EduQate: Generating Adaptive Curricula through RMABs in Education Settings

Sidney Tio[†,*]
Singapore Management University
Singapore
sidney.tio.2021@phdcs.smu.edu.sg

Dexun Li [*]
Singapore Management University
Singapore
dexunli.2019@phdcs.smu.edu.sg

Pradeep Varakantham
Singapore Management University
Singapore
pradeepv@smu.edu.sg

ABSTRACT

There has been significant interest in the development of personalized and adaptive educational tools that cater to a student's individual learning progress. A crucial aspect in developing such tools is in exploring how mastery can be achieved across a diverse yet related range of content in an efficient manner. While Reinforcement Learning and Multi-armed Bandits have shown promise in educational settings, existing works often assume the independence of learning content, neglecting the prevalent interdependencies between such content. In response, we introduce *Education Network Restless Multi-armed Bandits* (EdNetRMABs), utilizing a network to represent the relationships between interdependent arms. Subsequently, we propose *EduQate*, a method employing interdependency-aware Q-learning to make informed decisions on arm selection at each time step. We establish the optimality guarantee of EduQate and demonstrate its efficacy compared to baseline policies, using students modeled from both synthetic and real-world data.

KEYWORDS

Restless Multi-Armed Bandits; Agents in Education; Networked RMABs; Q-Learning

ACM Reference Format:
Sidney Tio[†,*], Dexun Li [*], and Pradeep Varakantham. 2025. EduQate: Generating Adaptive Curricula through RMABs in Education Settings. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

The COVID-19 pandemic has accelerated the adoption of educational technologies, especially on eLearning platforms. Despite abundant data and advancements in modeling student learning, effectively capturing the learning process with interdependent content remains a significant challenge [9]. The conventional rules-based approach to creating personalized learning curricula is impractical due to its labor-intensive nature and need for expert knowledge. Machine learning-based systems offer a scalable alternative, automatically generating personalized content to optimize learning [22, 25].

One possible approach to model the learning process is the Restless Multi-Armed Bandits (RMAB, [27]), where a teacher agent selects a subset of arms (concepts) to teach each round. However, RMAB's assumption that arms are independent is unrealistic in educational settings. For example, solving a math question on the area of a triangle requires knowledge of algebra, arithmetic, and geometry. Practicing this question should enhance proficiency in all three areas. Models that ignore such interdependencies may inaccurately predict knowledge levels by assuming each exercise impacts only a single area.

In response to this challenge, we introduce an interdependency-aware RMAB model to the education setting. We posit that by acknowledging and modeling the learning dynamics of interdependent content, both teachers and algorithms can strategically leverage overlapping utility to foster mastery over a broader range of topics within a curriculum. We advocate for RMABs as a fitting model for this context, as the inherent dynamics of such a model align closely with the learning process.

In this study, our objective is to derive a teacher policy that effectively recommends educational content to students, accounting for interdependencies among the content to enhance overall utility (that characterizes understanding and retention of content). Our contributions are as follows:

(1) We introduce Restless Multi-armed Bandits for Education (EdNetRMABs), enabling the modeling of learning processes with interdependent educational content.
(2) We propose EduQate, a Whittle index-based heuristic algorithm that uses Q-learning to compute an inter-dependency-aware teacher policy. Unlike previous methods, EduQate does not require knowledge of the transition matrix to compute an optimal policy.
(3) We provide a theoretical analysis of EduQate, demonstrating guarantees of optimality.
(4) We present empirical results on simulated students and real-world datasets, showing the effectiveness of EduQate over other teacher policies.

2 RELATED WORK AND PRELIMINARIES

2.1 Restless Multi-Armed Bandits

The selection of the right time and manner for limited interventions is a problem of great practical importance across various domains, including health intervention [5, 17], anti-poaching operations [20], education [2, 6, 13], etc. These problems share a common characteristic of having multiple arms in a Multi-armed Bandit (MAB) problem, representing entities such as patients, regions of a forest, or students' mastery of concepts. These arms evolve in an uncertain

[*]Equal contribution.
[‡]Corresponding author.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

manner, and interventions are required to guide them from "bad" states to "good" states. The inherent challenge lies in the limited number of interventions, dictated by the limited resources (e.g., public health workers, the number of student interactions). RMAB, a generalization of MAB, offers an ideal model for representing the aforementioned problems of interest. RMAB allows non-active bandits to also undergo the Markovian state transition, effectively capturing uncertainty in arm state transitions (reflecting uncertain state evolution), actions (representing interventions), and budget constraints (illustrating limited resources).

RMABs and the associated Markov Decision Processes (MDP) for each arm offer a valuable model for representing the learning process. Firstly, leveraging the MDPs associated with each arm provides the flexibility to adopt nuanced modeling of learning content, accommodating different learning curves for various content based on students' strengths and weaknesses. Secondly, the transition probabilities serve as a useful mechanism to model forgetting (through state decay due to passivity or negligence) and learning (through state transitions to the positive state from repeated practice). Considering these aspects, RMABs prove to be a beneficial framework for personalizing and generating adaptive curricula across a diverse range of students.

In general, computing the optimal policy for a given set of restless arms in RMABs is recognized as a PSPACE-hard problem [18]. The Whittle index [27] provides an approach with a tractable solution that is provably optimal, especially when each arm is indexable. However, proving indexability can be challenging and often requires specification of the problem's structure, such as the optimality of threshold policies [16, 17]. Moreover, much of the research on Whittle Index policies has focused on two-action settings or requires prior knowledge of the transition matrix of the RMABs. Meeting these conditions proves challenging in the educational context, where diverse students interact with educational systems, each possessing different prior knowledge and distinct learning curves for various topics.

WIQL [5], on the other hand, employs a Q-learning-based method to estimate the Whittle Index and has demonstrated provable optimality without requiring prior knowledge of the transition matrix. We utilize WIQL as a baseline method in our subsequent experiments.

