Robust Evaluation Matrix
Towards a More Principled Offline Exploration of Instructional Policies

Shayan Doroudi
Computer Science Department
Carnegie Mellon University
shayand@cs.cmu.edu

Vincent Aleven
Human-Computer Interaction Institute
Carnegie Mellon University
aleven@cs.cmu.edu

Emma Brunskill
Computer Science Department
Carnegie Mellon University
ebrun@cs.cmu.edu

ABSTRACT
The gold standard for identifying more effective pedagogical approaches is to perform an experiment. Unfortunately, frequently a hypothesized alternate way of teaching does not yield an improved effect. Given the expense and logistics of each experiment, and the enormous space of potential ways to improve teaching, it would be highly preferable if it were possible to estimate in advance of running a study whether an alternative teaching strategy would improve learning. This is true even in learning at scale situations, since even if it is logistically easier to recruit a large number of subjects, it remains a high stakes environment because the experiment is impacting many real students. For certain classes of alternate teaching approaches, such as new ways to sequence existing material, it is possible to build student models that can be used as simulators to estimate the performance of learners under new proposed teaching methods. However, existing methods for doing so can overestimate the performance of new teaching methods. We instead propose the Robust Evaluation Matrix (REM) method which explicitly considers model mismatch between the student model used to derive the teaching strategy and that used as a simulator to evaluate the teaching strategy effectiveness. We then present two case studies from a fractions intelligent tutoring system and from a concept learning task from prior work that show how REM could be used both to detect when a new instructional policy may not be effective on actual students and to detect when it may be effective in improving student learning.

Author Keywords
instructional policies; reinforcement learning; off-policy; policy estimation; policy selection

L@S 2017, April 20 - 21, 2017, Cambridge, MA, USA

ISBN 978-1-4503-4450-0/17/04…$15.00

DOI: http://dx.doi.org/10.1145/3051457.3051463

INTRODUCTION
The gold standard for identifying more effective pedagogical approaches is to perform an experiment. Unfortunately, frequently a hypothesized alternate way of teaching does not yield an improved effect. Given the expense and logistics of each experiment, and the enormous space of potential ways to improve teaching, it would be highly preferable if it were possible to estimate in advance of running a study whether an alternative teaching strategy would improve learning. This is true even in learning at scale situations, since even if it is logistically easier to recruit a large number of subjects, it remains a high stakes environment because the experiment is impacting real students, and likely many more than in standard classroom environments.

It is possible to build student models that can be used as simulators to estimate the performance of learners under a variety of proposed teaching methods. In particular, one important open question in education is whether and how the sequencing of a given set of course activities impacts student learning. There are an enormous possible set of ways to sequence material, including the use of adaptive policies (like cognitive mastery learning [5] or reinforcement learning based policies [1, 3, 21, 24, 17, 15] which map representation of the current student state to a next pedagogical activity). Indeed prior work has suggested that the pedagogical activity to provide in terms of maximizing learning gains may depend on the student state [12], and offer significant benefits over randomly or suboptimally selecting such activities [5, 3, 15, 21]. To help estimate the potential performance of new sequencing approaches, we can build a student model, and use it as a simulator to approximate what the student learning outcomes might be when taught using the new sequencing policy. In particular, a common approach is to estimate the efficacy of a policy a priori by simulating its performance using a student model that is identical to the one used to compute the policy itself[1] [3, 17, 25]. Unfortunately, such an approach can overestimate the performance of the policy [16, 15], and relies on students

[1]Note that this simulation process can in fact be done in multiple ways, and often the process used to compute the policy given a student model may itself directly also yield an estimate of the performance (the student learning outcomes) of the policy assuming the student model it uses is in fact how students learn in the real world. (For example, in reinforcement learning, value iteration or policy iteration would yield such an estimate.)

really learning in the same way as the student model assumes. While we hope that we have good models of student learning, student modeling is an active area of research, and different models of student learning with similar predictive accuracies can yield very different policies or outcomes for students [23, 13, 28].

In contrast, in this paper, we present the robust evaluation matrix (REM) method for estimating the potential impact of a new way of teaching (focusing on sequencing strategies) in advance of an experiment. REM seeks to make our predictions more robust to model mismatch, the situation where the student model used to derive a policy is not the same as the true model that underlies student learning where the policy will be deployed. We present two case studies to show how REM can be used in practice. In the first, we demonstrate that our method can help correctly predict when a new policy would not be effective in improving student learning while standard model-based evaluation predicts otherwise. In the second, we show that REM could be used to detect policies that will do better than baselines in a concept learning domain when deployed on actual students, and again can detect cases where policies are likely to be ineffective in the real world.

