.098**	**0.091**	**0.085**

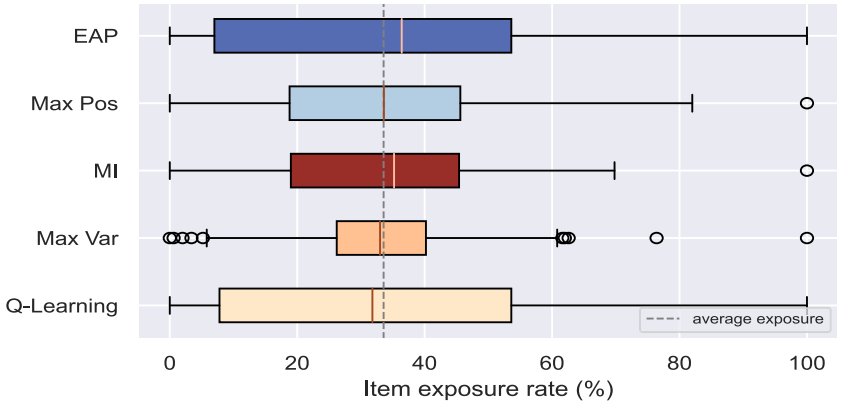


Figure 4. Distributions of item exposure rates.

7. Cognitive function measurements

7.1. pCAT-COG data and experiment design

We revisit the problem of designing a deep CAT system for the pCAT-COG study, as outlined in Section 2. Since item response data for all $N = 730$ examinees across $J = 57$ items are available, we can directly use real item responses during evaluation rather than simulating testing sessions. In this experiment, all $N = 730$ examinees are treated as online examinees. All CAT algorithms, including Q-learning, have access only to the estimated item parameters from the pCAT-COG study and do not observe the examinees' binary item responses beyond those revealed sequentially during adaptive administration.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Table 3. Comparison of termination efficiency, primary-factor accuracy, and computation for pCAT-COG

Algorithm	Avg termination (items)	Correlation	Win shares	Avg time (s/item)
EAP	15.1	0.956	17.7%	**0.020**
Max Pos	18.5	0.933	16.7%	0.030
MI	12.6	0.959	21.8%	0.038
Max Var	13.0	0.951	20.9%	0.021
Q-learning	**11.2**	**0.959**	**22.9**%	0.025

Note: Primary factor correlation and win shares (W.S) are computed after 20 administered items. Win shares denote the proportion of examinees for whom a method attains the smallest mean-squared error for the primary factor at this stage.



(a) Test Termination Speed (b) Test MSE Accuracy

Figure 5. pCAT-COG: Primary factor posterior variance reduction (left) and estimation accuracy (right).

Given that pCAT-COG is designed to measure global cognitive ability (first column in Figure 1) while accounting for five cognitive subdomains, we specified the Q-learning reward function to reduce posterior variance for the primary factor ($\mathcal{K} = \{1\}$). This demonstrates the flexibility of our CAT system, as it can be tailored to the specific cognitive assessment needs. As in Section 6, we also modified all baseline CAT algorithms so that they also target only the primary factor by integrating only over the primary dimension for fair comparison.

Finally, this real-data experiment also serves as a robustness check for the proposed Q-learning approach. As described in Algorithm 1, the policy is trained using simulated examinees with latent factors drawn from a standard multivariate normal distribution. In contrast, the empirical factor correlations in pCAT-COG need not follow this distributional assumption. Evaluating Q-learning on real response data therefore provides evidence that the learned policy remains effective when the latent factor structure deviates from the training distribution used in simulation.

7.2. Results

The left subplot of Figure 5 shows that Q-learning again achieves the fastest test termination compared to other methods. As before, the test is dynamically terminated when the posterior variance drops from 1 to below $\tau^2 = 0.16$, and the cumulative percentage of completed test sessions is computed over all 730 examinees. The second column of Table 3 shows that Q-learning reaches the desired posterior variance reduction threshold after only an average of 11.2 items.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Table 4. Mean-squared errors (MSEs) as a function of test length

Algorithm	10 items	20 items	30 items	40 items
EAP	0.183	0.079	**0.029**	**0.010**
Max Pos	0.217	0.113	0.067	0.016
MI	0.165	**0.070**	0.032	0.014
Max Var	0.165	0.085	0.038	0.014
Q-learning	**0.123**	0.072	0.038	0.016

