Importance Sampling to Identify Empirically Valid Policies and their Critical Decisions

Song Ju, Shitian Shen, Hamoon Azizsoltani, Tiffany Barnes, Min Chi
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
{sju2, sshen, hazizso, tmbarnes, mchi}@ncsu.edu

ABSTRACT

In this work, we investigated *off-policy* policy evaluation (OPE) metrics to evaluate Reinforcement Learning (RL) induced policies and to identify *critical decisions* in the context of Intelligent Tutoring Systems (ITSs). We explore the use of three common Importance Sampling based OPE metrics in *two* deployment settings to evaluate *four* RL-induced policies for a logic ITS. The two deployment settings explore the impact of using original or normalized rewards, and the impact of transforming deterministic to stochastic policies. Our results show that Per Decision Importance Sampling (PDIS), using soft max and original rewards, is the best metric, and the only metric that reached 100% alignment between the *theoretical* and *empirical* classroom evaluation results. Furthermore, we used PDIS to identify what we call *critical decisions* in RL-induced policies, where the policies successfully identify large differences between decisions. We found that the students who received more critical decisions significantly outperformed those who received less; more importantly, this result only holds on the policy that was identified to be effective using PDIS, not on ineffective ones.

Keywords

Reinforcement Learning, Off-policy Policy Evaluation, Importance Sampling

1. INTRODUCTION

Intelligent Tutoring Systems (ITSs) are a type of highly interactive e-learning environment that facilitates learning by providing step-by-step support and contextualized feedback to individual students [12, 30]. These step-by-step behaviors can be viewed as a sequential decision process where at each step the system chooses an action (e.g. give a hint, show an example) from a set of options, in which *pedagogical strategies* are policies that are used to decide what action to take next in the face of alternatives. Reinforcement Learning (RL) offers one of the most promising approaches to data-driven decision-making applications and RL algorithms are designed to induce effective policies that determine the best action for an agent to take in any given situation so as to maximize some predefined cumulative reward. A number of researchers have studied the application of existing RL algorithms to improve the effectiveness of ITSs [3, 26, 21, 4, 28, 8, 9, 34]. While promising, such RL work faces at least two major challenges discussed below.

One challenge is a lack of reliable yet robust evaluation metrics for RL policy evaluation. Generally speaking, there are two major categories of RL: online and offline. In the former category, the agent learns while interacting with the environment; in the latter case, the agent learns the policy from pre-collected data. Online RL algorithms are generally appropriate for domains where interacting with simulations and actual environments is computationally cheap and feasible. On the other hand, for domains such as e-learning, building accurate simulations or simulated students is especially challenging because human learning is a rather complex, poorly understood process. Moreover, learning policies while interacting with students may not be feasible, and more importantly, may not be ethical. Therefore, to improve student learning, much prior work applied offline RL approaches to induce effective pedagogical strategies. This is done by first collecting a training corpus and the success of offline RL is often heavily dependent on the quality of the training corpus. One common convention is to collect an exploratory corpus by training a group of students on an ITS that makes random yet *reasonable* decisions and then apply RL to induce pedagogical policies from that training corpus. Empirical study is then conducted from a new group of human subjects interacting with different versions of the system. The only difference among the system versions is the policy employed by the ITS. The students' performance is then statistically compared. Due to cost limitations, typically, only the *best* RL-induced policy is deployed and compared against some baseline policies. On the other hand, we often have a large number of RL algorithms (and associated hyperparameter settings), and it is unclear which will work *best* in our setting. In these high-stake situations, one needs confidence in the RL-induced policy before risking deployment. Therefore, we need to develop reliable yet robust evaluation metrics to evaluate these RL-induced policies without collecting new data before being tested in the real world. This type of evaluation is called *off-policy evaluation* (OPE) because the policy used to collect the training data, also referred to as the *behavior* policy, is different from the RL-induced policy, referred to as the *target* policy to be evaluated. To find reliable yet robust OPE metrics, we explored three Importance Sampling based off-policy evaluation metrics.

The second RL challenge is a lack of interpretability of the RL-induced policies. Compared with the amount of research

done on applying RL to induce policies, relatively little work has been done to analyze, interpret, or explain RL-induced policies. While traditional hypothesis-driven, cause-and-effect approaches offer clear conceptual and causal insights that can be evaluated and interpreted, RL-induced policies are often large, cumbersome, and difficult to understand. The space of possible policies is exponential in the number of domain features. It is therefore difficult to draw general conclusions from them to advance our understanding of the domain. This raises a major open question: *How can we identify the critical system interactive decisions that are linked to student learning?* In this work, we tried to identify key decisions by taking advantage of the reliable OPE metrics we discovered and the properties of the policies we induced.

2. MOTIVATION

Just like the fact that assessment sits at the epicenter of educational research [2], policy evaluation is indeed the central concern among the many stakeholders in applying offline RL to ITSs. As educational assessment should reflect and reinforce the educational goals that society deems valuable, our policy evaluation metrics should reflect the effectiveness of the induced policies. While various RL approaches such as policy iteration and policy search have shown great promise, existing RL approaches tend to perform poorly when they are actually implemented and evaluated in the real world.

In a series of prior studies on a logic ITS, RL and Markov Decision Processes (MDPs) were applied to induce *four* different pedagogical policies, named MDP1-MDP4 respectively, on one type of tutorial decision: whether to provide students with a *Worked Example* (WE) or to ask them to engage in *Problem Solving* (PS). In WEs, the tutor presents an expert solution to a problem step by step, while in PSs, students are required to complete the problem with the tutor's support. When inducing each of four policies, we explored different feature selection methods and used *Expected Cumulative Reward (ECR)* to evaluate the RL-induced policies. ECR of a policy is calculated by average over the value function of initial states and generally speaking, the higher the ECR value of a policy, the better the policy is supposed to perform.

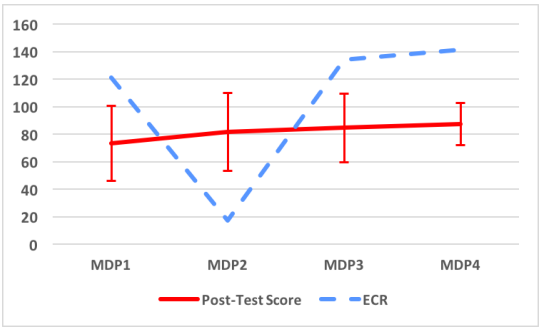

Figure 1: Post-test Score vs. ECR, showing no seeming direct relationship.