In a recent investigation by [12], RMABs were explored within a network framework, requiring the agent to manage a budget while allocating a high-cost, high-benefit resource to one arm to "unlock" potential lower-cost, intermediate-benefit resources for the arm's neighbors. The network effects emphasized in their work are triggered by an intentional, active action, enabling the agent to choose to propagate positive externalities to a selected arm's neighbors within budget constraints. In contrast, our study delves into scenarios where network effects are indirect results of an active action, and the agent lacks direct control over such effects. Thus, the challenge lies in accurately modeling these network effects and leveraging them when beneficial.

2.2 Reinforcement Learning in Education

In the realm of education, numerous researchers have explored optimizing the sequencing of instructional activities and content, assuming that optimal sequencing can significantly impact student learning. RL is a natural approach for making sequential decisions under uncertainty [1]. While RL has seen success in various educational applications, effectively sequencing interdependent content in a personalized and adaptive manner has yielded mixed or insignificant results compared to baseline teacher policies [8, 11, 21]. In general, these RL works focus on data-driven methods using student activity logs to estimate students' knowledge states and progress, assuming that the interdependencies between learning content are encapsulated in students' learning histories [3, 9, 19]. In contrast, our work focuses on modelling these interdependencies directly.

Of particular relevance are factored MDPs applied to skill acquisition introduced by [11]. While factored MDPs account for interdependencies amongst skills, decentralized policy learning is infeasible as policies must consider the joint state space. Our work leverages the advantage of decentralized policy learning provided by RMABs and introduces a novel decentralized learning approach that exploits interdependencies between arms.

Complementary to RL methods in education is the utilization of knowledge graphs to uncover relationships between learning content [9]. Existing research primarily focuses on establishing these relationships through data-driven methods (e.g. [7, 23]) often leveraging student-activity logs. In this work, we complement such research by presenting an approach where bandit methods can effectively operate with knowledge graphs derived by such methods.

3 MODEL

In this section, we introduce the Restless Multi-Armed Bandits for Education (EdNetRMABs). It is important to note that while we specifically apply EdNetRMABs to the education setting, the framework can be seamlessly translated to other scenarios where modeling the effects of active actions within a network is critical. For ease of access, a table of notations is provided in Table **??**.

In education, a teacher recommends learning content, or items, to maximize student education, often with content from online platforms. Items are grouped by topics, such as "Geometry," where exposure to one piece of content can enhance knowledge across others in the same group. This cumulative learning effect which we refer to as "network effects", implies that exposure to an item is likely to positively impact the student's success on items within the same group. A successful teacher accurately estimates a student's knowledge state over repeated interactions, leveraging these network effects to promote both breadth and depth of understanding through recommendations.

3.1 EdNetRMABs

The RMAB model tasks an agent with selecting k arms from N arms, constrained by a limit on the number of arms that can be pulled at each time step. The objective is to find a policy that maximizes the total expected discounted reward, assuming that the state of each arm evolves independently according to an underlying MDP.

The EdNetRMABs model extends RMABs by allowing for active actions to propagate to other arms dependent on the current

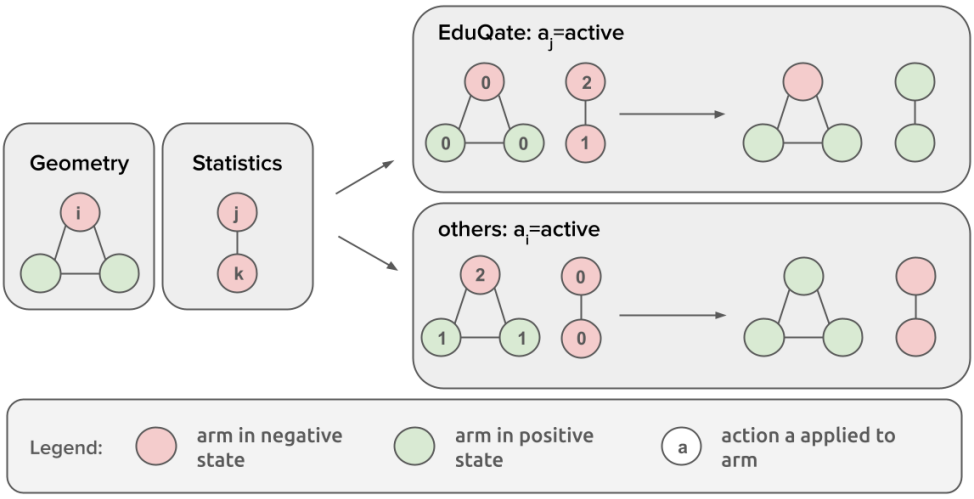

Figure 1: A comparative visualization of interdependency-aware EduQate versus interdependency-unaware teacher algorithms interacting with a EdNetRMAB with five arms organized into two topics. Note that actions 0,1,2 signifies passive, semi-active, and active actions accordingly. Top: EduQate, recognizing interdependencies, selects arm j or k, optimizing overall student knowledge state. Bottom: Interdependency-unaware algorithms treat arms i, j, and k as independent, considering each equally. Selecting arm i results in suboptimal learning outcomes.

arm when it is being pulled, thus relaxing the assumption of independent arms. This is operationalized by organising the arms in a network, and pulling of an arm results in changes for its neighbors, or members in the same group.

When applied to education setting, the EdNetRMABs is formalized as follows:

Arms. Each arm, denoted as $i \in 1, ..., N$, signifies an item. In the context of this networked environment, each arm belongs to a group $\phi \in \{1, ..., L\}$ representing the overarching topic that encompasses related items. It's important to note that arm membership is not mutually exclusive, allowing arms to be part of multiple groups. This flexibility enables a more nuanced modeling of interdependencies among educational content. For instance, a question involving the calculation of the area of a triangle may span both arithmetic and geometry groups.

