Get a Head Start: On-Demand Pedagogical Policy Selection in Intelligent Tutoring

Ge Gao[1], **Xi Yang**[2], **Min Chi**[1]

[1]Department of Computer Science, North Carolina State University
[2]IBM Research
ggao5@ncsu.edu, xi.yang@ibm.com, mchi@ncsu.edu

Abstract

Reinforcement learning (RL) is broadly employed in human-involved systems to enhance human outcomes. Off-policy evaluation (OPE) has been pivotal for RL in those realms since online policy learning and evaluation can be high-stake. Intelligent tutoring has raised tremendous attentions as highly challenging when applying OPE to human-involved systems, due to that students' subgroups can favor different pedagogical policies and the costly procedure that policies have to be induced fully offline and then directly deployed to the upcoming semester. In this work, we formulate on-demand pedagogical policy selection (ODPS) to tackle the challenges for OPE in intelligent tutoring. We propose a pipeline, EDUPLANNER, as a concrete solution for ODPS. Our pipeline results in an theoretically unbiased estimator, and enables efficient and customized policy selection by identifying subgroups over both historical data and on-arrival initial logs. We evaluate our approach on the Probability ITS that has been used in real classrooms for over eight years. Our study shows significant improvement on learning outcomes of students with EDUPLANNER, especially for the ones associated with low-performing subgroups.

Introduction

Reinforcement learning (RL) is extensively investigated in human-involved systems, such as healthcare and education, to facilitate decision-making process and enhance human outcomes. In the realm of modern education, intelligent tutoring systems (ITS) have been used worldwide to enhance students' engagement and improve learning outcomes (Chi et al. 2011; VanLehn 2006), where RL has been employed to induce automatic pedagogical policies (Liu et al. 2022; Zhou et al. 2022). Off-policy evaluation (OPE), which estimates the performance of target policies using historical data collected under a (different) behavior policy (Chandak et al. 2022; Gao et al. 2022, 2023a,b; Nie et al. 2022), has been pivotal for RL in human-involved systems, since online policy learning and evaluation can be high-stake. However, OPE, or off-policy selection (OPS) that selects policies based on OPE estimations (Yang et al. 2022), for ITS is highly challenging and prior works noted that directly applying some OPE/OPS methods can even lead to ineffective policy selection (Rowe, Mott, and Lester 2014; Shen and Chi 2016). That can be caused by two natures in real-world classrooms: *(i)* Prior works generally select the policies targeting the entire population by maximizing the OPE estimations, which may not fit the education domain well, where subgroups exist among students and can perform diversely under heterogeneous policies (Gao et al. 2024; Yang et al. 2020b); *(ii)* The entire process of deploying a policy is costly, where policies have to be induced fully offline and then directly deployed to the next semester. And only a small set of policies can be deployed due to the expensive recruitment of participants and safety concern. There is an urgent need for effective and efficient OPS in ITS.

In the context of ITS, the task of providing customized policies for each student can be formulated as the *on-demand pedagogical policy selection (ODPS)* problem, as illustrated in Figure 1, where the ODPS decides a policy to be deployed to each student from a small set of candidate policies that are trained using historical data, given the initial log of the student. ODPS is distinguished from typical OPS in three aspects: (*i*) OPS generally assumes only historical data is available, while ODPS can observe the initial log from each student, which makes it possible for ODPS to capture the specific needs from each student; (*ii*) OPS usually assumes selecting models or tuning hyperparameters with unlimited times during policy training (Kumar et al. 2022; Nie et al. 2022), while it is extremely inefficient for ODPS to retrain and evaluate a large number of candidate policies due to the expert sanity check on each policy, which is required and operated by independent departments or institutions for safety and ethics concerns and considered as 'black-box'; (*iii*) The major goal of prior OPS approaches is to select the best policy over the entire population, while ODPS aims to real-timely decide the best policy for each student who arrives the ITS depending on the initial log of the student.

In this work, we propose an on-demand policy selector, named **EDUPLANNER**, to provide tailored policies for different subgroups, as opposed to existing off-policy selection works that aim to select a single policy over the entire population. Specifically, we propose a novel subgroup partitioning technique, which only requires the initial states as inputs. It can efficiently capture various user behaviors, which is especially valuable in real-world scenarios where subgrouping based on prior knowledge is time-consuming and labor-intensive. Moreover, given limited data access in real-world

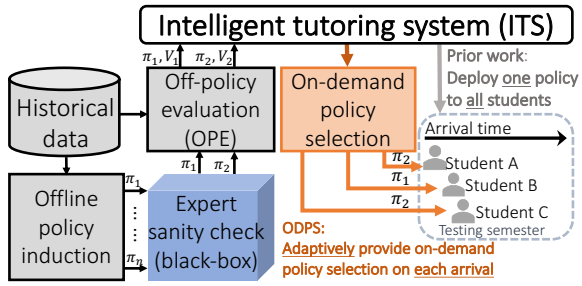

Figure 1: A conceptual illustration of on-demand pedagogical policy selection (ODPS) in ITS. ODPS decides a policy for each arrived student, given a small set of target policies trained with historical data and passed sanity check.

applications, we additionally introduce a variational autoencoder (VAE)-based augmentation model, to capture the user-system interactive dependencies within each subgroup, where the reconstructed samples can improve the state-action coverage of the dataset.

The key contributions of this work are summarized as follows: (*i*) To the best of our knowledge, we are the first to formulate the on-demand policy selection (ODPS) problem which is unmet but critical in the context of human-involved systems, especially education, and consequently provide a concrete pipeline for ODPS named EDUPLANNER to provide tailored policies for different subgroups. The resulting estimator is theoretically unbiased. (*ii*) We conduct extensive experiments to evaluate EDUPLANNER with students on a Probability ITS which has been used in colleges for over 8 years. EDUPLANNER not only significantly improved students' learning outcomes, compared to expert-designed policy and the policy selected by state-of-the-art OPE methods, respectively, but improved the performance of low-performing subgroups found in the ITS, on which existing baselines can even have negative effects. (*iii*) The proposed off-policy selector can be stand-alone to allow for any future build-on-top works such as policy optimization and representation learning in various applications.