Figure 1 shows the ECRs (blue dashed line) of the four RL-induced policies, MDP1-MDP4 (x-axis) and the *empirical* results of student learning performance (the solid red line) of the corresponding policies. Here the learning performance is measured by an in-class post-test after students were trained on the tutor with corresponding policies (the mean and standard errors of post-test scores are shown with a red solid line). Figure 1 shows that our *theoretical* evaluation (ECR) does not match the *empirical* results (post-test) evaluation in that there is no clear relationship between the ECRs' blue line and the corresponding post-test in red line across the four policies. This result shows that ECR is not a reliable OPE metric for evaluating RL-induced policy in ITSs. Indeed, Mandel et al. [15] pointed out that ECR tends to be biased, statistically inconsistent and thus it may not be the appropriate OPE metric in high stakes domains. In recent years, many state-of-the-art OPE metrics have been proposed and many of them are based on Importance Sampling.

Importance Sampling (IS) is a classic OPE method for evaluating a *target* policy on existing data obtained from an alternate *behavior* policy and thus can be handily applied to the task of evaluating the effectiveness of an offline RL-induced policy using pre-existing historical training datasets. Many IS-based OPE metrics are proposed and explored and it was shown that they gain significant performance in simulation environments like Grid World or Bandit [29, 6]. Among them, three IS-based OPE metrics, the original IS, Weighted IS (WIS), and Per-Decision IS (PDIS), are the most widely used. However, real-world human-agent interactive applications such as ITSs are much more complicated due to 1) individual differences, noise, and randomness during the interaction processes, 2) the large state space that can impact student learning, and 3) long trajectories due to the nature of the learning process.

In this work, we investigated the three IS-based offline OPE metrics on MDP1-MDP4 to investigate whether the three IS-based evaluation metrics are indeed effective OPE metrics for evaluating the four RL-induced policies mentioned beforehand. We believe an OPE is effective *if and only if* the *theoretical* results from the OPE evaluations are *completely aligned* with the *empirical* results from the classroom studies. Therefore, we explored different deployment settings for the IS-based metrics from two aspects: one is the transformation function used to convert the RL-induced deterministic policy to a stochastic policy used in IS-based metrics and the other is reward functions: the original reward function vs. the normalized reward function; the latter is supposed to reduce the variance. Our results showed that the theoretical and empirical evaluation results are aligned more or less for different deployment settings using different IS-based metrics. Only when using a soft-max transformation function and original reward function, the theoretical results of PDIS can reach 100% agreement with the empirical results. Based on results from the OPE metrics, we further explored using the properties of the RL-induced policy to identify *critical decisions* and our results showed that the critical decisions can be identified by using the theoretically "effective" policies that were identified by using PDIS with soft-max transformation and the original reward function.

In summary, we make the following contributions:

- We directly compared three IS-based policy evaluation metrics against the empirical results from real classroom studies across four different RL-induced policies. Our results showed that PDIS is the best one and its results can align with empirical results.
- As far as we know, this is the first study to compare different deployment settings (original/normalized rewards

or deterministic/stochastic policy transformation) on IS-based policy evaluation metrics. Our results showed that settings have a direct impact on the effectiveness of the evaluation metrics. Only PDIS with soft-max transformation and the original reward function agreed 100% with the empirical results.

- We investigated using information from the RL-induced policies to identify *critical decisions* to shed some light on the induced policies. As far as we know, this is the first attempt to differentiate the critical decisions from trivial ones.

3. RELATED WORK

3.1 Empirical Studies Applying RL to ITSs

In recent years, a number of researchers have applied RL to induce effective pedagogical policies for ITSs [3, 5, 13, 23]. Some previous work treated the user-system interactions as fully observable processes by applying Markov Decision Processes (MDPs) [14, 27, 1] while others utilized partially observable MDPs (POMDPs) [24, 32, 33, 15, 4], and more recently, deep RL Frameworks [31, 18]. Most of the previous work, including this work, took the offline RL approach in that it followed three general steps: 1) to collect an *exploratory* corpus by training a relatively large group of real and/or simulated students on an ITS that makes some random yet reasonable and rational tutorial decisions; 2) to apply RL to induce pedagogical policy directly from the historical exploratory corpus; and 3) to implement the RL-induced pedagogical policy back to the ITS and evaluate its effectiveness using simulations and/or in real classroom settings to investigate whether RL fulfills its promise. Oftentimes, when implementing and evaluating the RL-induced policies in real world, they tend to perform poorly compared to their theoretical performance.

Iglesias et al. [10] applied Q-learning to generate a policy in an ITS that teaches students database design. Their goal was to provide students with direct navigational support through the system's content. In the training phase, they used simulated students to induce an RL policy and online-evaluated the induced policy based upon three self-defined measures. In the test phase, they evaluated the induced policy with real students. The results showed that while students using the induced policies had more effective usage behaviors than their non-policy peers, there was no significant difference in student learning performance. Chi et al. [17] applied an offline RL approach to induce pedagogical policies for improving the effectiveness of an ITS that teaches students college physics. They used ECR to evaluate the induced policy in the training phase. However, when they applied the induced policy to real students, the induced policy did not outperform the random policy in empirical evaluation. Similarly, Shen et al. [26] explored immediate and delayed rewards based on learning gain and implemented an offline, MDP framework on a rule-based ITS for deductive logic. They selected the policy with the highest ECR and deployed it in an ITS for real students. The empirical results showed that the RL policies were no more effective than the random baseline policy. In addition, Rowe et al. [22] investigated an MDP framework for tutorial planning in a game-based learning system. They used ECR to theoretically evaluate induced policies but in an empirical study with real students, they found that students in the induced planner condition had significantly different behavior patterns from the control group but no significant difference was found between the two groups on the post-test [23].

In short, prior work on applying offline RL in ITSs primarily explored ECR as the OPE metric and one common phenomenon is that the theoretical evaluation results do not align with the empirical results from real students.