We believe REM could be used to more cautiously estimate the potential impact of new strategies, and thereby has the potential to help guide and potentially reduce the number of experiments needed to find promising new teaching methods.

OFF-POLICY POLICY ESTIMATION AND SELECTION

We are interested in the related problems of off-policy policy estimation and off-policy policy selection: the setting where we have access to prior data collected using some policy, and we want to use that data to make inferences about one or more *other* (instructional) policies. Off-policy policy estimation can be used to estimate the performance of a new instructional policy without (or in advance of) running an experiment. Such counterfactual reasoning is important not just in education, but in a wide swath of other areas including economics, healthcare, and consumer modeling [27, 29]. Off-policy policy estimation is often a critical part of off-policy policy selection: determining which policy from among a set of candidate policies would have the highest expected performance if deployed in the future. We are primarily interested in the problem of off-policy policy selection, as it can have practical implications with respect to what we do in practice. We consider the problem of off-policy policy estimation in so far as it helps us achieve the former. As we will show, while the two have been tightly coupled in the literature, we present a method that does not necessarily give us reliable estimates of the performance of instructional policies but could still be used to compare instructional policies.

An important class of instructional policies that we focus on in this paper are policies that determine what problem/activity to present to a student at any given time based on features of a student's state (e.g., the student's performance on past problems, how long the student has spent on the system, the student's level of prior knowledge etc.). The performance of an instructional policy might be how much it improves student learning, which is often assessed in educational experiments by how well students do on a posttest given after instruction, or how much faster it helps students learn a fixed amount of material, which is often what is optimized in mastery learning contexts. We now discuss approaches to doing off-policy policy estimation and selection, including how this problem has been tackled in education settings as well as in some of the broader reinforcement learning literature.

Model-Based Evaluation

To leverage the data collected from another policy, a common approach to doing off-policy policy estimation is to first use that data to fit the parameters of a statistical (student) model [3, 17, 21]. Given such a model, we can then use that model as a simulator to evaluate the performance of any compatible alternate (instructional) policy. Compatibility here involves two aspects. The first is that the student simulator model must be a generative model capable of simulating any observations required by the instructional policy. For example, for the following instructional policy,

---

if *student took* $>$ 100 *seconds to complete previous problem*
then
  give multiplication problem
else
  give division problem

---

the student simulator model must be capable of generating the amount of time a simulated student takes to do each problem. Second, the alternate policy can only select an activity given a student state which the model is capable of simulating an outcome for. Implicitly, this means that the data used to train the model parameters must have included making a similar decision for another student in that state, and observing some outcome. This indicates that the collected data supports the alternate policy. More intuitively, consider collecting data from an instructional policy (call it policy 0) that randomly either gives a student a worked example or a short video whenever the student first logs into the educational software. Now consider a new instructional policy (call it policy A) that provides a student with a quiz when he first logs into the educational software. The old policy never provided students with a quiz upon logging in to the system, and so a model of student learning and the impact of activities on the student's learning state will not include any estimate of what it would be like if the student were to get a quiz in this situation. In this situation, the alternate policy cannot be simulated. In contrast, consider another new instructional policy (policy B) that always provides students with a worked example when they first log in to the tutoring software. In this case, we can simulate the potential performance of policy B, because the statistical student model of learning that was estimated from policy 0 includes an estimate of potential outcomes that could occur in this setting. Throughout the rest of this paper, we will focus our attention on considering instructional policies that are compatible with the previously collected data– that is, policies whose outcomes could be simulated by a generative student learning model estimated from the collected data.

In model-based off-policy estimation, two immediate questions arise: (1) given a particular (student) model class and a dataset, how do we estimate the performance of a policy, and (2) how do we select which model class to choose? Here, model class refers to the type of statistical model used to represent student learning. There are many models of student learning considered in the literature, including Bayesian Knowledge Tracing [6], logistic regression models like performance factors analysis [18], Markov decision processes (MDPs) [3, 25], partially observable MDPs (POMDPs) [21], and Deep Knowledge Tracing [19].