Related Works

RL for ITS. In modern ITS, RL has raised attention to enhance the engagement and performance of students. For many forms of online learning environments, the system's behavior can be viewed as a sequential decision process; wherein, at each discrete step, the system is responsible for selecting the next action to take (Chi et al. 2011). *Pedagogical policies* decide the next system action when there are multiple ones available, with goals to support learners' cognition, emotions or outcomes (Chi et al. 2011; Abdelshiheed et al. 2023, 2020). Previous studies have demonstrated that RL can induce effective pedagogical policies in ITS (Shen and Chi 2016; Mandel et al. 2014; Wang et al. 2017; Zhou et al. 2022; Sanz Ausin et al. 2020). For example, Wang et al. (Wang et al. 2017) applied a variety of deep RL approaches to induce pedagogical policies with the goal to improve students' normalized learning gain in an educational game. The simulation evaluation revealed that the deep RL policies were more effective than a linear model-based RL policy. Zhou et al. (Zhou et al. 2022) applied hierarchical reinforcement learning (HRL) to improve students' normalized learning gain in a Discrete Mathematics course, and the HRL-induced policy was more effective than the Deep Q-Network induced policy. Since online evaluation is high-stakes in practice, effective off-policy evaluation and selection are in high demand when applying RL in ITS.

Off-policy selection (OPS). Off-policy selection is typically approached via off-policy evaluation (OPE), which estimates the expected return of target policies using historical data under a behavior policy. A variety of contemporary OPE methods has been proposed, which can be mainly divided into three categories (Voloshin et al. 2021): (*i*) Inverse propensity scoring (Precup 2000), such as Importance Sampling (IS) (Doroudi, Thomas, and Brunskill 2017). (*ii*) Direct methods that directly estimate the value functions of the evaluation policy (Nachum et al. 2019; Zhang et al. 2021; Yang et al. 2022), including but not limited to model-based estimators (MB) (Paduraru 2013; Zhang et al. 2021), value-based estimators (Le, Voloshin, and Yue 2019) such as Fitted Q Evaluation (FQE), and minimax estimators (Zhang, Liu, and Whiteson 2020; Voloshin, Jiang, and Yue 2021) such as DualDICE (Yang et al. 2020a). (*iii*) Hybrid methods combine aspects of both inverse propensity scoring and direct methods (Thomas and Brunskill 2016), such as DR (Jiang and Li 2016). In practice, due to expensive online evaluations, researchers generally selected the policy with the highest estimated returns via OPE (Mandel et al. 2014; Gao et al. 2023c). For example, Mandel et al. selected the policy with the maximum OPE estimations to be deployed to an educational game (Mandel et al. 2014). Recently, some works focused on estimator selection or hyperparameter tuning in OPS (Nie et al. 2022; Miyaguchi 2022; Kumar et al. 2022; Lee et al. 2022). However, retraining policies may not be feasible in ODPS due to the time- and resource-consuming procedure. More importantly, prior work generally selected policies targeting the entire population, while personalized policy is flavored towards individual needs.

The On-demand Pedagogical Policy Selection (ODPS) Problem

Formally, we consider framing an agent's interaction with the environment over a sequence of decision-making steps as a Markov decision process (MDP), which is formulated as a 6-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{S}_0, r, \gamma)$. $\mathcal{S}$ is the state space. $\mathcal{A}$ is the action space. $\mathcal{P}$ defines transition dynamics from the current state and action to the next state. $\mathcal{S}_0$ defines the initial state distribution. r is the reward function. $\gamma \in (0, 1]$ is discount factor. Episodes are of finite horizon T. At each time-step t, the agent observes the state $s_t \in \mathcal{S}$ of the environment, then chooses an action $a_t \in \mathcal{A}$ following a policy π. The environment accordingly provides a reward $r_t = r(s_t, a_t)$, and the agent observes the next state s_{t+1} determined by $\mathcal{P}$. A *trajectory* is defined as $\tau^{(i)} = [\ldots, (s_t, a_t, r_t, s'_t), \ldots]_{t=1}^{T}$.

Given an arrived person with an initial log s_0 and a set of target (evaluation) policies, the goal of on-demand ped-

agogical policy selection (ODPS) is to select one policy π with the highest estimation from off-policy evaluation (OPE), V^π, among all target policy trained using historical trajectories $\mathcal{D}$ collected over a *behavioral* policy $\beta \neq \pi$. OPE estimates the expected total return over the *target* policy π, $V^\pi = \mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t | a_t \sim \pi]$. The historical trajectory set $\mathcal{D} = \{..., \tau^{(i)}, ...\}_{i=1}^{N}$ consist of N trajectories.

ODPS via EDUPLANNER

In this work, we propose an ODPS pipeline, named EDUPLANNER, based on student subgroup partitioning over initial logs. EDUPLANNER mainly consists of three steps: (*i*) subgroup partitioning, which determines the subgroups from historical data; (*ii*) sample enrichment, which aims to improve OPE with augmented historical samples by providing more state-action visitation space considering the Markovian characteristics within trajectories; (*iii*) on-demand policy selection based on subgroup partitioning, which real-timely monitors arrived students and estimates the best policy for each student by determining the possible subgroup of the student using his or her initial log.