3.2 OPE Metrics

OPE is used to evaluate the performance of a target policy given historical data generated by an alternative behavior policy. A good OPE metric is especially important for real-world applications where the deployment of a bad or inefficient policy can be costly [19]. ECR is one of the most widely used OPE metrics, which is designed especially for the MDP framework. Tetreault et al. [11] estimated the reliability of ECRs by repeated sampling to estimate *confidence intervals* for ECRs. In simulation studies, they showed the policy induced by the confidence interval of ECR performed more reliably than the baseline policies but this phenomenon did not hold when evaluating the RL-induced policies in ITSs for empirical studies [17].

Importance Sampling (IS) [7] is a widely used OPE metric, which considers the mathematical characteristics of the decision making process and can be applied to any MDP, POMDP, or DeepRL framework. Precup [20] proposed four IS-based OPE metrics: IS, weighted importance sampling (WIS), per-decision importance sampling (PDIS), and weighted per-decision importance sampling (WPDIS). They used the IS-based estimator as the policy evaluation for Q-learning and then compared the effectiveness of estimators on a series of 100 randomly-constructed MDPs based on the mean square error (MSE). Their results showed that IS made the Q-learning process converge slowly and caused high variance, and WIS performed better than IS; but PDIS performed inconsistently and WPDIS performed the worst. Similarly, Thomas [29] compared the performance of several IS estimators using mean squared error in a grid-world simulation, showing PDIS outperformed all others.

In summary, previous work has explored the effectiveness of IS and its variants in simulation studies, which motivated the work reported here. Different from previous work, we mainly focus on comparing the theoretical evaluation with the empirical evaluation in order to determine whether IS-based methods are indeed reliable and robust for ITSs.

4. MARKOV DECISION PROCESS & RL

Some of the prior work on applying RL to induce pedagogical policies used Markov Decision Processes (MDP) frameworks. An MDP can be seen as a 4-tuple $\langle S,A,T,R\rangle$, where S denotes the observable state space, which is defined by a set of features that represent the interactive learning environment and A denotes the space of possible actions for the agent to execute. The reward function R represents the immediate or delayed feedback from the environment with respect to the agent's action(s); $r(s,a,s')$ denotes the expected reward of transiting from state s to state s' by taking action a. Once $\langle S,A,R\rangle$ is defined, T represents the transition probability where $P(s,a,s')=Pr(s'|s,a)$ is the probability of transitioning from state s to state s' by taking action a and it can be easily estimated from the training corpus. The optimal policy π for an MDP can be generated via dynamic programming approaches, such as Value Iteration. This algorithm operates by finding the optimal value for each

state $V^*(s)$, which is the expected discounted reward that the agent will gain if it starts in s and follows the optimal policy to the goal. Generally speaking, $V^*(s)$ can be obtained by the optimal value function for each state-action pair $Q^*(s,a)$ which is defined as the expected discounted reward the agent will gain if it takes an action a in a state s and follows the optimal policy to the end. The optimal value function $Q^*(s,a)$ can be obtained by iteratively updating $Q(s,a)$ via equation 1 until convergence:

$$Q(s,a) := \sum_{s'} p(s,a,s') \Big[r(s,a,s') + \gamma \max_{a'} Q(s',a') \Big] \quad (1)$$

where $0 \leq \gamma \leq 1$ is a discount factor. When the process converges, the optimal policy π^* can be induced corresponding to the optimal Q-value function $Q^*(s,a)$, represented as:

$$\pi^*(s) = \operatorname{argmax}_a Q^*(s,a) \quad (2)$$

Where π^* is the *deterministic* policy that maps a given state into an action. In the context of an ITS, this induced policy represents the pedagogical strategy by specifying tutorial actions using the current state.

5. THREE OPE METRICS & TWO SETTINGS

The following terms will be used throughout this paper.

- $H = s_1 \xrightarrow{a_1,r_1} s_2 \xrightarrow{a_2,r_2} s_3 \xrightarrow{a_3,r_3} \cdots s_L$; denotes *one* student-system interaction trajectory and H^L denotes a trajectory with length L.
- $G(H^L) = \sum_{t=1}^{L} \gamma^{t-1} r_t$ is the discounted return of the trajectory H^L, which generally reflects how good the trajectory is supposed to be.
- $D = \{H_1, H_2, H_3, ..., H_n\}$ denotes the historical dataset containing n student-system interaction trajectories.
- π_b denotes the *behavior* policy carried out for collecting the historical data D.
- π_e denotes the target policy to be *evaluated*.
- $\rho(\pi_e)$ represents the estimated performance of π_e.

5.1 Three IS-based OPE Metrics

Importance Sampling (IS) is an approximation method that allows the estimation of the expectation of a distribution, p, from samples generated from a different distribution, q. Suppose that we have sample space x and random variable $f(x)$ which is a measurable function from sample space x to another measurable space. We want to estimate the expectation of $f(x)$ over a strictly positive probability density function $p(x)$. Suppose also that we cannot directly sample from distribution $p(x)$, but we can draw Independent and Identically Distributed (IID) samples from probability density function $q(x)$ and evaluate $f(x)$ for these samples. The expectation of $f(x)$ over probability density function $p(x)$ can be calculated as:

$$\begin{aligned} E_p[f(x)] &= \int f(x)p(x)dx & (3)\\ &= \int f(x)\frac{p(x)}{q(x)}q(x)dx & (4)\\ &= E_q[f(x)\frac{p(x)}{q(x)}] & (5) \end{aligned}$$

where p is known as the target distribution, q is the sampling distribution, $E_p[f(x)]$ is the expectation of $f(x)$ under p, $p(x)/q(x)$ is the likelihood ratio weight and $E_q[f(x)p(x)/q(x)]$ is the expectation of $f(x)p(x)/q(x)$ under q. We can then approximate the expectation of $f(x)$ over probability density function $p(x)$ using the samples drawn from probability density function $q(x)$. In the context of OPE, the target distribution, p, is a probability event whose density function is determined by the target policy, and the sampling distribution, $p(x)$, is a probability event whose density function is determined by the behaviour policy.

Following the general IS technique, we approximated the expected reward of the target policy using the relative probability of the target and behavior policies. Because of the nature of the underlying MDP, samples in RL are sequential. Therefore, we assumed that each trajectory is a sequence of events whose probability density function is determined by its corresponding policy. Assuming the independence among trajectories and following the multiplication rule of independent events, the probability of occurrence of a trajectory, H^L, under policy π is

$$\mathcal{P}_\pi(H^L) = \mathcal{P}_1(s_1) \prod_{t=1}^{L} \pi(a_t|s_t) \mathcal{P}_T(s_{t+1}|s_t,a_t) \quad (6)$$

where $\mathcal{P}_\pi(H^L)$ is the probability of occurrence of the trajectory H^L following a policy π, $\mathcal{P}_1(s_1)$ is the probability of occurrence of the state s_1 in the beginning of the trajectory and $\mathcal{P}_T$ is the state transition probability function.