Given a model class and a dataset, the typical approach is to use machine learning to fit the parameters that best model the available data (such as finding the maximum likelihood model, or a model that minimizes a desired loss function). Using the resulting fit model parameters, we can then simulate how a student might learn under a desired compatible policy. To assess how good a policy is, we also need a way to evaluate the student learning outcomes generated under a specific policy. In reinforcement learning this is typically known as the reward model, which could, for example, provide a positive reward when a student gets a test question correct. Together the student learning model and reward model can be used to both simulate a student's learning under a compatible policy, and evaluate the quality of the resulting simulated outcomes. If the student learning model plus the reward model constitute a Markov decision process, there exist well known algorithms (such as value iteration or policy iteration) for computing an instructional policy that achieves the maximal expected policy performance under that student learning model and reward function. This approach has been leveraged in multiple educational data mining research projects and reinforcement learning settings [3, 17, 25].

There are at least two issues that arise with this approach. First, given finite data, the estimated model parameter values will be approximate, and these parameter uncertainties can result in error in the resulting estimated performance of a policy, especially when that policy is designed to maximize performance for that model [16]. There do exist multiple techniques to quantify the amount of error in the resulting estimated policy performance due to parameter uncertainty, some of which have been previously considered in the educational technology literature [16, 3]. In addition, large scale datasets like those collected when learning at scale reduce parameter uncertainty, as in general the more data we have the more precise our parameter estimates will be.

The other, larger issue, is that the chosen model class may be a poor approximation of how students learn, and may yield misleading estimates of the performance of a proposed policy. This relates to the second critical issue: how do we select which model class to choose?

Indeed, there is a vast amount of research on student modeling, and a common way to evaluate and compare potential student model classes is by their predictive accuracy, such as using cross validation root mean squared error on an input dataset. One natural idea then is to select the model class with the smallest predictive error, and then compute an instructional policy with the highest predicted performance for the given model class as fit to the available data. Unfortunately, even if two different student model classes have similar predictive accuracy when fit to a particular dataset, they may have very different implications for what instructional policy will be most effective [23, 13, 28]. Moreover, prior work has shown that just selecting the student model with the highest accuracy on an input dataset may not be the model whose best associated policy has the highest performance for real student learning [15]. Other work has shown the limitations of considering model accuracy and how it does not capture the information most meaningful for decision making [2, 9]. All of this suggests that model accuracy alone is not sufficient for deciding which model class to select.

A second approach is to select the student model and instructional policy that under that student model is expected to have the best performance (e.g. best student learning outcomes). This requires deriving a policy for each model under consideration (often the optimal policy or an approximation of the optimal policy for that model) and evaluating that policy by simulating it on the model used to derive it. We call this **direct model-based evaluation**. This approach has been used to compare and select among different student learning models and their optimal policies. Chi et al. used this approach to select an instructional policy, by comparing different student learning models represented as Markov decision processes with different student features and the resulting instructional policy that yielded the best expected performance for a given model [3]. Similarly, Rowe et al. estimated the predicted performance of instructional policies that were designed to maximize performance under particular student models and compared them to some hand designed baseline policies and a random policy by evaluating these policies under the same student models. Unsurprisingly, the policy that was computed to have the best predicted performance for a given student model was also estimated to to out-perform the baseline policies under that same model [25].

This approach is quite appealing, as it is more directly getting at what we often care about: estimating the performance of policies in order to select a policy with the best expected performance. Unfortunately, since any student model will not generally capture the way that real students learn (even given infinite amounts of data used to estimate the model parameters), evaluating a policy assuming the model it was derived under is correct will generally not provide an accurate estimate of the value of a policy if it were to be used with real students. Comparing the estimated performance of policies when each policy is evaluated using a different simulated student model can therefore yield misleading conclusions. Indeed Mandel et al. have shown that *even if* the real world can be accurately modeled as a complex Markov decision process, it is possible that the optimal policy for an alternate statistical model that is incorrect might have a higher estimated performance than the optimal policy of the true MDP, even with an infinite amount of data[2] [15]. Therefore, this is not a problem that learning at scale alone can solve.

[2]This is because the alternate statistical model may not satisfy the Markov property.