ODPS with Subgroup Partitioning

We first formally define the ODPS with subgroup partitioning, and prove its resulting estimator is unbiased, which is essential for general OPE problems. We assume a subgroup partitioning over the initial state space by $\mathbf{K} = \{K_1, \ldots, K_M\}$ such that $\cup_{i=1}^{M} K_i = \mathcal{S}_0$ and $K_i \cap K_j = \emptyset, \forall i, j$. The partition function $K(s_0; \mathbf{K}) = K_i$ such that $s_0 \in K_i$. Thus, given a partition $\mathbf{K}$, the value function for a target policy π is

$$V^\pi(s_0; \mathbf{K}) = \mathbb{E}_{s_0' \sim \mathcal{S}_0}[\sum_{t=1}^{T} \gamma^{t-1} r_t | s_0 = s_0', s_0' \in K(s_0; \mathbf{K}), a_t \sim \pi]. \quad (1)$$

In pedagogical policy design, a general goal is to induce a policy which outperforms expert policy that is defined by educational experts according to their prior knowledge in instructing students (Mandel et al. 2014; VanLehn 2006; Zhou et al. 2022). Therefore, under a subgroup partitioning $\mathbf{K}$ over initial state s_0, the goal for ODPS is to select the policy π with the greatest value distance $G^{\beta,\pi}(s_0; \mathbf{K})$:

$$G^{\beta,\pi}(s_0; \mathbf{K}) = V^\pi(s_0; \mathbf{K}) - V^\beta(s_0; \mathbf{K}). \quad (2)$$

The value distance is constant within each subgroup K_i. Therefore, following (Keramati et al. 2022), we can define an estimator $\hat{G}^{\beta,\pi}(s_0; \mathbf{K})$ of $G^{\beta,\pi}(s_0; \mathbf{K})$, such that

$$\hat{G}^{\beta,\pi}(s_0; \mathbf{K}) = \frac{1}{|\{s_0^{(i)} | s_0^{(i)} \in K(s_0; \mathbf{K})\}|} \sum_{i | s_0^{(i)} \in K(s_0; \mathbf{K})} (w_i \sum_{t=1}^{T} \gamma^{t-1} r_t^{(i)} - \sum_{t=1}^{T} \gamma^{t-1} r_t^{(i)}), \quad (3)$$

where $w_i := \Pi_{t=1}^{T} \pi(a_t^{(i)} | s_t^{(i)}) / \beta(a_t^{(i)} | s_t^{(i)})$. $\hat{G}^{\beta,\pi}(s_0; \mathbf{K})$ is an unbiased estimator of $G^{\beta,\pi}(s_0; \mathbf{K})$ in each subgroup, which is an important characteristic for OPE.

Subgroup Partitioning

We design to use initial logs of students for subgroup partitioning for two main reasons. First, initial logs may reflect not only the background knowledge of students but their interaction habits (Gao, Marwan, and Price 2021), without specific information related to behavior policies that may be distracting for subgroup partitioning. Though some existing works utilize demographics or grades of students from their prior taken courses to identify student subgroups (Castro-Wunsch, Ahadi, and Petersen 2017; Sinclair, Martin, and Michel 1999), it may not be feasible in practice due to the protection of student information by institutions. Second, prior works have found that initial logs can be informative to indicate learning outcomes of students (Mao 2019), which makes it possible for ITS to customize the policies with the goal of improving learning outcomes for each subgroup.

However, there is a challenge with subgroup partitioning over the initial logs of students. The state space of student logs in ITS is usually high-dimensional, due to the detailed capture of each step taken during interaction and associated timing information (Chi et al. 2011; Mandel et al. 2014). For example, in this study, 142 features have been recorded. While some features might be irrelevant for downstream data mining tasks, it is challenging to determine their relevance a priori (Mandel et al. 2014). To solve this, we introduce a data-driven feature taxonomy over the state features of students, then perform subgroup partitioning with distilled features based on the feature taxonomy.

A data-driven feature taxonomy. Educational researchers have used feature taxonomy in qualitative ways to support instructors subgroup students and understand behaviors of students (Marwan, Dombe, and Price 2020). Unlike prior approaches that are expensive requiring much effort from human experts, we design a data-driven feature taxonomy for a straightforward student subgroup partitioning that may reflect the knowledge background and dynamic learning progress of students. We define three possible categories of a given feature x according to its temporal and cross-sample characteristics:

$$x \in \begin{cases} \text{system-static}, & \text{if } x_t^{(i)} = x_t^{(j)}, \forall i, j \in N, t \in T \\ \text{student-centric}, & \text{if } x_0^{(i)} \neq x_0^{(j)}, \exists i, j \in N \\ \text{interaction-driven}, & \text{otherwise.} \end{cases} \quad (4)$$

Specifically, (i) *System-static*: the features, which are static across students on the same problem, are assumed to be assigned by the system; (ii) *Student-centric*: the features, which differ across students from the initial logs and may change over time, is assumed to be students-centric and reflect both students' initial knowledge background and the changes of individual underlying mindset during learning; (iii) *Interaction-driven*: the features, which contain characteristics from both system-assigned and student-centric types, are assumed to be mixed-style features that are affected by both system and individuals. Table 1 shows examples in Probability ITS with the defined data-driven taxonomy.

Subgroup partitioning with distilled features via feature taxonomy. Since system-assigned features are mainly dominated by system design and remain static across students

Taxonomy	Examples	Perc.
System-static	Problem difficulty	18%
Student-centric	Number of hints requested	48%
Interaction-driven	Number of `tells` since `elicit`	34%

Table 1: Feature examples and percentage in Probability ITS using the data-driven taxonomy.

on each problem, for the purpose of subgroup partitioning, we focus on the two categories of features, student-centric and interaction-driven, since both may be highly associated with students' underlying mental status and behaviors, for which we call *student-sensitive* features. In this work, we identify $82\%(117)$ from overall 142 features as student-sensitive features and used them for subgroup partitioning. Specifically, to learn the subgroups, we leverage Toeplitz inverse covariance-based clustering (TICC) (Hallac et al. 2017) to map initial logs $\mathcal{S}_0$ into M clusters based on the values of student-sensitive features, where each $s_0 \in \mathcal{S}_0$ is associated with a cluster from the set $\mathbf{K} = \{K_1, \ldots, K_M\}$. The initial logs that are mapped to the same cluster can be considered to share the graphical connectivity structure of cross-features information captured by TICC. We consider using TICC because of its superior performance in clustering compared to traditional distance-based methods such as K-means, especially with human behavior-related tasks (Hallac et al. 2017; Yang, Zhang, and Chi 2021). The number of clusters can be determined by silhouette scores following (Hallac et al. 2017). Note that we exhibit TICC as an example in our proposed pipeline, while it can be replaced by other partitioning approaches if needed. Then, we assume subgroup partitioning is consistent with cluster assignments associated with initial logs, *i.e.*, students whose initial logs are associated with the same cluster index are considered from the same subgroup.