Similar to the original IS technique, the proportional probability of each trajectory occurring under the target policy, π_e, and behavior policy, π_b, is used as the likelihood ratio weight for that trajectory. Thus, following the importance sampling technique, the importance sampling discounted return is defined as:

$$\begin{aligned} IS(\pi_e|H^L,\pi_b) &= \frac{\mathcal{P}_{\pi_e}(H^L)}{\mathcal{P}_{\pi_b}(H^L)} \cdot G(H^L) & (7)\\ &= \frac{\prod_{t=1}^{L} \pi_e(a_t|s_t)}{\prod_{t=1}^{L} \pi_b(a_t|s_t)} \cdot G(H^L) & (8) \end{aligned}$$

Substituting $G(H^L)$ with the discounted return and we have:

$$IS(\pi_e|H^L,\pi_b) = \Big(\prod_{t=1}^{L} \frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}\Big)\Big(\sum_{t=1}^{L} \gamma^{t-1} r_t\Big) \quad (9)$$

After having the individual IS estimator for each trajectory, we can calculate the expected reward of the dataset D by averaging

the individual IS estimators for each trajectory as

$$IS(\pi_e|D)=\frac{1}{n_D}\sum_{i=1}^{n_D}\left[(\prod_{t=1}^{L^i}\frac{\pi_e(a_t^i|s_t^i)}{\pi_b(a_t^i|s_t^i)})(\sum_{t=1}^{L^i}\gamma^{t-1}r_t^i)\right] \quad (10)$$

where s_t^i, a_t^i, and r_t^i refer to the ith trajectory at time t and n_D is the number of trajectories in D.

Weighted Importance Sampling (WIS) is a variant of the IS estimator, a biased but consistent estimator which has lower variance than IS. It normalizes the IS in order to produce a lower variance. At first, it calculates the weight W_D for the dataset as the summation of the likelihood ratios for each trajectory as shown in equation 11. Then, it normalizes the IS estimator as shown in equation 12. Finally, the WIS is simply the weighted average of the estimated reward for each sequence in the dataset D, as shown in equation 13.

$$W_D=\sum_{i=1}^{n_D}(\prod_{t=1}^{L^i}\frac{\pi_e(a_t^i|s_t^i)}{\pi_b(a_t^i|s_t^i)}) \quad (11)$$

$$WIS(\pi_e|H^L,\pi_b)=\frac{IS(\pi_e|H^L,\pi_b)}{W_D} \quad (12)$$

$$WIS(\pi_e|D)=\frac{\sum_{i=1}^{n_D}\left[(\prod_{t=1}^{L^i}\frac{\pi_e(a_t^i|^i s_t)}{\pi_b(a_t^i|s_t^i)})(\sum_{t=1}^{L^i}\gamma^{t-1}r_t^i)\right]}{\sum_{i=1}^{n_D}(\prod_{t=1}^{L^i}\frac{\pi_e(a_t^i|s_t^i)}{\pi_b(a_t^i|s_t^i)})} \quad (13)$$

Per-Decision Importance Sampling (PDIS) is also a variant of IS. Like IS, it is also unbiased and consistent. IS has very high variance because for each reward, r_t, it uses the likelihood ratio of the entire trajectory. However, the reward at step t should only depend on the previous steps. The variance can be reduced by using the likelihood ratio of the trajectory before step t for the reward r_t. It means that the importance weight for a reward at step t is $\prod_{j=1}^{t}\frac{\pi_e(a_j|s_j)}{\pi_b(a_j|s_j)}$. Therefore, the individual PDIS estimator for a trajectory is given in equation 14 and the PDIS for the whole historical dataset D is given in equation 15.

$$PDIS(\pi_e|H^L,\pi_b)=\sum_{t=1}^{L}\gamma^{t-1}(\prod_{j=1}^{t}\frac{\pi_e(a_j|s_j)}{\pi_b(a_j|s_j)})r_t \quad (14)$$

$$PDIS(\pi_e|D)=\frac{1}{n_D}\sum_{i=1}^{n_D}\left[\sum_{t=1}^{L^i}\gamma^{t-1}(\prod_{j=1}^{t}\frac{\pi_e(a_j^i|s_j^i)}{\pi_b(a_j^i|s_j^i)})r_t^i\right] \quad (15)$$

5.2 Two Different Settings

To evaluate the three IS-based metrics, we explored different deployment settings from two aspects: one is the transformation function to convert the RL-induced deterministic policy to a stochastic policy used in IS-based metrics and the other is reward functions: the original reward function vs. the normalized reward.

5.2.1 *Three Transformation Functions:*

As described above, IS-based metrics require the policy to be stochastic but our MDP induced policies are deterministic, i.e. given the current state s, the agent should take the deterministic optimal action a^* following the optimal policy π^*. To transform the deterministic policy into a stochastic policy, we explore three types of transformation functions: Hard-code, Q-proportion, and Soft-max. The basic idea behind them is: for any given state s, the assigned stochastic probability for an action a should reflect its value of $Q(s,a)$.

1. **Hard-code Transformation**

$$\pi(a|s)=\begin{cases}1-\varepsilon & \text{optimal action}\\ \varepsilon & \text{otherwise}\end{cases} \quad (16)$$

In eqn 16, ε is a fixed small probability like $\varepsilon = 0.001$ which is assigned to actions with smaller Q-value while $1-\varepsilon$ is assigned to the action with the largest Q-value.

2. **Q-proportion Transformation**

$$\pi(a|s)=\frac{Q(s,a)}{\sum_{a'\in A}Q(s,a')} \quad (17)$$

As shown in eqn 17, the probability to take action a given state s is the proportion of a's Q-value among all possible actions a' in the state s. Thus, the action which has the highest Q-value is guaranteed to have the highest probability. In practice, because some of the Q-values are smaller than 0, we add a constant to Q-values in the same state so that $Q(s,a)\geq 1$.