Indeed, the limitations of evaluating the performance of a policy with the student model used to derive the policy has been observed previously. In simulation, Rowe et al. estimated a new instructional policy would have a performance of 25.4 in contrast to a random policy that was estimated to have a performance of 3.6, where performance was measured as a function of students' normalized learning gains[3] beyond the median student and the performance of both policies was simulated with the student model used to derive the new instructional policy[25]. In contrast, in an experiment with real students, there was no significant difference between the performance of students taught by the two policies [24]. While there are many factors in any experiment with real students, estimating performance using the assumed student model may particularly lead to overly optimistic estimates of the resulting performance. In this paper, we will present other situations where doing so incorrectly predicts a difference in performance between policies that is not found in an actual experiment, but where our alternate procedure (to be described shortly) would have correctly anticipated no significant difference in performance among the policies.

Importance Sampling

Using prior data to obtain an estimator of an instructional policy's performance in advance of deploying the new policy that is not biased by assuming particular statistical student model could seem rather difficult. However, there does exist an elegant solution: importance sampling, an approach that does not require building a student model, but rather re-weighs past data to compute an estimate of the performance of a new policy [20]. Importance sampling is statistically consistent and unbiased. In prior work, Mandel et al. used importance sampling to find an instructional policy in an educational game that significantly outperformed a random policy and even an expert-designed instructional policy [15]. Unfortunately, importance sampling tends to yield highly variable estimates of a new policy's performance when evaluating instructional policies that are used for many sequential decisions, such as students interacting with a tutoring system across many activities. Intuitively this issue arises when a new policy is quite different from a previous policy, and so the old data consists of quite different student trajectories (sequences of pedagogical activities given and student responses) than what would be expected to be observed under a new policy. Mathematically, this is because importance sampling yields unbiased but high variance estimates, unlike direct-model based evaluation which can yield very biased estimates (due to choosing an inaccurate model class) with potentially low variance (when we have enough data).

It is true that with more data, the variance of the importance sampling estimator will decrease, so one may assume this should be the method of choice for learning at scale, but this is not the case when one has to make a large number of sequential decisions. For example, consider some educational software that presents 20 activities to students and only needs to choose between one of two options at any given time (for example, whether to give the student a worked example or a problem-solving exercise). Suppose we have collected existing data from a policy that randomly chose each option for each of the 20 decisions and want to use this for off-policy policy estimation. If we want to evaluate a deterministic instructional policy (i.e., a policy with no randomness), then only one out of every 2^{20} (over one million) students would encounter a trajectory that matches the policy of interest, which means we need millions of students to get a decent estimate of the policy. If the software were to make 50 decisions, then we would need over 10^{15} students!

Finding a statistical estimator that offers the best of both approaches (model-based evaluation and importance sampling estimators) is an active area of research in the reinforcement learning community [8, 11, 26] but remains a challenge whenever the (instructional) policies may be used to make a large number of decisions, as highlighted above.

ROBUST EVALUATION MATRIX (REM)

Ideally we want a method for off-policy policy estimation that combines the statistical efficiency of (student) model based estimators with the agnosticism of importance sampling techniques which allows them to be robust to the choice of student model used to derive a particular policy. As we previously argued, this is important even given an enormous amount of data. One potential avenue is to focus on designing better student models, a key effort in the educational data mining and artificial intelligence in education communities. However, since these model classes will still likely be approximate models of student learning, we propose an alternative approach that may not enable us to achieve accurate estimates, but can still help inform comparisons among different policies: using many models we expect to be wrong, rather than using one model we hope to be right.

Our robust evaluation matrix (REM) is a tool for more conservatively evaluating the potential performance of a new policy in relation to other policies during off-policy policy selection. As shown in Algorithm 1, the simple idea is to estimate the performance of different instructional policies by simulating them using multiple plausible student models whose model parameters were fit using previously collected data. The rows of the matrix are different student models and the columns of the matrix are the various policies one wants to estimate the performance of. An entry in the matrix represents the expected performance of a particular instructional policy when simulated under a particular student model. As the student model simulators have parameters that are fit based on the previously collected data, they will often represent reasonable possible ways of modeling the dynamics of student learning. If we restrict our comparison to models with similar predictive accuracy (e.g., as evaluated using cross validation or a test set constructed from the available data), it is unclear which model is better, but the REM method can be used to assess trends in performance across policies that are consistent across multiple possible ways that students may learn in the real environment (e.g., Bayesian Knolwedge Tracing, Performance Factors Analysis, Deep Knowledge Tracing etc.).