Samples Enrichment

The amount of student samples within each subgroup is usually limited, due to the high cost of recruiting participants. For example, in this work, one subgroup only contains 45 students from the training set. Such training data can contain limited visitation of state and action spaces and have a substantial influence on the downstream policy selection (Nie et al. 2022). Latent-model-based data augmentation has been commonly employed in previous offline RL (Hafner et al. 2020; Lee et al. 2020; Rybkin et al. 2021). However, prior works generally collect online interaction data, which may not be the case for on-demand pedagogical policy selection problems. To enrich samples, we introduce an example of offline trajectory augmentation by adapting a variational autoencoder (VAE) to capture the MDP transitions underlying each subgroup.

Specifically, given offline trajectories $\mathcal{T}$ from samples in one subgroup K (we omit the subscript for expressional conciseness), the formulation of VAE in MDP consists of three major components, *i.e.*, (i) the latent prior $p(z_0)$ that represents the distribution of the initial latent states over $\mathcal{T}$; (ii) the encoder $q_\alpha(z_t|s_{t-1}, a_{t-1}, s_t)$ that encodes the MDP transitions into the latent space; (iii) the decoders $p_\zeta(z_t|z_{t-1}, a_{t-1})$, $p_\zeta(s_t|z_t)$, $p_\zeta(r_{t-1}|z_t)$ that reconstructs new samples.

Algorithm 1: EDUPLANNER.

Input: A set of target policies $\mathbf{\Pi}$, N students' trajectories from historical data $\mathcal{D}$.

Begin:

1: Calculate the number of subgroups M using $\mathcal{D}$.
2: Conduct feature distillation via data-driven taxonomy.
3: Obtain the subgroup partitioning $\mathbf{K} = \{K_1, \ldots, K_M\}$ on distilled features.
4: **for** each subgroup K_i **do**
5: Augment subgroup samples $\mathcal{T}$ with $\widehat{\mathcal{T}}$.
6: Calculate the value of each target policy $\pi \in \mathbf{\Pi}$ on $\mathcal{T} \cup \widehat{\mathcal{T}}$.
7: Select the best target policy $\pi_i^* \in \mathbf{\Pi}$ in subgroup K_i.
8: **end for**
9: **while** the ITS receives initial log s_0 from a new student **do**
10: Check the subgroup K_i of the student with $s_0 \in K_i$.
11: Deploy the policy π_i^* associated with subgroup K_i to the student.
12: **end while**

The training objective for the VAE in MDP is to maximize the evidence lower bound (ELBO), which consists of the log-likelihood of reconstructing the states and rewards, and regularization of the approximated posterior, *i.e.*,

$$\begin{aligned} ELBO(\alpha, \zeta) = \mathbb{E}_{q_\alpha}\Big[& \sum\nolimits_{t=0}^{T} \log p_\zeta(s_t|z_t) \\ & + \sum\nolimits_{t=1}^{T} \log p_\zeta(r_{t-1}|z_t) - KL\big(q_\alpha(z_0|s_0)||p(z_0)\big) \\ & - \sum\nolimits_{t=1}^{T} KL\big(q_\alpha(z_t|z_{t-1}, a_{t-1}, s_t)||p_\zeta(z_t|z_{t-1}, a_{t-1})\big)\Big]. \end{aligned} \tag{5}$$

Consequently, given a set of subgroup samples, $\mathcal{T}$, the VAE in MDP to the set can be trained to induce a set of new samples, denoted as $\widehat{\mathcal{T}}$. The need for augmented samples is further investigated and justified in Section .

EDUPLANNER

The framework of the pipeline is described in Algorithm 1. With EDUPLANNER, ITS can real-timely monitors each arrived student, estimate the best policy for each subgroup, and decide the policy to be deployed to each student according to his or her initial log. Such real-time operating manner is important for ITS in practice, which is different from prior works in policy selection that assume either only historical data is accessible or targeting samples/distributions are known (Keramati et al. 2022; Yang et al. 2022; Zhong et al. 2022). In practice, students may start learning irregularly according to their own schedules, which may create discrepancies in their start times. It poses a challenge when selecting policies based on population information or subgroup distribution in the target semester, which requires waiting for

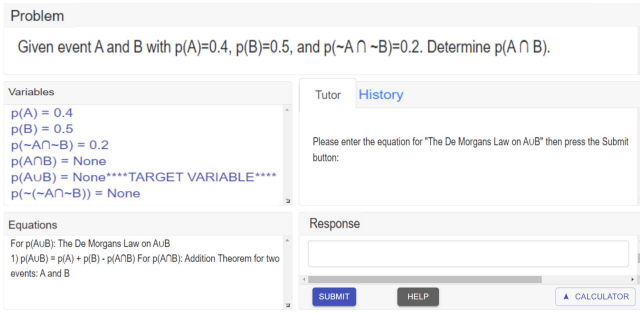

Figure 2: Probability ITS GUI. The problem statement window (*top*) presents the statement of the problem. The dialog window (*middle right*) shows the message the tutor provides to the students. Responses, e.g., writing an equation, are entered in the response window (*bottom right*). Any variables and equations generated through this process are shown on the variable window (*middle left*) and equation window (*bottom left*).

collecting all students' data. Note that we are the first to formally formulate the ODPS problem and this work is the first try to solve the problem with a pipeline framework that can cooperate with ITS in practice. In the prior sections, we exhibit technical examples that can be used in EDUPLANNER step-by-step, which can be substituted by advanced techniques in the future and allows build-on-top work with EDUPLANNER.