State space. In this framework, each arm possesses a binary latent state, denoted as $s_i \in \{0, 1\}$, where "0" represents an "unlearned" state, and "1"indicates a "learned" state. Considering all arms collectively, these states serve as a representation of the student's overall knowledge state. In the current work, it is assumed that the states of all arms are fully observable, providing a comprehensive model of the student's understanding of the various educational concepts.

Action space. To capture the network effects associated with arm pulls, we depart from the conventional RMAB framework with a binary action space $A = \{0, 1\}$ by introducing a pseudo-action. In this modified setup, the action space is extended to $A = \{0, 1, 2\}$, where actions 0 and 2 represent "no-pull" and "pull", as commonly used in bandit literature. Notably, in EdNetRMABs, a third action 1 is introduced to simulate the network effects resulting from pulling another arm within the same group. It is important to clarify that agents do not directly engage with action 1 but we employ it solely for modeling network effects, hence the term "pseudo-action".

Transition function. For a given arm i, let $P^{a,i}_{s,s'}$ represent the probability of the arm transitioning from state s to s' under action a. It's noteworthy that, in typical real-world educational settings, the actual transition functions governing the states of the arms are often unknown and, even for the same concept, may vary among students due to differences in prior knowledge [9]. To address this challenge, we adopt model-free approaches in this study, devising methods to compute the teacher policy without relying on explicit knowledge of these transition functions. In the following experiments, we maintain the assumption of non-zero transition probabilities, and enforce constraints that are aligned with the current domain [17]: (i) The arms are more likely to stay in the positive state than change to the negative state: $P^0_{0,1} < P^0_{1,1}$, $P^1_{0,1} < P^1_{1,1}$ and $P^2_{0,1} < P^2_{1,1}$; (ii) The arm tends to improve the latent state if more efforts is spent on that arm, i.e., it is active or semi-active: $P^0_{0,1} < P^1_{0,1} < P^2_{0,1}$ and $P^0_{1,1} < P^1_{1,1} < P^2_{1,1}$.

With the formalization of the EdNetRMABs model provided, we now apply it to an educational context. In this scenario, the agent assumes the role of a teacher and takes actions during each time step $t \in \{1, ..., T\}$. Specifically, at each time step, the teacher recommends an item for the student to study. We represent the vector of actions taken by the teacher at time step t as $\mathbf{a}^t \in \{0, 1, 2\}^N$. Here, arm i is considered to be active at time t if $a^t_i = 2$ and passive when $a^t_i = 0$. When arm i is pulled, the set of arms that share the same group membership as arm i, denoted as ϕ^-_i under goes the pseudo-action, represented as $a^t_j = 1$ for all $j \in \phi^-_i$. In our framework, the teacher agent acts on exactly one arm per time step to simulate the real-world constraint that the teacher can only recommend one

concept to students ($\sum_i I_{a_i^t=2} = 1, \forall t$). Subsequent to taking action, the teacher receives $\mathbf{s}^t \in \{0,1\}^N$, a vector reflecting the state of all arms, and reward $r_t = \sum_{i=1}^{N} s^t(i)$. The vector $\mathbf{s^t}$ represents the overall knowledge state of the student. The teacher agent's goal, therefore, is to maximize the long term rewards, either discounted or averaged.

While previous studies (e.g. [7, 24]) focused on directed relations for prerequisite learning, EdNetRMABs consider undirected relations between learning content. This approach reflects bidirectional learning relationships in education: prerequisites aid future learning, while advanced topics reinforce foundational knowledge. Additionally, EdNetRMABs can be extended to unidirectional structures ($\phi_i \neq \phi_j$ for $j \in \phi_i$).

4 EDUQATE

Q-learning [26] is a popular reinforcement learning method that enables an agent to learn optimal actions in an environment by iteratively updating its estimate of state-action value, $Q(s,a)$, based on the rewards it receives. At each time step t, the agent takes an action a using its current estimate of Q values and current state s, thus received a reward of $r(s)$ and new state s'.

Expanding upon Q-learning, we introduce *EduQate*, a tailored Q-learning approach designed for learning Whittle-index policies in EdNetRMABs. In the interaction with the environment, the agent chooses a single item, represented by arm i, to recommend to the student. In this context, the agent possesses knowledge of the group membership ϕ_i of the selected arm and observes the rewards generated by activating arm i and semi-activating arms in ϕ_i^-. EduQate utilizes this interaction to learn the Q-values for all arms and actions.

To adapt Q-learning to EdNetRMABs, we propose leveraging the learned Q-values to select the arm with the highest estimate of the Whittle index, defined as:

$$\lambda_i = Q(s_i, a_i = 2) - Q(s_i, a_i = 0) + \sum_{u \in \phi_i^-} (Q(s_u, a_u = 1) - Q(s_u, a_u = 0)) \quad (1)$$

Here, λ_i is the Whittle Index estimate for arm i. In essence, the Whittle Index of arm i is computed as the linear combination of the value associated with taking action on arm i over passivity and the value of associated with semi-actively engaging with members from same group, compared to passivity.

To improve the convergence of Q-learning, we incorporate Experience Replay [15]. This involves saving the teacher algorithm's previous experiences in a replay buffer and drawing mini-batches of samples from this buffer during updates to enhance convergence. In Section 4.1, we prove that EduQate will converge to the optimal policy. However, in practice, we may not have enough episodes to fully train EduQate. Therefore, we propose Experience Replay to mitigate the cold-start problem common in RL applications, a common problem where initial student interactions with sub-optimal teachers can lead to poor learning experiences [3].