[3]The normalized learning gain for a student is the difference between the posttest score and pretest score of the student divided by the maximum possible difference.

Simulating the potential performance of instructional policies under multiple student models to inform off-policy policy selection has been previously underexplored. There has been some prior work that analyzes the interaction of student models and instructional policies (that may have been derived with a particular student model) [21, 13, 23, 9, 4], but such work has often been done to understand the general differences between policies run on various models, rather than as a tool to inform whether a new policy may offer benefits over previous ones before conducting experiments or embedding a policy in a tutoring system. One exception is work by Clement et al., where they investigate the case where the knowledge graphs (i.e., prerequisite relations between knowledge components) used to learn models used to compute policies are not the same as the ones underlying student learning [4]. The authors found that a particular model that does not have parameters fine-tuned to the knowledge graph performs best when there is a mismatch in the policy's representation of knowledge graph and true knowledge graphs of students. Their work differs from our current paper in that the authors only consider robustness of policy's of varying complexity in light of the knowledge graph changing but do not consider student models that differ more wildly and the authors do not present a general method for off-policy policy estimation or selection. Moreover, they only presented results from simulations with hand-crafted parameters rather than models and policies fit to real data. Nonetheless, we can consider this work as an example of REM being used in the past to inform policy selection. The most closely related work is by Rafferty et al.[21], which analyzed the potential performance of various instructional policies derived from different models of student concept learning under various student concept learning models that were fit from a previously collected dataset. However, unlike our current paper, they presented this idea primarily to understand the interaction between the policies and the models of student learning (e.g. could a policy assuming a very simple model of student learning still do well if the real student exhibits much more complicated student learning), rather than as a generic tool for off-policy policy estimation and selection. In the next section, we reinterpret their results as a positive use case of REM. Moreover, while Rafferty et al. consider simulating policies only on models of student learning that were used to derive some of the policies, REM could simulate policies on other models of student learning, even if one does not derive any policies from those student models. We present one example of this in the next section.

REM can be used in several ways. If one or more student models in the matrix suggest that a new policy is no better or even worse than other (baseline) policies, then it would suggest a new policy may not yield a significant improvement in learning outcomes. On the other hand, if the student models agree that one policy appears to be better than others (and these student models are indeed quite different from each other[4]), then it should increase our confidence that the policy will actually out-perform the other policies. Recall that we are

Input: Set of models $m = 1 \ldots M$ and policies $p = 1 \ldots P$
REM $\leftarrow m \times p$ matrix
for *model* $m = 1 \ldots M$ **do**
  for *policy* $p = 1 \ldots P$ **do**
    if *model m compatible with policy p* **then**
      mean, stddev $\leftarrow$ Estimate performance of policy p on model m `// For example by simulating many times`
      REM$[m][p] \leftarrow$ mean, stddev
return *REM*

Algorithm 1: Pseudocode for algorithm to fill in robust evaluation matrix.

interested in the joint problems of off-policy policy estimation and off-policy policy selection. We propose that REM can help with addressing the second problem, even though it does not necessarily help us with the first. That is, if we find a policy that robustly does better than another policy according to various student models, then we may decide to choose to implement that policy in practice; however, if different student models have very different predictions as to how well the new policy will perform, then we may not have a good estimate of its performance a priori. But having an estimate of a policy we are confident will do well a priori may not be necessary if we are planning on testing it on actual students anyways. This makes REM differ from off-policy policy selection techniques in the existing literature, which aim to use imperfect methods of policy estimation as a way to do policy selection. Rather, REM aims to help the researcher make decisions about what policy to select without directly trying to get a good estimate of a policy's performance.

CASE STUDIES

We now present two case studies to ground the discussion and illustrate how REM can inform what instructional policies may yield improved performance, given prior data. The first is an experimental study we ran in which we used old data to derive a new policy we estimated to be better than a standard baseline, but which yielded equivalent performance in a subsequent student study. Our post hoc analysis suggests we could have predicted this result by using a REM analysis. In the second case study, we will look at the results of a paper by Rafferty et al. where they perform an analogue of REM to better understand how various instructional policies might perform under different student models [21]. Although their paper did not suggest using such a method for off-policy policy selection, we show two examples of how it could have been used both to predict that several policies were likely to do well when tested on real students and to predict that another policy may perform poorly (a result that would not have been predicted if a policy's performance was only estimated assuming that the student model used to derive the policy was in fact how students truly learn).