The Probability ITS

Though the problem setting and our method are general and can be applied to other interactive ITSs, we primarily focus on a Probability ITS used in an undergraduate STEM course at North Carolina State University, which was designed by domain experts and overseen by department committees, and has been extensively used by over $2,000$ students with $\sim$800k recorded interaction logs through eight academic years. It is designed to teach entry-level undergraduate students with ten major probability principles, including complement theorem, Bayes' rule, etc. Figure 2 presents the GUI of the Probability ITS.

Since students' underlying learning states are inherently unobservable (Mandel et al. 2014), the Probability ITS defined its state space with 142 features that could possibly capture students' learning status based on their interaction logs, as consulted with domain experts. The size of action space is 3 to decide the manner of solving the next problem, including a worked example (`WE`) (Sweller and Cooper 1985), problem-solving (`PS`), and a collaborative problem-solving worked example (`CPS`) (Schwonke et al. 2009). In `WE`s, students, observe how the tutor solves a problem; in `PS`s, students solve the problem themselves; in `CPS`s, the students and the tutor co-construct the solution. The rewards are sparse and defined as the normalized learning gain (Chi et al. 2011), $NLG = \frac{score_{posttexam} - score_{preexam}}{\sqrt{1 - score_{preexam}}}$, calculated by scores of students' pre- and post- exams taken before and after tutoring.

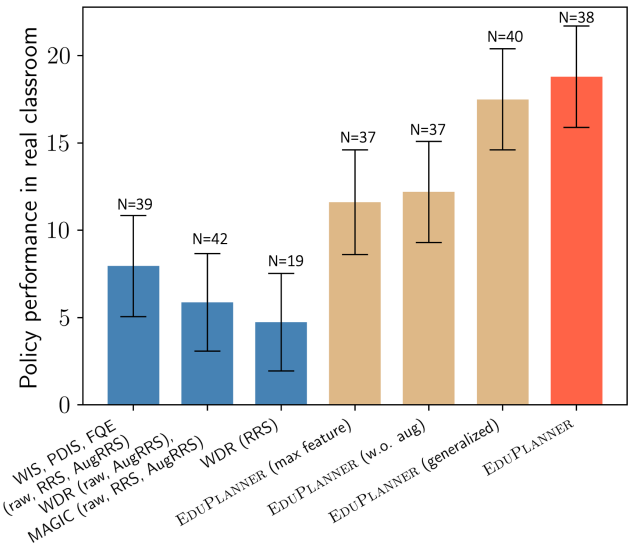

Figure 3: Students' performance under policies selected by different policy selection methods. OPE RRS represents incorporating OPE methods with repeated random sampling (RRS). OPE AugRRS represents OPE methods on augmented samples with RRS. Methods that selected the same policy are merged in one bin.

Empirical Experiments

In this study, we used $459k$ historical logs from $1,148$ students across five years for offline policy training and evaluation, and targeted the following semester using EDUPLANNER with the major goal to improve learning outcomes.

Classroom setup. All participants were entry-level undergraduates majoring in STEM and enrolled in the Probability course, while no demographics data or course grades were collected under an IRB protocol. They were recruited by an invitation email introducing procedure and purpose of this study. And they can opt-in without any influence on their course grades, and were allowed to opt-out any time during the study.

In the testing semester, due to fairness concern, each student was randomly assigned one policy from 4 target policies (i.e., 3 DQN-based policies and 1 expert policy). The chi-squared test was employed to check the relationship between policy assignment and subgroups, and it showed that the policy assignment cross subgroups were balanced with no significant relationship (p-value=0.479). In total, 140 students accomplished all procedures of the study.

Comparison Methods

OPE. Since our proposed method is an off-policy selector via OPE, we compared it to six popular OPE methods: Weighted IS (WIS) (Precup 2000), Per-Decision IS (PDIS) (Precup 2000), Fitted Q Evaluation (FQE) (Le, Voloshin, and Yue 2019), Weighted DR (WDR) (Thomas and Brunskill 2016), MAGIC (Thomas and Brunskill 2016), and Dual stationary DIstribution Correction Estimation (DualDICE) (Nachum et al. 2019). For FQE, as in (Le, Voloshin, and Yue 2019), we

OPE	OPE RRS	OPE AugRRS
9	6	**3**
EDUPLANNER (max. feature)	EDUPLANNER (w.o. aug.)	EDUPLANNER (generalized)
10	6	-4
EDUPLANNER		
3		

Table 2: Distance from estimated to true policy returns using each method. For *OPE*, *OPE RRS*, *OPE AugRRS*, results with the least gap between estimated and true returns among OPE methods (*i.e.*, WIS, FQE RRS, and FQE AugRRS, respectively).

train a neural network to estimate the value of the evaluation policy π_e by bootstrapping from $Q(o', a')$. For DualDICE, we use the open-sourced code in its original paper. For MAGIC, we use the implementation of (Voloshin et al. 2021).

Repeated random sampling. Considering the limited data collected from real students while it's hard to conduct online validation or retrain policies due to sanity checks by domain experts (Krishnan et al. 2016), we investigate a sample enrichment way, repeated random sampling (RRS) with replacement of the historical data to perform OPE, which has shown superior performance in some human-related tasks, such as disease treatment (Nie et al. 2022). We repeated 20 times of OPE with RRS to obtain an average value of policy estimations.

OPE on augmented samples with RRS. Since RRS doesn't assume MDP nature nor provide more state-action visitations in historical data, we perform OPE with RRS on augmented samples using our adapted VAE in MDP. We set the amount of augmented data identical to the amount of original historical data, *i.e.*, $|\widehat{\mathcal{T}}| = |\mathcal{T}| = N$, and RRS N samples from both set $\widehat{\mathcal{T}} \cup \mathcal{T}$ to perform OPE. 20 repeated times is used to obtain average values of estimations.

EDUPLANNER (max features). It ablates our feature distillation using proposed feature taxonomy, but uses all recorded features for subgroup partitioning.