3. **Soft-max Transformation**

$$\pi(a|s)=\frac{e^{\theta\cdot Q(s,a)}}{\sum_{a'\in A}e^{\theta\cdot Q(s,a')}} \quad (18)$$

Soft-max is a classical function used to calculate the probability distribution of one event over n possible events. In our application, given state s, the soft-max function will calculate the probability of action a over all possible actions using equation 18. The main advantage of soft-max is the output probability range is 0 to 1 and the sum of all the probabilities will be 1 and it can handle negative Q-values. θ is a weight parameter to the Q-value.

5.2.2 *Two Types of Reward Functions:*

The effectiveness of RL-induced policies is very sensitive to the reward functions. In our application, the range of the reward function is very large [−200,200], which may cause large variances for IS, especially when a trajectory is long. One effective way to reduce this variance is to use normalized rewards. Therefore, both original rewards and normalized rewards are considered. More specifically, the normalized reward z is defined as: $z=\frac{x-min(x)}{max(x)-min(x)}$ where min and max are the minimum and maximum values of original reward function x and $z\in[0,1]$.

6. ITS & FOUR MDP POLICIES

6.1 Our Logic Tutor

Figure 2 shows an interface of the logic tutor, which is a data-driven ITS used in the undergraduate Discrete Mathematics (DM) course at a large university. It [16] provides students with a graph-based representation of logic proofs which allows students to solve problems by applying rules to derive new statements, represented as nodes. The system automatically verifies proofs and provides immediate feedback on logical errors. Every problem in the tutor can be presented in the form of either

Figure 2: Tutor Problem Solving Interface

WE or PS. By focusing on the pedagogical decisions of WE and PS, the tutor allows us to strictly control the content to be *equivalent* for all students.

6.2 Four MDP Policies and Empirical Study

Four MDP policies, MDP_1 - MDP_4 were induced from an exploratory pre-collected dataset following different feature selection procedures. The detailed descriptions on our policy induction process are described in [26, 25] and must be omitted here because of page limits. The effectiveness of each MDP policy was empirically evaluated against the Random policy in strictly controlled studies during three consecutive semesters. In each strictly controlled study, students were randomly assigned to two conditions: MDP policy, or a Random baseline policy which makes random yet reasonable decisions because both *PS* and *WE* are always considered to be *reasonable* educational interventions in our learning context. Moreover, all students went through the identical procedure on the tutor and the only difference was the pedagogical policy employed. After completing the tutor, students take a post-test which involved two proof questions in a midterm exam. They were 16 points each and graded by one TA using a rubric. Overall, no significant difference was found between our four RL-induced policies and the Random policy on students' *post-test scores* across all studies.

There are many possible explanations for our results. First, while the Random policy is generally a weak policy for many RL tasks, in our situation both of our action choices: WE vs. PS are considered reasonable and more importantly for each decision point, there is a 50% chance that the random policy would carry out the better of the two. Second, non-significant statistical results do not mean non-existence. Small sample size may play an important role in limiting the significance of statistical comparisons. A post hoc power analysis revealed that in order to be detected as significant at the 5% level with 80% power, MDP1 vs. Random needed a total sample of 1382 students; MDP2 vs. Random needed 1700 students; MDP3 vs. Random needed 212 students; MDP4 vs. Random needed 394 students. However, in each empirical study, our student sample sizes are much smaller, 59, 50, 57, and 84 respectively. And last but not least, it turned out that all four RL-induced policies were only partially carried out. All of the training problems in our tutor are organized into six strictly ordered levels and in each level students are required to complete 3–4 problems. In level 1, all participants receive the same set of PS problems and in the levels 2–6, our tutor has two hard-coded *action-based constraints* that are required by the class instructors: students must complete at least one PS and one WE and the last problem on each level must be PS. Therefore, over the entire training process, only $\sim$50% of actions are actually decided by the pedagogical policy and the rest are decided by hard-coded system rules.

In short, despite the fact that ECR showed that our four RL-induced policies should be more effective than the Random, empirical results showed otherwise because of various potential reasons. So we explored other OPE metrics to evaluate MDP_1 - MDP_4.

7. EXPERIMENT SETUP

Our dataset contains a total of 450 students' interaction logs involved in the strictly controlled studies mentioned above. The goals of our experiment were to: 1) investigate whether any of the three IS-based metrics can be used to align the theoretical and empirical results for our four MDP policies and 2) identify critical decisions that are linked to student learning.

7.1 Three IS-based Metrics Evaluation

We will describe how we determine whether or not the three IS-based metrics can align the theoretical results with the empirical results for MDP1-MDP4.

For a given RL-induced policy π, we first split all students into High vs. Low based on the actual carry-out percentages according to π. Since there are only two tutorial choices: WE vs. PS, there is a possibility that each actual decision that the tutor made would agree with the decision according to π. For the Random policy, for example, the probability is 50-50. In other words, we can measure each trajectory by the percentage of the tutorial decisions that agree with π. If π is indeed effective, we would expect that the more the tutorial decisions in a trajectory agree with π, the better the corresponding student performance would be. We thus treat all 450 students' interaction log data *equally* regardless of their original assigned conditions and for each student-ITS interaction log we can calculate a carry-out percentage for π using the formula: $percentage = \frac{N_{agree}}{N_{total}}$, where N_{total} is the total number of tutorial decisions in the trajectory and N_{agree} is the number of decisions that agree with π. Then, students are divided into High Carry-out (High) vs. Low Carry-out (Low) by a median split on carry-out percentages.

Then we *empirically evaluate* the effectiveness of π by checking whether there is a significant difference between the High and Low groups on their post-test scores. Similarly, we compare the High and Low groups' *theoretical* results. To do so, for each student-ITS interaction trajectory, we estimate its reward by exploring different combinations of the three IS-based OPE metrics with the three policy transformation functions and the two reward functions. More specifically, we treat all the trajectories as generated by Random policy regardless of their original behavior policy. So for each student, we have a total of 18 theoretical evaluations for a given π. If π is indeed effective and our OPE metric is reliable, we would expect that the more the tutorial decisions in a trajectory agree with π, the higher the corresponding theoretical rewards would be and vice versa.