For further comparison, we consider adaptively selecting 20 items for each test taker without dynamic termination. We chose the number 20 because the Max Pos approach required the largest average of 19.8 items for termination. The third column of Table 3 shows that the Q-learning approach achieves the highest 0.959 correlation between the estimated primary factor posterior means and the ground truth after only 20 items. Additionally, the Q-learning approach attains the highest win shares in estimating the primary dimension across all $N = 730$ examinees after 20 items. As defined in Section 6, win shares denote the proportion of examinees for whom a CAT method attains the smallest MSE. Since all 57 items are ultimately administered in the real-data study, exposure rates are degenerate under full-length testing. We therefore report item exposure patterns based on the first 30 administered items in Appendix G.1 of the Supplementary Material.

Additionally, the right subplot of Figure 5 highlights the rapid decay of MSE in estimating the primary dimension with Q-learning. As summarized in Table 4, Q-learning achieves the smallest MSE at the early stages when $T = 10$ and $T = 20$. At longer horizons when $T = 30$ and $T = 40$, the gains in MSE for the Q-learning approach diminish. This behavior is consistent with the simulation findings in Section 6 and reflects the reward specification used for Q-learning, which emphasizes rapid early-stage variance reduction. Because most simulated training trajectories generated in Algorithm 1 terminate well before 20 items, such performance behavior for Q-learning is expected.

This experiment highlights the effectiveness of our Q-learning approach in high-dimensional cognitive function measurement. Unlike approaches that learn a fixed test design offline (Krantsevich et al., 2023), our framework enables fully dynamic item selection, maximizing the utilization of the item bank by exploring diverse testing trajectories. This termination efficiency is particularly valuable in cognitive assessment, where item development is costly, and can help mitigate practice effects and preserve items for future use.

8. Discussion

This work advocates for a deep RL perspective in the design of MCAT systems. We make two key contributions to the existing CAT literature: (1) a computational advancement that accelerates existing online item-selection rules within a flexible Bayesian MIRT framework and (2) a novel RL-based approach that mitigates myopic decision-making and prioritizes the assessment of primary factors of interest.

Theorem 4.1 not only provides an efficient parameterization of the latent factor posterior distributions for our proposed deep Q-learning approach, but also improves existing CAT item-selection algorithms, as detailed in Section 4. Leveraging direct sampling from unified skew-normal distributions, our methodology scales efficiently with a large number of factors and items, achieving near-instantaneous online selection by circumventing MCMC sampling and data augmentation. Additionally, our approach naturally extends to fully Bayesian item selection by accounting for uncertainties in item parameters, which is an essential consideration when the item bank is not well-calibrated as detailed in Appendix C of the Supplementary Material. Throughout the testing trajectory, our approach precisely characterizes the evolution of posterior distributions at each time step, providing a more robust measurement process beyond point estimates with difficult-to-compute standard errors.

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Another key contribution of this work is the development of a robust deep double Q-learning algorithm with a customized reward structure that directly minimizes test length. As demonstrated in both simulations and real-data studies, our Q-learning algorithm consistently achieves the fastest posterior variance reduction while rapidly decreasing estimation bias. Moreover, its flexible reward function allows adaptation to different testing objectives, providing a principled framework for designing customized tests and overcoming the myopic nature of traditional CAT methods.

Our work also offers practical guidance for selecting the appropriate item-selection algorithms. Unlike the existing heuristic rules, one limitation of our deep CAT system is the requirement of offline training (Algorithm 1) before online deployment. For the pCAT-COG study, offline training took approximately 12 hours on a single GPU as a one-time investment, but no GPU is needed for subsequent online item selection. Even when offline training is undesirable, our framework significantly accelerates existing methods. Experiments show that MI (Equation (3.5)) and our Max Var (Equation (3.6)) approaches often outperform other online item-selection rules, echoing the findings presented in Wang and Chang (2011).