The pseudo-code is provided in Algorithm 1. Similar to WIQL [5], we employ a ϵ-decay policy that facilitates exploration and learning in the early steps, and proceeds to exploit the learned Q-values in later stages. A visualization of how EduQate interacts with EdNetRMABs compared to other teacher algorithms can be found at Figure 1.

Algorithm 1 Q-Learning for EdNetRMABs (EduQate)

Input: Number of arms N
Initialize $Q_i(s,a) \leftarrow 0$ and $\lambda_i(s) \leftarrow 0$ for each state $s \in S$ and each action $a \in \{0,1,2\}$, for each arm $i \in 1, ..., N$.
Initialize replay buffer D with capacity C.
for t in $1, ..., T$ **do**
  $\epsilon \leftarrow \frac{N}{N+t}$
  With probability ϵ, select one arm uniformly at random. Otherwise, select arm with highest Whittle Index, $i = \arg\max_i \lambda_i$.
  for arm n in $1, ..., N$ **do**
    if $n \neq i$ **then**
      Set arm n to passive, $a_n^t = 0$
    else
      Set arm n to active, $a_n^t = 2$
      for $j \in \phi_i^-$ **do**
        Set arms in same group as i to semi-active, $a_j^t = 1$
      end for
    end if
  end for
  Execute actions $\mathbf{a^t}$ and observe reward r^t and next state s^{t+1} for all arms
  Store experience $(s^t, \mathbf{a^t}, \mathbf{r^t}, \mathbf{s^{t+1}})$in replay buffer D.
  Sample minibatch B of Experience from replay buffer D.
  for Experience in minibatch B **do**
    Update $Q_n(s,a)$ using Q-learning update in Equation **??**.
    Compute λ_n using Equation 1
  end for
end for

4.1 Analysis of EduQate

In this section, we analyze EduQate closely, and show that EduQate does not alter the optimality guarantees of Q-learning under the constraint that maximum number of arms that can be pulled on each timestep $k = 1$ (Theorem 1). Our method relies on the assumption that teachers are limited to assign 1 item to the student at each time step. Theorem 2 analyzes EduQate under the conditions that $k > 1$. Since our setting involves the semi-active actions, we should compute Equation 1. To reiterate, ϕ_i here refers to the group that arm i belongs to, and ϕ_i^- is the same group but does not include arm i. If arm i is selected, then all the remaining arms in group ϕ_i^- should be semi-active.

Theorem 1. *Choosing the top arm with the largest λ value in Equation 1 is equivalent to maximizing the cumulative long-term reward.*

Proof. According to the approach, we select the arm according to the λ value. Assume arm i has the highest λ value, then for any arm j where $j \neq i$, we have

$$\lambda_i \geq \lambda_j$$

$$\begin{aligned} & Q(s_i, a_i = 2) - Q(s_i, a_i = 0) \\ & \quad + \sum_{u \in \phi_i^-} (Q(s_u, a_u = 1) - Q(s_u, a_u = 0)) \\ & \geq Q(s_j, a_j = 2) - Q(s_j, a_j = 0) \\ & \quad + \sum_{w \in \phi_j^-} (Q(s_w, a_w = 1) - Q(s_w, a_w = 0)) \end{aligned} \tag{2}$$

According to the definition of λ in Equation 1, we move the negative part to the other side, and the left side becomes:

$$\begin{aligned} & Q(s_i, a_i = 2) + \sum_{u \in \phi_i^-} (Q(s_u, a_u = 1)) \\ & \quad + Q(s_j, a_j = 0) + \sum_{w \in \phi_j^-} (Q(s_w, a_w = 0)) \end{aligned}$$

and the right side is similar. There are three cases:

- arm i and arm j are not connected, and group ϕ_i and ϕ_j has no overlap, i.e., $\phi_i \cap \phi_j = \emptyset$. We add $\sum_{z \notin \phi_i \wedge z \notin \phi_j} Q(s_z, a_z = 0)$ on both sides. This denotes the addition of $Q(s_z, a_z = 0)$ for all arm z that are not included in the set of ϕ_i or ϕ_j. We have the left side:

$$\begin{aligned} & Q(s_i, a_i = 2) + \sum_{u \in \phi_i^-} (Q(s_u, a_u = 1)) + Q(s_j, a_j = 0) \\ & \quad + \sum_{w \in \phi_j^-} (Q(s_w, a_w = 0)) + \sum_{z \notin \phi_i \wedge z \notin \phi_j} Q(s_z, a_z = 0) \\ & = Q(s_i, a_i = 2) + \sum_{u \in \phi_i^-} (Q(s_u, a_u = 1)) \\ & \quad + \sum_{w \notin \phi_i} (Q(s_w, a_w = 0)) \\ & = Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{a_i = 2\}}) \end{aligned} \tag{3}$$

  Similarly, we do the same for the right side and thus, the equation 4.1 becomes

$$Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{a_i = 2\}}) \geq Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{a_j = 2\}})$$

- arm i and arm j are not connected, but group ϕ_i and ϕ_j has overlap, i.e., $\phi_i \cap \phi_j \neq \emptyset$. In this case, we add $\sum_{z \notin \phi_i \wedge z \notin \phi_j} Q(s_z, a_z = 0) - \sum_{z \in \phi_i \cap \phi_j} Q(s_z, a_z = 0)$ on both sides.
- arm i and arm j are connected, and group ϕ_i and ϕ_j has overlap, i.e., $\phi_i \cap \phi_j \neq \emptyset$, and $\{i, j\} \subset \phi_i \cap \phi_j$. This case is similar to the previous one, we add $\sum_{z \notin \phi_i \wedge z \notin \phi_j} Q(s_z, a_z = 0) - \sum_{z \in \phi_i \cap \phi_j} Q(s_z, a_z = 0)$ on both sides.

□

Thus when $k = 1$, selecting the top arm according to the λ value is equivalent to maximizing the cumulative long-term reward, and is guaranteed to be optimal.