Finally, for each of the 18 OPE metric settings, we conduct an alignment test between the theoretical and empirical results on π. This is done by comparing the empirically evaluated results and the theoretical rewards using the corresponding OPE metric. More specifically, they are considered to be **aligned** when:

1. Both empirical and theoretical results were not significant, that is $p >= 0.05$, or

2. Both results were significant, and the direction of the comparison was the same, that is $p < 0.05$, and the sign of the t values are both positive or both negative.

All the remaining cases are considered as not aligned. Thus, for each of 18 OPE metrics, we can test whether its theoretical results would align with the empirical results for π. Since we have four RL-induced policies, MDP1-MDP4, robust and reliable OPE metrics should align the two types of evaluation results across all four policies.

7.2 Critical Decision Identification

Next, we will explain how the critical interactive decisions are identified and empirically examined. Note that, there may be critical decisions over which the RL policies have no influence. Hence, we focus only on interactive decisions that are critical. For many RL algorithms, the fundamental approach to induce an optimal policy can be seen as recursively estimating the Q-values: $Q(s,a)$ for any given state-action pair until the Bellman equation is converged. More specifically, $Q(s,a)$ is defined as the expected discounted reward the agent will gain if it takes an action a in a state s and follows the corresponding optimal policy to the end. Thus, for a given state s, a large difference between the values of $Q(s, \text{“}PS\text{”})$ and $Q(s, \text{“}WE\text{”})$ indicates that it is more important for the ITS to follow the optimal decision in the state s. We, therefore, used the absolute difference between the Q-values for each state s to identify critical decisions. Our procedure can be divided into two steps:

Step 1: Identify Critical Decisions: Given an MDP policy, for each state, we calculated the absolute Q-value difference between the two actions (PS vs. WE) associated with it. Figure 3 shows the Q-value difference (y-axis) for each state (x-axis) sorted in descending order for MDP1-MDP4 policies respectively. It clearly shows that, across the four MDP policies, the Q-value differences for different states can vary greatly. We used the median Q-value difference to split the x-axis *states* into critical vs. non-critical states. The states with the larger Q-value differences were *critical states* and the rest were non-critical ones. For a given RL-induced policy, the *critical decisions* are defined as those in critical states where the actual carried-out tutorial action agreed with the corresponding policy.

Step 2: Evaluate Critical Decisions: For each of the four RL-induced policies, we counted the number of critical decisions that each student encountered during his/her training. Then for each policy, students were split into: More vs. Less groups, by a median split on the number of critical decisions experienced in the training process. A t-test was conducted on the post-test scores of More vs. Less groups to investigate whether the students with More critical decisions would indeed perform better than those with Less.

8. RESULTS

8.1 Three IS-based Metrics Evaluation

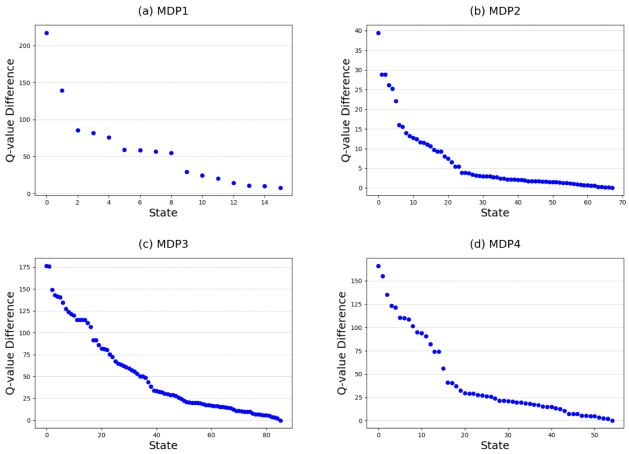

Figure 3: Q-value difference in MDP1-MDP4

Table 1 shows the empirical evaluation results by comparing the High vs. Low carry-out groups' post-test scores based on the corresponding RL-induced policy. The motivation is that, if a policy is indeed effective, students in the High group should significantly out-perform their Low peers on the post-test. In Table 1, the first column indicates the name of the RL-induced policy; columns 2 and 3 show the mean and standard deviation of classroom post-test scores for High and Low carry-out groups, respectively, and the last column shows the t-test results when comparing post-test between groups. Rows 2-4 show that there is no significant difference between the High vs. Low groups in terms of post-test scores for MDP1, MDP2, and MDP3 policies, but there is a significant difference between the two groups for the MDP4 policy (row 5): $t(448) = 2.19, p = .029$. This result suggests that, among the four MDP policies, only MDP4 seems to be effective in that the students in MDP4's High carry-out group performed significantly better than those in the Low carry-out group.

Table 2 shows the overall IS-based metrics evaluation results showing the impact of policy transformations and original versus normalized rewards on the outcomes of each IS metric. The first column indicates the type of policy transformation applied and the second column shows whether the rewards are normalized. The third through fifth columns show the performance of each IS metric, where performance is determined by the percent alignment between the IS policy predictions and empirical post-test results. Among the three policy transformations, Q-proportion is the worst, with none of its six performance results better than the corresponding results of Soft-max or Hard-code. Hard-code performs slightly better than Soft-max in most cases but never reaches a 100% match. For reward normalization, the original reward performs better than the normalized reward for the PDIS metric, but there was no effect on IS or WIS.

Comparing the three IS metrics, PDIS shows the greatest performance with all 12 of the performance results being the best. For the last two metrics, IS and WIS, the results are exactly the same because WIS is much like multiplying IS by a constant and this kind of re-scale won't change the result of the t-tests. The metric with the best performance is PDIS with Soft-max policy transformation and the original reward, whose performance is 100%. This means that all the t-test results on the PDIS predictions aligned with those on the empirical results

Table 1: Empirical Post-test Evaluation Results for High and Low Carry-out Groups

Policy	High	Low	T-test Result
MDP1	79.06(24.64)	83.13(23.69)	$t(448) = -1.78, p = .076$
MDP2	81.46(24.90)	80.44(23.66)	$t(448) = .44, p = .658$
MDP3	82.55(23.48)	79.30(25.00)	$t(448) = 1.24, p = .156$
MDP4	83.43(22.83)	78.43(25.44)	$\mathbf{t(448) = 2.19, p = .029}^{**}$

bold and ** denote significance at $p < 0.05$.