A promising direction for future work is to address several limitations of the current Q-learning framework for CAT. First, the learned policy is tied to a specific item bank and must be retrained when the items are substantially modified or replenished, which may limit its immediate applicability in settings with frequent item updates. Second, as a model-free approach, Q-learning can be computationally expensive to train, particularly for long-horizon testing scenarios in which optimal policies depend on extended future trajectories. In addition, although the proposed approach does not rely on tabular exploration of the full state space and instead uses function approximation based on a compact posterior representation, scalability remains an important practical consideration as the state space grows exponentially. Our simulation results suggest that the method works well for moderate 150-item banks, but additional methodological development may be useful for extending the approach to substantially larger banks. Finally, while the proposed method yields reasonable item exposure patterns in our experiments, additional work is needed to incorporate explicit exposure control mechanisms directly into the learning objective.

Another exciting avenue is to explore alternative reward structures for our Q-learning algorithm. While the 0–1 reward structure is interpretable and stabilizes Q-network training, it provides sparse feedback, which may limit empirical performance. It is also important to understand when RL truly improves on myopic rules, since Chan and Farias (2009) showed that greedy policies can be near-optimal in some dynamic stochastic optimization problems. This requires careful assumptions about the reward function and the item bank properties, providing deeper insights into the trade-offs between RL and traditional item-selection strategies.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10106.

Data availability statement. The data and code that support the findings of this study are publicly available at the GitHub repository: https://github.com/JiguangLi/deep_CAT/tree/deep_cat.

Funding statement. R.G. acknowledges support from the National Institute on Aging (NIA) under Grant R56 AG084070-01, titled Adaptive Testing of Cognitive Function based on Multi-Dimensional Item Response Theory (2023–2025; R.G. – PI). V.R. acknowledges support from the National Science Foundation (NSF-DMS 1944740).

Competing interests. Dr. Gibbons founded the company Adaptive Testing Technologies that distributes traditional MIRT-based CATs for which competing interests are managed by the University of Chicago. The remaining authors declare none.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679.

Alzheimer's Association. (2024). 2024 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *20*(5). https://doi.org/10.1002/alz.13809

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Arellano-Valle, R. B., & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, *33*(3), 561–574.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561. https://doi.org/10.1007/BF02296195

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. (1st ed.) Athena Scientific.

Bock, R., & Gibbons, R. (2021). *Item response theory*. Wiley.

Botev, Z. I. (2016). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(1), 125–148. https://doi.org/10.1111/rssb.12162

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304. https://doi.org/10.1214/ss/1177009939

Chan, C. W., & Farias, V. F. (2009). Stochastic depletion problems: Effective myopic policies for a class of dynamic optimization problems. *Mathematics of Operations Research*, *34*(2), 333–350.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*(1), 37–52. https://doi.org/10.1007/BF02294469

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing *Applied Psychological Measurement*, *20*, 213–229.

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, *106*(4), 765–779. https://doi.org/10.1093/biomet/asz034

Foster, A., Ivanova, D. R., Malik, I., & Rainforth, T. (2021). Deep adaptive design: Amortizing sequential Bayesian experimental design. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 139, pp. 3384–3395). PMLR.

Gibbons, R., Lauderdale, D., Wilson, R., Bennett, D., Arar, T., & Gallo, D. (2024). Adaptive measurement of cognitive function based on multidimensional item response theory. *Alzheimer's & Dementia Translational Research & Clinical Interventions*, *10*. https://doi.org/10.1002/trc2.70018

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436. https://doi.org/10.1007/BF02295430

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders [Epub 2015 Nov 20]. *Annual Review of Clinical Psychology*, *12*, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Gordon, N. J., Salmond, D., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. https://api.semanticscholar.org/CorpusID:12644877

Huang, A. R., Strombotne, K. L., Horner, E. M., & Lapham, S. J. (2018). Adolescent cognitive aptitudes and later-in-life Alzheimer disease and related disorders. *JAMA Network Open*, *1*(5), e181726. https://doi.org/10.1001/jamanetworkopen.2018.1726

Jiang, Z., & Templin, J. (2019). Gibbs samplers for logistic item response models via the pólya–gamma distribution: A computationally efficient data-augmentation strategy. *Psychometrika*, *84*(2), 358–374. https://doi.org/10.1007/s11336-018-9641-x

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., & Levine, S. (2018). Scalable deep reinforcement learning for visionbased robotic manipulation. In A. Billard, A. Dragan, J. Peters, & J. Morimoto (Eds.), *Proceedings of The 2nd Conference on Robot Learning* (Proceedings of Machine Learning Research, Vol. 87, pp. 651–673). PMLR.