Theorem 2. *When $k > 1$, selecting the k arms is a NP-hard problem. The non-asymptotic tight upper bound and non-asymptotic tight lower bound for getting the optimal solution are $o(C(n, k))$ and $\omega(N)$, respectively.*

Proof Sketch. This problem can be considered as a variant of the knapsack problem. If we disregard the influence of the shared neighbor nodes for two selected arms, then selecting arm i will not influence the future selection of arm j. In such instances, the problem of selecting the k arms is simplified to the traditional 0/1 knapsack problem, a classic NP-hard problem. Therefore, when considering the effect of shared neighbor nodes for two selected arms, this problem is at least as challenging as the 0/1 knapsack problem. □

When $k > 1$, it is difficult to compute the optimal solution, but a heuristic greedy algorithm with the complexity of $O(\frac{(2N-k)*k}{2})$ exists.

5 EXPERIMENT

In this section, we demonstrate the effectiveness of EduQate against benchmark algorithms on synthetic students and students derived from a real-world dataset, the Junyi Dataset and the OLI Statics dataset. All experiments are run on CPU only. In our experiments, we compare EduQate with the following policies:

- **Threshold Whittle (TW)**: This algorithm, proposed by [17], utilizes an efficient closed-form approach to compute the Whittle index, considering only the pull action as active. It operates under the assumption that transition probabilities are known and stands as the state-of-the-art in RMABs.
- **WIQL**: This algorithm employs a Q-learning-based Whittle Index approach [5]. It learns Q-values using the pull action as the only active strategy and calculates the Whittle Index based on the acquired Q-values.
- **Myopic**: This strategy disregards the impact of the current action on future rewards, concentrating solely on predicted immediate rewards. It selects the arm that maximizes the expected reward at the immediate time step.
- **Random**: This strategy randomly selects arms with uniform probability, irrespective of the underlying state.

Inspired by work in healthcare settings [12, 14], we compare policies using the *Intervention Benefit* (*IB*), which we modify for the education settings and is defined as:

$$IB_{Random}(\pi) = \frac{\mathbb{E}_{\pi}(R(.)) - \mathbb{E}_{Random}(R(.))}{\mathbb{E}_{Random}(R(.))} \tag{4}$$

where *Random* represents a policy where arms are pulled at random. Previous research in educational settings has shown that random policies can produce robust learning outcomes through spaced repetition [9, 10]. Thus, effective algorithms must demonstrate superiority over random policies. Our modified metric effectively compares the extent to which a challenger algorithm π outperforms a random policy.

5.1 Experiment setup

In all experiments, we commence by initializing all arms in state 0 and permit the teacher algorithms to engage with the student for a total of 50 actions, pulling exactly 1 arm (i.e. $k = 1$) at each time step. Following the completion of these actions, the episode concludes, and the student state is reset. This process is iterated across 800 episodes, for a total of 30 seeds. The datasets used in our experiment are described below:

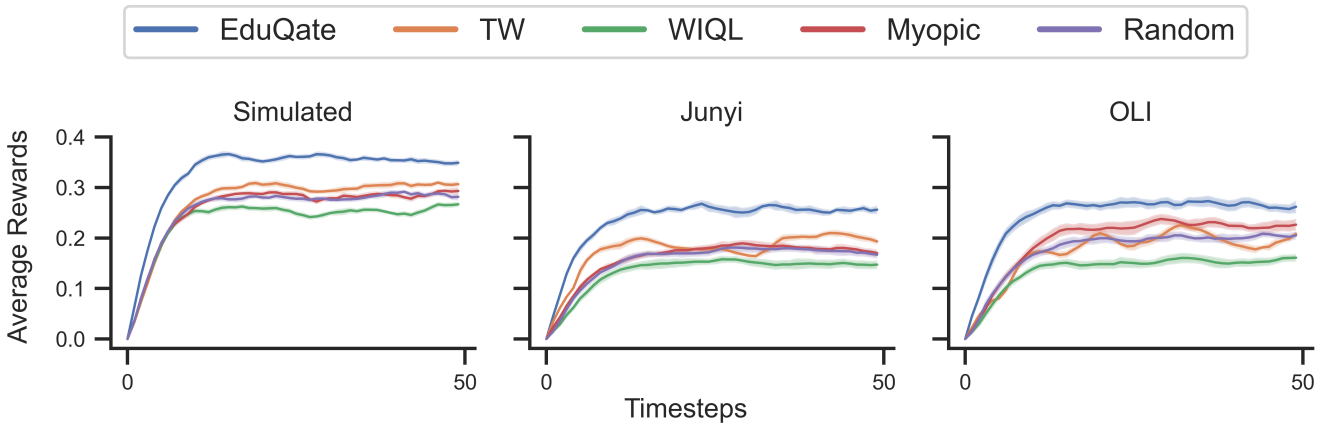

Figure 2: Average rewards for the respective algorithms on 3 datasets, averaged across 30 runs. Shaded regions represent standard error.

5.1.1 Synthetic dataset. Given the domain-motivated constraints on the transition functions highlighted in Section 3.1, we create a simulator based on $N = 50$, $S \in \{0, 1\}$, $N_{\text{topics}} = 20$. We randomly assign arms to topic groups, and allow arms to be assigned to be more than one topic. Under this method, number of arms under each group may not be equal. For each trial, a new transition matrix is generated to simulate distinct student scenarios.

5.1.2 Junyi dataset. The Junyi dataset [7] is an extensive dataset collected from the Junyi Academy [1], an eLearning platform established in 2012 on based on the open-source code released by Khan Academy. In this dataset, there are nearly 26 million student-exercise interactions across 250 000 students in its mathematics curriculum, organized into 21 topics and 9 areas. For this experiment, we selected the top 100 exercises with the most student interactions to create our student models and assign these exercises to their group based on the topic. In addition, Junyi dataset provides expert annotated similarity ratings between exercise pairs, which we use to further enrich the groupings. This results in a more complex network beyond simple topical groups. Using our method to generate groups, the resultant EdNetRMAB has $N = 100$ and $N_{topics} = 21$.