[4]The difference in student models could be based difference in theory, for example a Bayesian Knowledge Tracing model and a Deep Knowledge Tracing model make rather different assumptions about student learning—or based on empirically observing that simulating the same instructional policy on two different models results in reasonably different trajectories quantified in some way.

Case Study 1: Fractions Tutor Experiment

We ran an experiment to test five instructional policies in an intelligent tutoring system (ITS) designed to teach fractions to elementary school students [22, 7].

There were two main goals to the experiment: (1) to test whether adaptive problem selection based on an individual student's knowledge state makes a difference (in terms of improving student learning), and (2) to test whether supporting a variety of activity types in an ITS leads to more robust learning. Additionally, we were interested in testing whether we could improve upon the traditional form of adaptive instruction used in ITSs: cognitive mastery learning using Bayesian Knowledge Tracing (BKT). Namely, we were interested in testing whether reasoning about (prerequisite) relationships between skills when deciding what problem to give a student to solve improves student learning beyond simply giving problems until a student masters each skill independently. We therefore developed a new student model that treats the correctness on the last two steps of each skill as the state of a student's knowledge of that skill, and then predicts the student's next state of a skill based on the student's knowledge of that skill as well as prerequisite skills. Prerequisite skills were identified using the G-SCOPE algorithm [10]. Our models used a skill model that was inferred using the weighted Chinese restaurant process technique developed by Lindsey et al. [14], which was seeded with a hand-crafted skill model. Model parameters were fit given access to data that was previously collected using a semi-random instructional policy to teach over 1,000 students, who used the tutor for four to six days, with most students completing between 20 and 100 problems out of a potential set of 156 problems. Student learning was assessed using identical pretests and posttests composed of 16 questions.

We iterated over multiple potential adaptive instructional policies, seeking to identify a policy that we estimated would yield improved performance over both strong baseline non-adaptive policies, and equal or better performance to a state-of-the-art policy based on mastery teaching. Since each student completed many problems using the tutor, typically more than 20, importance sampling techniques for estimating the student learning outcomes under an alternate instructional policy (that adaptively sequenced activities in a different way) were infeasible (see example above). Instead, we relied on simulating a policy's performance based on a student learning model. We choose adaptive policies that we estimated would yield a significant improvement over the non-adaptive baselines. This lead us to choose the following adaptive policies for use in a future experiment, policies that we believed had a good chance of yielding a significant improvement,

- Adaptive Policy 1 (**AP-1**): greedily maximize the number of skills that students learn with each problem assuming the fit G-SCOPE model.
- Adaptive Policy 2 (**AP-2**): Selects problems to myopically maximize the student's posttest score under a fit G-SCOPE student model.

These were to compared to the following baselines

- Baseline 1: Instructional policy that selects standard (induction and refinement) problems, in a reasonable non-adaptive order, based on spiralling through the curriculum.
- Baseline 2: Instructional policy that selects among a diverse set of problem types, in a reasonable non-adaptive order, based on spiralling through the curriculum.
- BKT Mastery Policy (**BKT-MP**): This is a state-of-the-art cognitive mastery learning policy used with a Bayesian Knowledge Tracing model which has been previously shown to yield substantial improvements in student learning [6].

Row 1 of Table 1 shows the estimated performance of the above policies, where each adaptive policy was simulated using the student model used to derive the policy. Since the first two policies are non-adaptive, they were not derived using a student model. We used the G-SCOPE student model to simulate the performance of these baseline non-adaptive policies. All evaluations assumed each (simulated) student completed 40 problems, and we repeated this process with 1,000 simulated students.

Using these off-policy policy performance estimates, the predicted Cohen's *d* effect size of AP-2 vs. Baseline 2 is 3.66 and the predicted effect size of AP-2 vs. Baseline 1 is 4.14, indicating that the new adaptive policies may yield a large improvement in robust student learning.

However, in our subsequent experiments there was no significant difference in the performance of students taught in the different policies as shown in Row 2 of Table 1.

We now consider the insight we could have obtained by using REM. We apply REM to our policies by evaluating them on three models: (1) the G-SCOPE model (which was used to derive AP-1 and AP-2), (2) the BKT student model (which was used to derive BKT-MP), and (3) a Deep Knowledge Tracing (DKT) model [19]. The results are shown in Table 2.