Table 2: Policy Transformation and Normalization Impacts on IS Metric Alignment to Post-test Outcomes

Policy Transformation	Rewards	Metrics		
		IS	WIS	PDIS
Soft-max	Original	50%	50%	**100%**$*$
	Normalized	50%	50%	50%
Q-proportion	Original	25%	25%	75%
	Normalized	25%	25%	25%
Hard-code	Original	75%	75%	75%
	Normalized	75%	75%	75%

Table 3: Detailed IS-based Metrics Evaluation Results Using Original Reward

Transform	Policy	Empirical Result	IS & WIS Result	PDIS Result
Soft-max	MDP1	$t(448) = -1.78, p = .076$	$t(448) = .89, p = .376$	$t(448) = .89, p = .375$
	MDP2	$t(448) = .44, p = .658$	$\mathbf{t(448) = 2.60, p = .010}^{**}$	$t(448) = 1.84, p = .067$
	MDP3	$t(448) = 1.42, p = .156$	$\mathbf{t(448) = 2.50, p = .013}^{**}$	$t(448) = .53, p = .594$
	MDP4	$\mathbf{t(448) = 2.19, p = .029}^{**}$	$\mathbf{t(448) = 2.18, p = .030}^{**}$	$\mathbf{t(448) = 3.23, p = .001}^{**}$
Q-proportion	MDP1	$t(448) = -1.78, p = .076$	$\mathbf{t(448) = 3.78, p < .001}^{**}$	$t(448) = 1.77, p = .077$
	MDP2	$t(448) = .44, p = .658$	$\mathbf{t(448) = 3.26, p = .001}^{**}$	$\mathbf{t(448) = 2.13, p = .034}^{**}$
	MDP3	$t(448) = 1.42, p = .156$	$\mathbf{t(448) = 2.71, p = .007}^{**}$	$t(448) = .31, p = .760$
	MDP4	$\mathbf{t(448) = 2.19, p = .029}^{**}$	$\mathbf{t(448) = 3.69, p < .001}^{**}$	$\mathbf{t(448) = 2.32, p = .021}^{**}$
Hard-code	MDP1	$t(448) = -1.78, p = .076$	$t(448) = 1.19, p = .233$	$t(448) = .96, p = .337$
	MDP2	$t(448) = .44, p = .658$	$t(448) = .71, p = .479$	$t(448) = .84, p = .404$
	MDP3	$t(448) = 1.42, p = .156$	$\mathbf{t(448) = 2.34, p = .020}^{**}$	$\mathbf{t(448) = 2.06, p = .040}^{**}$
	MDP4	$\mathbf{t(448) = 2.19, p = .029}^{**}$	$\mathbf{t(448) = 2.83, p = .005}^{**}$	$\mathbf{t(448) = 3.73, p < .001}^{**}$

Electric-blue cells denote that the theoretical t-test results align with the empirical t-test results (Column 3);
Grey cells denote misaligned t-test results.

in terms of significance match.

Table 3 shows the detailed results for metric evaluations using the original reward, providing t-test results when comparing post-test results between High and Low carry-out groups. The first column in table 3 shows the type of policy transformation functions applied. The second column shows the four MDP policies considered when splitting the dataset into High and Low carry-out groups. The third column shows the t-test results of the empirical evaluation of High vs. Low carry-out, which served as the ground truth. The fourth and fifth columns show the t-test results for the prediction of the three IS metrics: IS, WIS, and PDIS respectively. In the tables, electric-blue cells denote that the theoretical t-test results align with the empirical t-test results (Column 3) while grey cells denote mismatched t-test results. From this table, we can see that only PDIS with soft-max transformation and the original reward results in all four t-tests aligning with the corresponding empirical results. IS and WIS are more likely to predict the significant difference between High vs. Low. Meanwhile, Q-proportion tends to cause the metric to predict more significant difference, while Hard-code tends to predict less.

In summary, when comparing groups in the RL-induced policy, our results showed that for the MDP4 policy, the students in the High carry-out group significantly outperformed the students in the Low group, but no significant difference was found in the other three policies. This suggests that the MDP4 policy is an effective policy in that the more it is carried out, the better it performs. However, the partially carry-out situation reduced the power of the MDP4 policy so that it did not significantly outperform the baseline random policy. When comparing the empirical evaluation results with theoretical evaluation results, PDIS is the best among the three IS-based metrics, reaching

Table 4: Critical Decision Evaluation Results

Policy	More	Less	T-test Result
MDP1	78.49(23.08)	83.01(25.44)	$t(448) = -1.97, p = .049$
MDP2	83.22(23.53)	78.86(24.79)	$t(448) = 1.91, p = .057$
MDP3	79.45(24.59)	82.08(24.00)	$t(448) = -1.14, p = .257$
MDP4	83.54(22.89)	78.74(25.21)	$\mathbf{t(448) = 2.10, p = .036}^{**}$

bold and ** denote significance at $p < 0.05$.

100% agreement. Our results suggested that proper deployment settings have an impact on the performance of IS-based metrics. When transforming the deterministic policy to stochastic policy, soft-max is the best one, while Q-proportion is the worst, and Hard-code is stable. The comparison between the original reward and normalized reward indicates that the original reward can better reflect the empirical results despite having larger variance.

8.2 Critical Decision Identification

Recall that each policy impacts its own critical decisions: those with higher differences between Q-values for possible decisions are considered to be critical, and we split each group of students according to whether the student received More decisions aligned with the critical decisions. Table 4 shows the t-test results comparing the post-test scores between the More vs. Less critical decisions groups. The first column indicates the MDP policies considered when identifying the critical decision. The second and third columns show the average post-test scores of students in the More and Less groups, showing mean(sd). The fourth column shows the t-test results when comparing the post-test scores of the More and Less critical decisions groups. The MDP1 row shows a significant difference between the More vs. Less critical decisions groups for the MDP1 policy, with $t(448) = -1.97, p = .049$. However, the students in More group perform worse than the Less critical decisions group. The MDP2 and MDP3 rows show that there is no significant difference between the two critical decisions groups in terms of post-test scores for the MDP2 or MDP3 policies. Finally, the MDP4 policy shows a significant difference between the two groups, $t(448) = 2.10, p = .036$, which means that students with More critical decisions performed significantly better than students with Less.