5.1.3 OLI Statics dataset. The OLI Statics dataset [4] comprises student interactions with an online Engineering Statics course[2]. In this dataset, each item is assigned one or more Knowledge Components (KCs) based on the related topics. After filtering for the top 100 items with the most student interactions, the resultant EdNetRMAB includes $N = 100$ items and $N_{topics} = 76$ distinct topics.

5.2 Creating student models

In this section, we outline the procedure for generating student models aimed at simulating the learning process. To clarify, a student model in this context is defined as a set of transition matrices for all items. These matrices are employed with EdNetRMABs to simulate the learning dynamics.

We employ various strategies to model transitions within the RMAB framework. Active transitions are determined by assessing the average success rate on a question before and after a learning intervention. Passive transitions are influenced by difficulty ratings, with more challenging questions more prone to rapid forgetting. Semi-active transitions, on the other hand, are computed as proportion of active transition, guided by similarity scores.

Active Transitions. We use data on students' correct response rate after interacting with an item to create the transition matrix for action 2, based on the change in correctness rates before and after a learning intervention.

Passive Transitions. To construct passive transitions for items, we use relative difficulty scores to determine transitions based on difficulty levels. We assume that higher difficulty correlates with a greater likelihood of forgetting, resulting in higher failure rates. Specifically, higher difficulty values correspond to higher $P^0_{1,0}$ values, indicating a greater likelihood of forgetting. The transition matrix for the passive action $a = 0$ is then randomly generated, with values influenced by difficulty levels.

Semi-active Transitions. To derive semi-active transitions, we use similarity scores between exercises from the Junyi dataset. We first normalize these scores to the range $[0, 1]$. Then, for any chosen arm, we compute its transition matrix under the semi-active action $a = 1$ as a proportion of its active action transitions, $P^1_{0,1} = \sigma(P^2_{0,1})$, where σ signifies the similarity proportion. The arm's transition matrix for the semi-active action varies due to different similarity scores between pairs in the same group. To address this, we use the average similarity score to determine the proportion. Since the OLI dataset does not contain similarity ratings, we assume a constant similarity rating of $\sigma = 0.8$ for all pairs.

6 RESULTS

The experimental results for the synthetic, Junyi, and OLI datasets, presented in Table 1 and Figure 2, demonstrate the performance

[1] http://www.Junyiacademy.org/
[2] https://oli.cmu.edu/courses/engineering-statics-open-free/

Table 1: Comparison of policies on synthetic, Junyi, and OLI datasets. $\mathbb{E}[R]$ represents the average reward obtained in the final episode of training. Statistic after ± represents standard error across 30 trials.

Policy	Synthetic		Junyi		OLI	
	$\mathbb{E}[IB](\%)\pm$	$\mathbb{E}[R]\pm$	$\mathbb{E}[IB](\%)\pm$	$\mathbb{E}[R]\pm$	$\mathbb{E}[IB](\%)\pm$	$\mathbb{E}[R]\pm$
Random	-	26.84 ± 0.46	-	15.82 ± 0.34	-	18.46 ± 0.35
WIQL	-7.93 ± 1.59	24.60 ± 0.43	-10.79 ± 4.93	14.01 ± 0.97	-21.34 ± 2.82	14.33 ± 0.42
Myopic	1.02 ± 1.35	27.07 ± 0.52	7.02 ± 1.73	16.86 ± 0.36	12.25 ± 3.39	20.51 ± 0.48
TW	6.45 ± 1.23	28.50 ± 0.47	17.75 ± 1.77	18.53 ± 0.28	-1.124 ± 2.12	18.07 ± 0.211
EduQate	**28.61 ± 2.11**	**34.33 ± 0.49**	**56.33 ± 3.10**	**24.53 ± 0.31**	**39.54 ± 3.74**	**25.47 ± 0.47**

of five algorithms: EduQate, TW, WIQL, Myopic, and Random. We report the average *IB* and final episode rewards from thirty independent runs for each algorithm. Across all datasets, EduQate consistently outperforms the other policies, showcasing higher intervention benefits and average rewards.

A notable observation is that in some cases, WIQL and Myopic policies report negative or negligible *IB* values, indicating their inability to surpass the performance of the random policy. This aligns with prior research by Doroudi et al. [9], which highlighted the robustness of random policies in educational settings. Our results further confirm that random policies can be challenging to outperform, even when algorithms are equipped with knowledge of the learning dynamics.

The superior performance of our interdependency-aware EduQate over random policies and other algorithms underscores the importance of considering network effects and interdependencies in EdNetRMABs. This suggests that accounting for the complex relationships between learning topics can lead to more effective educational interventions.

WIQL, which relies solely on Q-learning for active and passive actions, performs worse than a random policy as noted by its negative *IB* across the three datasets. This underperformance is likely due to its tendency to misattribute positive network effects to passive actions, highlighting the limitations of traditional reinforcement learning approaches in this context where assuming independence amongst arms can fail.

Interestingly, despite having access to the transition matrix, TW does not perform as well as EduQate. In particular, TW was not able to beat the random policy on the OLI dataset. While TW has demonstrated effectiveness in traditional RMABs, its weaknesses become evident in the current setting, where pulling an arm has wider implications for other arms. This observation emphasizes the unique challenges posed by educational environments and the need for specialized algorithms like EduQate.

Figure 2 provides a visual representation of the average rewards obtained in the final episode for each algorithm, further illustrating EduQate's superior performance across different datasets.