Using the BKT student model, we see that all the policies appear to have much more similar expected performance than when using the G-SCOPE student model, though the new adaptive policies are still expected to be as good or better than the state-of-the-art BKT mastery policy in either situation, and an improvement over the non-adaptive policies. Therefore, were we only to simulate policies under the models used to derive the policies, we might still expect that the new adaptive policies would yield improved performance.

The key distinction comes up when we also simulate under another plausible student model, which was not used to derive a particular student policy. In contrast to the other student models, simulating using a Deep Knoweldge Tracing student model actually predicts that Baseline 1 will yield the highest expected student learning performance, and be substantially higher than the predicted performance of the adaptive instructional policies.[5] Since three student models (BKT, G-SCOPE

[5]This Deep Knowledge Tracing model was introduced by Piech et al. [19] after these experiments were conducted, so interestingly, we could not have done this analysis prior to running our experiment.

	Instructional Policies				
	Baseline 1	Baseline 2	BKT-MP	AP-1	AP-2
Direct Model-Based Evaluation Results	5.87 ± 0.90	6.10 ± 0.97	7.03 ± 1.00	7.85 ± 0.98	9.10 ± 0.80
Actual Experimental Results	5.52 ± 2.61	5.14 ± 3.22	5.46 ± 3.0	5.57 ± 3.27	4.93 ± 1.8

Table 1. The first row shows the estimated expected performance of a student when taught under each policy, assuming either the student model used to derive the policy, or, in the case of the non-adaptive policies, using the estimated G-SCOPE student model. The second row shows the results of our actual experiment. Note that the posttest was out of sixteen points.

		Instructional Policies				
		Baseline 1	Baseline 2	BKT-MP	AP-1	AP-2
Student Models	New Student Model	5.87 ± 0.90	6.10 ± 0.97	N/A	7.85 ± 0.98	9.10 ± 0.80
	BKT Student Model	6.46 ± 0.78	6.65 ± 0.95	7.03 ± 1.00	6.82 ± 0.94	7.04 ± 0.96
	DKT Student Model	9.89 ± 1.45	8.69 ± 1.82	8.55 ± 2.08	8.31 ± 2.22	8.58 ± 2.13

Table 2. Robust evaluation matrix showing predictions of the five policies in our experiment according to the new student model as well as the BKT student model and a DKT student model. Notice that BKT-MP was not simulated on the new student model since they were not exactly compatible due to a nuance in the way they represent steps.

and DKT) are all seemingly reasonable choices of student models with similar predictive accuracies (RMSE between 0.41 and 0.44), our robust evaluation matrix suggests that we should not have been confident that new adaptive policies would yield a large effect size improvement over non-adaptive baselines or even necessarily be better than the non-adaptive policies (thus consistent with the lack of difference in the true experimental results).

Therefore, in this case REM could have served as a diagnostic tool to identify that our new proposed adaptive policies might not yield the significant improvement we hoped for, by explicitly considering whether this improvement is robust across many plausible student models.

Case Study 2: Concept Learning

In Rafferty et al. [21], the authors consider three instructional policies for concept learning. The models are derived under three different partially observable Markov decision process (POMDP) student learning models of varying complexity inspired from the cognitive science literature: a memoryless model in which a learner maintains a single potential concept until evidence contradicts the correctness of this concept, a discrete model with memory which augments the memoryless model to prevent the learner from forgetting prior negative evidence about the potential concepts, and a continuous model which assigns probabilities to different potential concepts [21]. The model parameters were fit with data the authors collected from students given a random policy. The performance of a policy is measured in how long (time in seconds) it takes for students to learn a series of rules or a concept.

Like REM, the authors first simulate each policy on each of the three student models, but unlike REM, the authors only consider models that are used to derive some instructional policy (and no other student learning models). This is because the authors are interested in the interaction of student models with policies derived from student models and what that says about human learning, rather than using this simulation as an off-policy policy selection tool to help decide which instructional policies may offer a benefit over existing benchmarks. Indeed, Rafferty et al. test all policies with real students. We reinterpret their results in terms of insights REM would have offered about the relative expected performance among the policies.