EDUPLANNER (without augmentation). It ablates our sample enrichment module.

EDUPLANNER (generalized). It ablates policy selection on subgroup level, but uses the best policy estimated during off-policy evaluation over all subgroups. The policy that achieves the highest average estimated value across subgroups in historical data is selected.

Main Results

Figure 3 presents students' performance under policies selected by different methods in real classrooms. The overall performance of an OPS method was evaluated by averaging NLGs of students *who followed the selected policy* by the method, *i.e.*, students from each subgroup who followed the policies selected by EDUPLANNER, or students who followed one policy selected by a baseline. Overall, EDUPLANNER is the most effective policy selection leading to the greatest average NLGs. OPE-related methods tend to select sub-optimal policies which outperform domain expert policy. However, DualDICE estimates the performance of all target policies equally, thus unable to perform policy selection and we omit its results. With RRS and augmented samples, most OPE methods remain their choice with sub-optimal policy, suggesting that only using the modern data enrichment approaches that were effective in prior work with simulations (Nie et al. 2022) may not be enough to improve the policy selection in ITS.

We are also interested in the accuracy of estimation used for policy selection. Table 2 presents the distance from the estimated to true rewards of the policies selected by different methods. EDUPLANNER provides more accurate policy estimation with the smallest gap between true and estimated policy rewards, similar to OPE AugRRS (FQE AugRRS). With RRS and augmented samples, most OPE and their related methods could become more accurate in estimating policy performance, which may benefit from the richer state-action visitations provided by the augmented samples. However, OPE-related methods remain their choice with sub-optimal policy as shown in Figure 3. It suggests that the goal of maximizing the estimation accuracy may not be satisfying enough in tackling the on-demand policy selection problem, it is considerable that one can select the optimal policy from the candidates to facilitate students' learning.

The two variations of EDUPLANNER, EDUPLANNER (max features), and EDUPLANNER (generalized), select policies with better students' performance than compared OPE methods and their combinations with RRS and augmented samples, indicating the effectiveness of subgroup partitioning and the need of feature taxonomy. Though EDUPLANNER (max features) achieves the good performance of policy selection, it may easily overestimate a policy, probably due to the noise introduced by the system-assigned features. Moreover, both variations underperform EDUPLANNER in terms of both performance of policy selection and accuracy of estimations, indicating the effectiveness of our design with feature taxonomy and on-demand policy selection.

Further Discussions

For a more comprehensive understanding of our proposed work and students' behaviors in ITS, we further investigate the following two questions:

How does EDUPLANNER perform within students subgroups, especially low-performing subgroups? EDUPLANNER identified four subgroups (*i.e.*, K_1, K_2, K_3, K_4) using historical data from the Probability ITS. Specifically, $K_1(N_{his} = 345, N_{test} = 30)$ and $K_2(N_{his} = 678, N_{test} = 92)$ are majority groups with average NLG of 2 and 1, respectively, where EDUPLANNER selected the same policy as the best OPE-related methods with NLGs 18 and 14, respectively, in the testing semester. Both subgroups can be considered as high-performing subgroups with positive averaged NLGs across semesters. On the contrast, the other two subgroups, $K_3(N_{his} = 101, N_{test} = 12)$ and $K_4(N_{his} = 24, N_{test} = 6)$, contained less samples with respect to the entire population. In historical data, K_3 and K_4 showed negative average NLGs of -1 and -2, respectively. Both subgroups can be considered low performers, while they

	OPE (raw, RRS, AugRRS)	EDUPLANNER
K_3	-12_3 (N=4)	24_7 (N=4)
K_4	-20_{27} (N=3)	32_{11} (N=2)

Table 3: Performance (average NLGs$_{std}$) of students from low-performing subgroups under selected policies in the testing semester.

presented different learning behaviors. K_3 performed carelessly by keeping making quick movements within a short period, while K_4 abused hints but made much more mistakes during problem-solving.

Table 3 presents the NLGs of students from the low-performing subgroups K_3 and K_4 under policies selected by the best OPE-related methods and EDUPLANNER. With EDUPLANNER, both subgroups achieved significant improvement (average NLGs 24 and 32, respectively) compared to students in historical semesters. However, the sub-optimal policy chosen by comparison methods had a negative effect on both subgroups (average NLGs -12 and -20, respectively), while it had an overall positive effect across all students under that policy (see Figure 3). That particularly indicates the need for tremendous attention from educators to provide careful support for the low performers, since they can be more sensitive to policy selection.

Are OPEs with subgroup partitioning sensitive to information accumulated over time? Recall that OPE AugRRS selected sub-optimal policy, while their estimation accuracy (*i.e.,* absolute error) was improved compared to OPE and OPE RRS, and even similar to EDUPLANNER, where the augmentation was performed over all historical trajectories. We further investigate the effects of subgroup partitioning with longer trajectory information on OPE AugRRS performance. We conduct subgroup partitioning over the length of trajectories, *i.e.*, perform subgroup partitioning on the averaged states' features associated with the first Δ problems across historical trajectories, where $\Delta \in [1, 11]$ excluding the final problem. Then we augment the same amount of samples for each subgroup K, *i.e.*, $|\widehat{\mathcal{T}}_K| = |\mathcal{T}_K| = |K|$ and perform OPE with RRS. We observe that in all 55 conditions except the five (*i.e.*, WIS AugRRS $\Delta = 4, 11$, PDIS AugRRS $\Delta = 8$, and FQE AugRRS $\Delta = 7, 8$), all OPE AugRRS still select the sub-optimal policy. Figure 4 presents the mean absolute error (MAE) of the OPE AugRRS methods over the four target policies. It shows the trend of improved MAE over the number of problems for most methods. Those indicate that more information over a longer trajectory does have some positive effects on OPE AugRRS estimation, but their policy selection is hard to be improved and stabilized. More students-centric and robust OPE methods are needed for ITS planning.