The synthetic dataset produces networks with distinct isolated groups, in contrast to the more intricate and interconnected networks from the Junyi and OLI datasets. Compared to synthetic dataset, real-world educational environments presents a greater degree of intricacy and challenges for teacher algorithms. Despite that, EduQate demonstrates robust and effective performance in maximizing rewards. This consistency across different network setups further validates EduQate's adaptability and efficacy in diverse educational contexts.

Table 2: Performance comparison under corrupted arm groupings across 5 trials. Corrupted groupings were created by randomly reassigning 30% or 50% of arms to different groups.

Policy	30% Corruption		50% Corruption	
	$\mathbb{E}[IB](\%)\pm$	$\mathbb{E}[R]\pm$	$\mathbb{E}[IB](\%)\pm$	$\mathbb{E}[R]\pm$
Random	-	7.36 ± 0.46	-	7.46 ± 0.49
WIQL	-12.44 ± 6.16	6.41 ± 0.49	-9.34 ± 5.65	6.74 ± 0.54
TW	17.62 ± 4.71	8.58 ± 0.33	15.94 ± 5.62	8.55 ± 0.38
EduQate	8.68 ± 5.22	7.92 ± 0.33	11.59 ± 6.68	8.23 ± 0.46

In conclusion, our results not only demonstrate the superiority of EduQate but also highlight the importance of considering network effects and interdependencies in educational settings. Future research could explore the impact of different network structures on algorithm performance and investigate ways to further optimize EduQate for specific educational contexts. We explore the effects of different network topologies by varying the number of topics while limiting the membership of each item. We find that as network interdependencies are reduced, the network effects diminish, and such EdNetRMABs can be approximated to traditional RMABs with independent arms. Under these conditions, our algorithm's advantage is reduced.

6.1 Ablation Studies

6.1.1 Misspecified Network Relations. In our previous analyses, we assumed that the network and relations between arms accurately reflected true groupings. However, real-world scenarios may involve misspecified relationships due to incomplete data mining methods or insufficient data for accurate modeling. This scenario is particularly likely when new items are added to the pool, and data mining methods lack sufficient data to establish accurate relations. To evaluate EduQate's robustness to such misspecifications, we conducted an ablation study with altered arm groupings, hypothesizing that the interdependency-aware EduQate would be adversely affected, while other methods would remain relatively stable.

To test this hypothesis, we modified the original Junyi dataset by randomly reassigning the relations of a percentage of arms, creating a 'corrupted' dataset. Table 2 presents the performance metrics for this scenario. The results show that EduQate outperforms WIQL

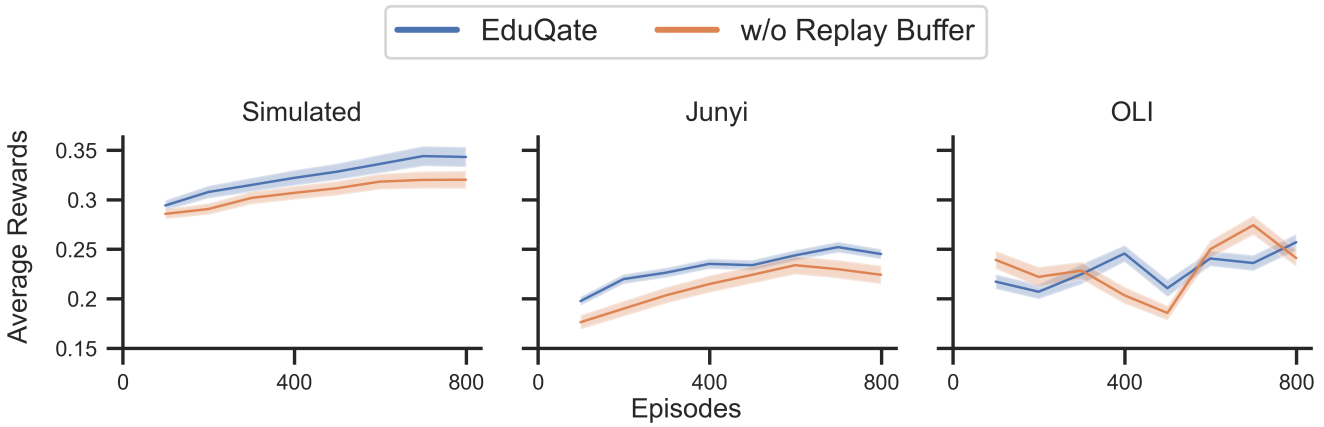

Figure 3: Average rewards across 800 episodes of training, across 30 seeds. Experience Replay Buffer helps EduQate achieve stronger results across all datasets.

and Random policies, but does not surpass TW under these conditions. Similar to WIQL, this reduced performance likely stems from erroneous attributions in EduQate: positive effects may be wrongly assigned to passive actions, while negative effects may be incorrectly attributed to pseudo-actions. However, EduQate still maintains superior performance over the random policy, a robust baseline in educational settings, demonstrating its resilience even with imperfect information.

WIQL can be interpreted as an extreme case of misspecification in the EduQate framework, where $\phi_i = \emptyset$ for all i. This perspective provides insight into the relative performance of these algorithms under varying degrees of network misspecification and highlights the importance of considering interdependencies in educational recommendation systems.

Crucially, EdNetRMABs allow for easy updates to the network structure through expert intervention or improved data mining methods. This flexibility enables system refinement over time, potentially mitigating initial misspecifications and enhancing real-world performance. These findings highlight the importance of accurate relationship modeling for optimal EduQate performance while demonstrating its robustness in suboptimal conditions. The adaptability of EdNetRMABs suggests promising avenues for ongoing improvement in educational applications.