Krantsevich, C., Hahn, P. R., Zheng, Y., & Katz, C. (2023). Bayesian decision theory for tree-based adaptive screening tests with an application to youth delinquency. *Annals of Applied Statistics*, *17*(2), 1038–1063. https://doi.org/10.1214/22-AOAS1657

Lam, R., Willcox, K., & Wolpert, D. H. (2016). Bayesian optimization with a finite budget: An approximate dynamic programming approach. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Li, J., Gibbons, R., & Rockova, V. (2025). Sparse Bayesian multidimensional item response theory. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2025.2476786

Li, X., Xu, H., Zhang, J., & Chang, H. (2023). Deep reinforcement learning for adaptive learning systems. *Journal of Educational and Behavioral Statistics*, *48*(2), 220–243. https://doi.org/10.3102/10769986221129847

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. A. (2013). *Playing Atari with deep reinforcement learning*. Preprint, ArXiv, abs/1312.5602. https://api.semanticscholar.org/CorpusID:15238391

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533. https://doi.org/10.1038/nature14236

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection [Epub 2008 Dec 23]. *Psychometrika*, *74*(2), 273–296. https://doi.org/10.1007/s11336-008-9097-5

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback–Leibler information item selection. In *Elements of adaptive testing* (pp. 77–101). https://doi.org/10.1007/978-0-387-85461_84

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Rainforth, T., Foster, A., Ivanova, D. R., & Smith, F. B. (2024). Modern Bayesian experimental design. *Statistical Science*, *39*(1), 100–114. https://doi.org/10.1214/23-STS915

A Rényi. (1961). On measures of entropy and information. https://api.semanticscholar.org/CorpusID:123056571

Schaul, T.,Horgan, D., Gregor, K., & Silver, D. (2015). Universal value function approximators. In F. Bach, & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 1312–1320). PMLR.

Sebastiani, P., & Wynn, H. P. (2000). Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(1), 145–157.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354. https://doi.org/10.1007/BF02294343

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–73). Kluwer Academic.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Smith, A. F. M., & Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, *46*(2), 84–88.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. W. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In J. Fürnkranz & T. Joachims (Eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)* (pp. 1015–1022). Omnipress.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. (2nd ed.) The MIT Press. http://incompleteideas.net/book/the-book-2nd.html

Tan, C., Han, R., Ye, R., & Chen, K. (2020). Adaptive learning recommendation strategy based on deep q-learning. *Applied Psychological Measurement*, *44*(4), 251–266. https://doi.org/10.1177/0146621619858674

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion *Journal of Educational and Behavioral Statistics*, *24*(4), 398–412.

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

van der Linden, W. J., & Ren, H. (2020). A fast and simple algorithm for Bayesian adaptive testing. *Journal of Educational and Behavioral Statistics*, *45*(1), 58–85. https://doi.org/10.3102/1076998619858970

van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30). https://doi.org/10.1609/aaai.v30i1.10295

Veldkamp, B., & van der Linden, W. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575–588. https://doi.org/10.1007/BF02295132

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*(2), 428–449. https://doi.org/10.1007/s11336-013-9399-0

Wang, C., & Chang, H. (2011). Item selection in multidimensional computerized adaptive testing–gaining information from different angles [copyright—The psychometric society 2011; last updated - 2023-12-03]. *Psychometrika*, *76*(3), 363–384.

Wang, C., Chang, H.-H., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, *76*(1), 13–39. https://doi.org/10.1007/s11336-010-9186-0

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*(1), 41–58. https://doi.org/10.1177/0013164406288164

Zaheer, M., Kottur, S., Ravanbhakhsh, S., Póczos, B., Salakhutdinov, R., & Smola, A. J. (2017). Deep sets. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, & S. V. N. Vishwanathan (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3394–3404). Curran Associates, Inc.

Cite this article: Li, J., Gibbons, R. and Ročková, V. (2026). Deep Computerized Adaptive Testing. *Psychometrika*, 1127–1148. https://doi.org/10.1017/psy.2026.10106

https://doi.org/10.1017/psy.2026.10106 Published online by Cambridge University Press