Conclusion, Limitation, & Social Impact

In this work, we formulate ODPS problem to tackle the challenges when selecting RL-induced policy for ITS. We proposed EDUPLANNER to provide tailored policies for different subgroups, resulting in an unbiased off-policy evaluation estimator. EDUPLANNER was extensively evaluated on an ITS that had been employed over 8 years in universities. EDUPLANNER achieved significant improvement in the performance of students in terms of increased normalized learning gain, especially for low-performers. Note that our proposed method doesn't depend on specific RL algorithms for policy induction and the improvements resulting from the proposed estimator can be isolated, thus our work can be potentially generalizable to different ITSs and broader human-involved applications, such as healthcare, FinTech, etc.

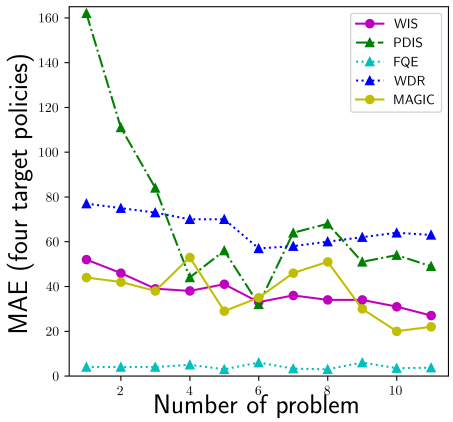

Figure 4: Mean absolute error (MAE) of OPE AugRRS with subgroup partitioning over problems in historical data.

In the proposed method, we assume that initial states are informative enough to guide subgroup partitioning, while there may exist some extreme conditions such as a cold start in a different human-involved system (Nie, Brunskill, and Piech 2021). An intuitive solution is deploying the expert policy to jump-start, or the optimal policy selected by OPE where subgroup partitioning can be conducted over longer historical trajectories to capture different human behaviors. The exhibited technical examples used in each step could be substituted by advanced techniques in the future and allows build-on-top work. In future work, we plan to investigate our framework in other classrooms or different domains, such as healthcare, to explore whether it consistently supports improvement on students' performance and various human-involved systems.

All educational data was obtained anonymously through an exempt IRB-approved protocol and were scored using established rubrics. No demographic data or class grades were collected. All data were shared within the research group under IRB, and were de-identified and automatically processed for grading. This research seeks to remove societal harms that come from lower engagement and retention of students who need more personalized interventions in ITS.

Acknowledgments

This research was supported by the NSF Grants: Integrated Data-driven Technologies for Individualized Instruction in STEM Learning Environments (1726550), CAREER: Improving Adaptive Decision Making in Interactive Learning Environments (1651909), and Generalizing Data-Driven

Technologies to Improve Individualized STEM Instruction by Intelligent Tutors (2013502). We would also like to thank the anonymous reviewers for insightful comments that leads to improved paper presentations.

References

Abdelshiheed, M.; Hostetter, J. W.; Barnes, T.; and Chi, M. 2023. Bridging declarative, procedural, and conditional metacognitive knowledge gap using deep reinforcement learning.

Abdelshiheed, M.; Zhou, G.; Maniktala, M.; Barnes, T.; and Chi, M. 2020. Metacognition and Motivation: The Role of Time-Awareness in Preparation for Future Learning.

Castro-Wunsch, K.; Ahadi, A.; and Petersen, A. 2017. Evaluating neural networks as a method for identifying students in need of assistance. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*, 111–116.

Chandak, Y.; Shankar, S.; Bastian, N.; da Silva, B.; Brunskill, E.; and Thomas, P. S. 2022. Off-Policy Evaluation for Action-Dependent Non-Stationary Environments. *Advances in Neural Information Processing Systems*, 35: 9217–9232.

Chi, M.; VanLehn, K.; Litman, D.; and Jordan, P. 2011. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1): 137–180.

Doroudi, S.; Thomas, P. S.; and Brunskill, E. 2017. Importance Sampling for Fair Policy Selection. *Grantee Submission*.

Gao, G.; Gao, Q.; Yang, X.; Ju, S.; Pajic, M.; and Chi, M. 2024. On Trajectory Augmentations for Off-Policy Evaluation. In *The Twelfth International Conference on Learning Representations*.

Gao, G.; Ju, S.; Ausin, M. S.; and Chi, M. 2023a. HOPE: Human-Centric Off-Policy Evaluation for E-Learning and Healthcare. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 1504–1513.

Gao, G.; Marwan, S.; and Price, T. W. 2021. Early performance prediction using interpretable patterns in programming process data. In *Proceedings of the 52nd ACM technical symposium on computer science education*, 342–348.

Gao, Q.; Gao, G.; Chi, M.; and Pajic, M. 2022. Variational Latent Branching Model for Off-Policy Evaluation. In *The Eleventh International Conference on Learning Representations*.

Gao, Q.; Gao, G.; Dong, J.; Tarokh, V.; Chi, M.; and Pajic, M. 2023b. Off-Policy Evaluation for Human Feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gao, Q.; Schmidt, S. L.; Chowdhury, A.; Feng, G.; Peters, J. J.; Genty, K.; Grill, W. M.; Turner, D. A.; and Pajic, M. 2023c. Offline Learning of Closed-Loop Deep Brain Stimulation Controllers for Parkinson Disease Treatment. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, 44–55.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*.

Hallac, D.; Vare, S.; Boyd, S.; and Leskovec, J. 2017. Toeplitz inverse covariance-based clustering of multivariate time series data. In *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 215–223.

Jiang, N.; and Li, L. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 652–661. PMLR.

Keramati, R.; Gottesman, O.; Celi, L. A.; Doshi-Velez, F.; and Brunskill, E. 2022. Identification of Subgroups With Similar Benefits in Off-Policy Policy Evaluation. In *Conference on Health, Inference, and Learning*, 397–410. PMLR.

Krishnan, S.; Haas, D.; Franklin, M. J.; and Wu, E. 2016. Towards reliable interactive data cleaning: A user survey and recommendations. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, 1–5.