6.1.2 Ablation of Replay Buffer. We investigate the importance of the Experience Replay buffer in EduQate, as shown in Figure 3 and Table 3. For the Simulated and Junyi datasets, EduQate without Experience Replay buffer does not achieve the performance levels of the full EduQate algorithm within 800 episodes, highlighting the importance of methods that aid Q-learning convergence. In real-world applications, slow convergence can result in students experiencing a curriculum similar to a random policy, leading to sub-optimal learning experiences during the early stages. This issue is known as the cold-start problem [3]. Future work in EdNetRMABs should explore methods to overcome cold-start problems and improve convergence in Q-learning-based methods.

Table 3: Comparison of EduQate with and without Experience Replay Buffer policies across different datasets.

Policy	$\mathbb{E}[IB]$ (%) ±		
	Synthetic	Junyi	OLI
EduQate	28.61 ± 2.11	56.33 ± 3.10	39.54 ± 3.74
w/o Replay Buffer	19.73 ± 1.35	40.35 ± 2.42	37.55 ± 3.20
Policy	$\mathbb{E}[R]$ ±		
	Synthetic	Junyi	OLI
EduQate	34.33 ± 0.49	24.53 ± 0.31	25.47 ± 0.47
w/o Replay Buffer	32.03 ± 0.47	22.13 ± 0.54	24.71 ± 0.44

7 CONCLUSION AND LIMITATIONS

We introduced EdNetRMABs, a MAB variant designed for modeling interdependencies in educational content, and proposed EduQate, a novel Whittle-based learning algorithm. EduQate computes optimal policies without requiring transition matrix knowledge while accounting for network effects. We demonstrated its optimality theoretically and effectiveness empirically through synthetic and real-world experiments. While our work assumes fully observable student knowledge states, it provides a foundation for future research. Future directions include extending EduQate to handle partially observable states, addressing the cold-start problem, and incorporating additional learning factors such as student motivation and engagement.

ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation, Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Awards No: AISG2-RP-2020-017) and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2022-01-025).

REFERENCES

[1] Richard C Atkinson. 1972. Ingredients for a theory of instruction. *American Psychologist* 27, 10 (1972), 921.

[2] Aqil Zainal Azhar, Avi Segal, and Kobi Gal. 2022. Optimizing Representations and Policies for Question Sequencing Using Reinforcement Learning. *International Educational Data Mining Society* (2022).

[3] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C Mitchell. 2020. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[4] Norman Bier. 2011. OLI Engineering Statics - Fall 2011 (114 students). https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=590

[5] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learn to Intervene: An Adaptive Learning Policy for Restless Bandits in Application to Preventive Healthcare. *arXiv preprint arXiv:2105.07965* (2021).

[6] Colton Botta, Avi Segal, and Kobi Gal. 2023. Sequencing Educational Content Using Diversity Aware Bandits. (2023).

[7] Haw-Shiuan Chang, Hwai-Jung Hsu, and Kuan-Ta Chen. 2015. Modeling Exercise Relationships in E-Learning: A Unified Approach.. In *EDM*. 532–535.

[8] Shayan Doroudi, Vincent Aleven, and Emma Brunskill. 2017. Robust evaluation matrix: Towards a more principled offline exploration of instructional policies. In *Proceedings of the fourth (2017) ACM conference on learning@ scale*. 3–12.

[9] Shayan Doroudi, Vincent Aleven, and Emma Brunskill. 2019. Where's the reward? a review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education* 29 (2019), 568–620.

[10] Hermann Ebbinghaus. 1885. *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

[11] Derek Green, Thomas Walsh, Paul Cohen, and Yu-Han Chang. 2011. Learning a skill-teaching curriculum with dynamic bayes nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25. 1648–1654.

[12] Christine Herlihy and John P. Dickerson. 2022. Networked Restless Bandits with Positive Externalities. arXiv:2212.05144 [cs.LG]

[13] Andrew S Lan and Richard G Baraniuk. 2016. A Contextual Bandits Framework for Personalized Learning Action Selection.. In *EDM*. 424–429.

[14] Dexun Li and Pradeep Varakantham. 2023. Avoiding Starvation of Arms in Restless Multi-Armed Bandits. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1303–1311.

[15] Long-Ji Lin. 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning* 8 (1992), 293–321.

[16] Keqin Liu and Qing Zhao. 2010. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory* 56, 11 (2010), 5547–5567.

[17] Aditya Mate, Jackson A Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing bandits and their application to public health interventions. *arXiv preprint arXiv:2007.04432* (2020).

[18] Christos H Papadimitriou and John N Tsitsiklis. 1994. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*. IEEE, 318–322.

[19] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in neural information processing systems* 28 (2015).

[20] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. 123–131.

[21] Avi Segal, Yossi Ben David, Joseph Jay Williams, Kobi Gal, and Yaar Shalom. 2018. Combining difficulty ranking with multi-armed bandits to sequence educational content. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19*. Springer, 317–321.

[22] Shitian Shen, Markel Sanz Ausin, Behrooz Mostafavi, and Min Chi. 2018. Improving learning & reducing time: A constrained action-based reinforcement learning approach. In *Proceedings of the 26th conference on user modeling, adaptation and personalization*. 43–51.

[23] Anni Siren and Vassilios Tzerpos. 2022. Automatic learning path creation using OER: a systematic literature mapping. *IEEE Transactions on Learning Technologies* (2022).

[24] Shiwei Tong, Qi Liu, Wei Huang, Zhenya Hunag, Enhong Chen, Chuanren Liu, Haiping Ma, and Shijin Wang. 2020. Structure-based knowledge tracing: An influence propagation view. In *2020 IEEE international conference on data mining (ICDM)*. IEEE, 541–550.

[25] Utkarsh Upadhyay, Abir De, and Manuel Gomez Rodriguez. 2018. Deep reinforcement learning of marked temporal point processes. *Advances in Neural Information Processing Systems* 31 (2018).

[26] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3 (1992), 279–292.

[27] Peter Whittle. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* (1988), 287–298.