For the MDP4 policy, the identified critical decisions comprised 25% of all decisions. This shows that, although critical decisions are a small proportion of all decisions, they can significantly impact the outcome. Also, the results also show that the Q-values in the MDP4 policy can be used to identify critical decisions aligned with empirical results, but the other three cannot. Based on the results from Table 1, MDP4 was identified as the only effective policy, since its empirical post-test results aligned, with students in the High carry-out group performing significantly better than the Low carry-out group. The critical decisions results suggest that MDP4 is also the only policy where larger differences in Q-values had larger impacts on post-test results. Taken together, these results suggest that only Q-values in effective policies work to influence decisions that impact actual post-test performance. This further inspires us to investigate whether we could verify the effectiveness of a policy in reverse: Given a policy, if the decisions with larger Q-value difference are significantly linked to the student performance, then this policy may be more likely to be effective.

9. CONCLUSION AND FUTURE WORK

In this work, we explored three IS-based OPE metrics with two deployment settings in a real-world application. Through comparing the effectiveness of four RL-induced policies empirically and theoretically, our results showed that PDIS is the best one for interactive e-learning systems and appropriate deployment settings (i.e., where policy decisions are carried out) are required to achieve reliable, robust evaluations. We also proposed a method to identify critical decisions by the Q-value differences in a policy. In order to verify our method, we investigated the relationship between the number of identified critical decisions and student post-test scores. The results revealed that the identified critical decisions are significantly linked to student learning, and further, that critical decisions can be identified by an effective policy but not by ineffective policies. In the future, we will apply the PDIS metric with soft-max transformation and original rewards to help us induce better RL policies which further improve students' learning in the ITS. Also, when inducing policies, we will consider constraints to avoid the partially carry-out situation that limits the impact a policy can have on outcomes. Furthermore, with the identification of critical decisions, we can reduce the size of resulting policies still further by focusing only on the most monumental decisions rather than meaningless ones.

10. REFERENCES

[1] J. Beck, B. P. Woolf, and C. R. Beal. Advisor: A machine learning architecture for intelligent tutor construction. In *AAAI/IAAI*, pages 552–557, 2000.

[2] J. D. Bransford and D. L. Schwartz. Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, pages 61–100, 1999.

[3] M. Chi, K. VanLehn, D. Litman, and P. Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1-2):137–180, 2011.

[4] B. Clement and et al. A comparison of automatic teaching strategies for heterogeneous student populations. In *EDM 16-9th EDM*, 2016.

[5] S. Doroudi, K. Holstein, V. Aleven, and E. Brunskill. Towards understanding how to leverage sense-making, induction and refinement, and fluency to improve robust learning. *JEDM*, 2015.

[6] L. J. Dudik, Miroslav and L. Li. Doubly robust policy evaluation and learning. *the 28th International Conference on Machine Learning*, 2011.

[7] J. Hammersley and D. Handscomb. General principles of the monte carlo method. *Springer*, 1964.

[8] A. Iglesias, P. Martínez, R. Aler, and F. Fernández. Learning teaching strategies in an adaptive and intelligent educational system through reinforcement

learning. *Applied Intelligence*, 31(1):89–106, 2009.

[9] A. Iglesias, P. Martínez, R. Aler, and F. Fernández. Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems. *Knowledge-Based Systems*, 22(4):266–270, 2009.

[10] A. Iglesias, P. Martínez, R. Aler, and F. Fernández. Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems. *Knowledge-Based Systems*, 22(4):266–270, 2009.

[11] D. J. L. Joel R. Tetreault, Dan Bohus. Estimating the reliability of mdp policies: A confidence interval approach. *The Association for Computational Linguistics*, pages 276–283, 2007.

[12] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark. Intelligent tutoring goes to school in the big city. 1997.

[13] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[14] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23, 2000.

[15] T. Mandel, Y.-E. Liu, S. Levine, E. Brunskill, and Z. Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pages 1077–1084, 2014.

[16] Z. L. Mostafavi Behrooz and T. Barnes. Data-driven proficiency profiling. In *Proc. of the 8th International Conference on Educational Data Mining*, 2015.

[17] M. Chi, K. VanLehn, D. J. Litman, and P. W. Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Model. User-Adapt. Interact.*, 21(1-2):137–180, 2011.

[18] K. Narasimhan, T. Kulkarni, and R. Barzilay. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*, 2015.

[19] E. B. Philip S. Thomas. Data-efficient off-policy policy evaluation for reinforcement learning. In *In International Conference on Machine Learning*, pages 2139–2148, 2016.

[20] S. R. S. S. S. Precup, D. Eligibility traces for off-policy policy evaluation. *17th International Conference on Machine Learning*, pages 759–766, 2000.

[21] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching via pomdp planning. *Cognitive science*, 40(6):1290–1332, 2016.

[22] J. P. Rowe and J. C. Lester. Optimizing player experience in interactive narrative planning: A modular reinforcement learning approach. In *the Tenth International Conference on AIIDE*, pages 160–166, 2014.

[23] J. P. Rowe and J. C. Lester. Improving student problem solving in narrative-centered learning environments: A modular reinforcement learning framework. In *AIED*, pages 419–428. Springer, 2015.

[24] N. Roy, J. Pineau, and S. Thrun. Spoken dialogue management using probabilistic reasoning. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 93–100. Association for Computational Linguistics, 2000.

[25] S. Shen and M. Chi. Aim low: Correlation-based feature selection for model-based reinforcement learning. In *EDM*, pages 507–512, 2016.

[26] S. Shen and M. Chi. Reinforcement learning: the sooner the better, or the later the better? In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 37–44. ACM, 2016.

[27] S. Singh, D. Litman, M. Kearns, and M. Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16:105–133, 2002.

[28] J. C. Stamper, M. Eagle, T. Barnes, and M. Croy. Experimental evaluation of automatic hint generation for a logic tutor. In *International Conference on Artificial Intelligence in Education*, pages 345–352. Springer, 2011.

[29] P. Thomas. Safe reinforcement learning. *PhD thesis*, 2015.

[30] K. Vanlehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.

[31] P. Wang, J. Rowe, W. Min, B. Mott, and J. Lester. Interactive narrative personalization with deep reinforcement learning. In *IJCAI*, 2017.

[32] J. D. Williams and S. Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

[33] B. Zhang, Q. Cai, J. Mao, E. Chang, and B. Guo. Spoken dialogue management as planning and acting under uncertainty. In *INTERSPEECH*, pages 2169–2172, 2001.

[34] G. Zhou and et al. Towards closing the loop: Bridging machine-induced pedagogical policies to learning theories. In *EDM*, 2017.