Kumar, A.; Singh, A.; Tian, S.; Finn, C.; and Levine, S. 2022. A Workflow for Offline Model-Free Robotic Reinforcement Learning. In *Conference on Robot Learning*, 417–428. PMLR.

Le, H.; Voloshin, C.; and Yue, Y. 2019. Batch policy learning under constraints. In *International Conference on Machine Learning*, 3703–3712. PMLR.

Lee, A. X.; Nagabandi, A.; Abbeel, P.; and Levine, S. 2020. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in Neural Information Processing Systems*, 33: 741–752.

Lee, J.; Tucker, G.; Nachum, O.; and Dai, B. 2022. Model selection in batch policy optimization. In *International Conference on Machine Learning*, 12542–12569. PMLR.

Liu, E.; Stephan, M.; Nie, A.; Piech, C.; Brunskill, E.; and Finn, C. 2022. Giving Feedback on Interactive Student Programs with Meta-Exploration. *Advances in Neural Information Processing Systems*, 35: 36282–36294.

Mandel, T.; Liu, Y.-E.; Levine, S.; Brunskill, E.; and Popovic, Z. 2014. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077.

Mao, Y. 2019. One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In *In: Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

Marwan, S.; Dombe, A.; and Price, T. W. 2020. Unproductive help-seeking in programming: What it is and how to address it. In *Proceedings of the 2020 ACM conference on innovation and technology in computer science education*, 54–60.

Miyaguchi, K. 2022. A Theoretical Framework of Almost Hyperparameter-free Hyperparameter Selection Methods for Offline Policy Evaluation. *arXiv e-prints*, arXiv–2201.

Nachum, O.; Chow, Y.; Dai, B.; and Li, L. 2019. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32.

Nie, A.; Brunskill, E.; and Piech, C. 2021. Play to grade: testing coding games as classifying Markov decision process. *Advances in Neural Information Processing Systems*, 34: 1506–1518.

Nie, A.; Flet-Berliac, Y.; Jordan, D.; Steenbergen, W.; and Brunskill, E. 2022. Data-Efficient Pipeline for Offline Reinforcement Learning with Limited Data. *Advances in Neural Information Processing Systems*, 35: 14810–14823.

Paduraru, C. 2013. Off-policy evaluation in Markov decision processes.

Precup, D. 2000. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, 80.

Rowe, J.; Mott, B.; and Lester, J. 2014. Optimizing player experience in interactive narrative planning: a modular reinforcement learning approach. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 10, 160–166.

Rybkin, O.; Zhu, C.; Nagabandi, A.; Daniilidis, K.; Mordatch, I.; and Levine, S. 2021. Model-based reinforcement learning via latent-space collocation. In *International Conference on Machine Learning*, 9190–9201. PMLR.

Sanz Ausin, M.; Maniktala, M.; Barnes, T.; and Chi, M. 2020. Exploring the impact of simple explanations and agency on batch deep reinforcement learning induced pedagogical policies. In *International Conference on Artificial Intelligence in Education*, 472–485. Springer.

Schwonke, R.; Renkl, A.; Krieg, C.; Wittwer, J.; Aleven, V.; and Salden, R. 2009. The worked-example effect: Not an artefact of lousy control conditions. *Computers in human behavior*, 25(2): 258–266.

Shen, S.; and Chi, M. 2016. Reinforcement learning: the sooner the better, or the later the better? In *Proceedings of the 2016 conference on user modeling adaptation and personalization*, 37–44.

Sinclair, R. R.; Martin, J. E.; and Michel, R. P. 1999. Full-time and part-time subgroup differences in job attitudes and demographic characteristics. *Journal of Vocational Behavior*, 55(3): 337–357.

Sweller, J.; and Cooper, G. A. 1985. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and instruction*, 2(1): 59–89.

Thomas, P.; and Brunskill, E. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2139–2148. PMLR.

VanLehn, K. 2006. The Behavior of Tutoring Systems. *International Journal Artificial Intelligence in Education*, 16(3): 227–265.

Voloshin, C.; Jiang, N.; and Yue, Y. 2021. Minimax model learning. In *International Conference on Artificial Intelligence and Statistics*, 1612–1620. PMLR.

Voloshin, C.; Le, H. M.; Jiang, N.; and Yue, Y. 2021. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Wang, P.; Rowe, J.; Min, W.; Mott, B.; and Lester, J. 2017. Interactive narrative personalization with deep reinforcement learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Yang, M.; Dai, B.; Nachum, O.; Tucker, G.; and Schuurmans, D. 2022. Offline policy selection under uncertainty. In *International Conference on Artificial Intelligence and Statistics*, 4376–4396. PMLR.

Yang, M.; Nachum, O.; Dai, B.; Li, L.; and Schuurmans, D. 2020a. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33: 6551–6561.

Yang, X.; Zhang, Y.; and Chi, M. 2021. Multi-series time-aware sequence partitioning for disease progression modeling. In *IJCAI*.

Yang, X.; Zhou, G.; Taub, M.; Azevedo, R.; and Chi, M. 2020b. Student Subtyping via EM-Inverse Reinforcement Learning. *International Educational Data Mining Society*.

Zhang, M. R.; Paine, T. L.; Nachum, O.; Paduraru, C.; Tucker, G.; Wang, Z.; and Norouzi, M. 2021. Autoregressive dynamics models for offline policy evaluation and optimization. *arXiv preprint arXiv:2104.13877*.

Zhang, S.; Liu, B.; and Whiteson, S. 2020. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, 11194–11203. PMLR.

Zhong, R.; Zhang, D.; Schäfer, L.; Albrecht, S.; and Hanna, J. 2022. Robust On-Policy Sampling for Data-Efficient Policy Evaluation in Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35: 37376–37388.

Zhou, G.; Azizsoltani, H.; Ausin, M. S.; Barnes, T.; and Chi, M. 2022. Leveraging granularity: Hierarchical reinforcement learning for pedagogical policy induction. *International journal of artificial intelligence in education*, 32(2): 454–500.