In the first experiment, the authors find that in simulation, all three student models agree that the three policies induced by the POMDPs would enable student to learn the rules faster than a random policy (i.e., the memoryless, discrete with memory, and continuous policies do better than the random policy *in all three rows* of the robust evaluation matrix). We propose this should lead a practitioner to believe that these three policies will likely do better than a random policy when presented to actual students (if the student models are believed to be decent). Indeed, in their experiments, the authors found that all three POMDP policies induced a smaller average time to mastering the rules than the policy which selects activities randomly, two of which were statistically significantly faster.

In this situation REM consistently estimated that the adaptive policies would have higher performance than the random activity selection policies, under 3 different student models, and this result was confirmed experimentally. This shows a situation where REM consistently identified a predicted improvement, under a variety of student models.

We now consider another example from this work where REM could have helped predict that a policy would likely not work well in practice, but evaluating policies only under the models used to derive that policy would fail to identify this issue.

In their Experiment 3, Rafferty et al. compare various policies on three concept learning tasks both in simulation (under all three student models) and in an actual experiment. The following result is of most interest to us: when using the continuous POMDP model to simulate student learning, they find that a heuristic greedy policy derived from this model—the maximum information gain policy—does significantly better than both the random-action-selection policy and the two POMDP polices derived from other POMDP models. This was estimated to hold in all three concept learning tasks. However,

in the actual experiment with students, the maximum information gain policy yields lower student performance than the random action selection policy and all the POMDP policies for all three concept learning tasks. This result could have been detected using REM, as both the memoryless model and the discrete model with memory estimated that the performance of the maximum information gain policy would be lower than the estimated performance of the random-action-selection instructional policy in at least one concept learning task. In this situation REM would have restricted the confidence with which one could expect the new policy to yield a big improvement in performance.

DISCUSSION

In some cases, REM might result in one being overly-conservative by not deploying an instructional policy that is actually worthwhile, but at the end of the day, it is up to each researcher to decide if they want to try a policy they think might result in improved student learning, even if they do not have strong evidence that it will, or if they would rather find a policy they are confident would result in an improvement. One can attain such confidence (although not in any statistically precise sense) if one finds a policy that does very well under various student models as we saw an example of in Case Study 2. However, as we have emphasized several times, this confidence depends on being convinced that our choice of student models to use in the matrix was good. As we mentioned, we do not expect any of these student models to be correct, so what does it mean for a model to be "good"? A necessary condition is that such a model should be able to differentiate between different policies. For example, a model that predicts students are always in the same state (perhaps determined by their prior knowledge or pretest scores) and never learn would not be a good model to use in REM, because it would predict all instructional policies result in equal student outcomes. One way to avoid such "bad" models is to avoid models with bad predictive accuracy; even if high predictive accuracy is not a good indicator of a model's ability to suggest good instructional policies, an especially low predictive accuracy should be a red flag.

So far we have been discussing how REM can help address the problem of wrong classes of student models. But notice that REM can also help address two other related issues that may arise in educational contexts and certainly did arise in Case Study 1. First, recall that in the fractions tutor case study, the off-policy estimation was based on assuming students would do 40 problems each (i.e., we simulated trajectories of 40 problems). In reality, trajectories will be of varying length due to a number of factors: some students work faster than others, some students spend less time working or may be absent on certain days of our experiment, etc. However, even if we consider the variance in trajectory lengths that existed in our past data, the evaluation results would be similar. But one thing we did not consider is that the distribution of trajectory lengths varies for different instructional policies. For example, students who had the Baseline 1 policy, did around 48 problems on average, whereas for all the other policies, the average was 28 problems or less. This is, at least in part, because Baseline 1 only gives problems of a particular activity type (induction and refinement), which tended to be the activity type that took the least amount of time on average. This could explain why Baseline 1 did as well as the other policies in our experiment; these students simply had more problems, which could make up for the lack of diversity or adaptivity of problems. To tackle this problem, we can consider different generative models of how many problems students will do given a particular instructional policy (for example by taking into account how long problems took students in our past data); we can then use these various models as different student models (i.e., different rows in our matrix) and see if any policies robustly do well with respect to these differences.

The second issue is that the classrooms that we ran this experiment in were very different from the classrooms we had collected data from previously to fit the models (and hence policies) used in this experiment. This mismatch in student population could mean that our student models learned from students of one population may not generalize to other student populations. For example, students in low-performing schools may have lower learning rates than students in high-performing schools, even if the model class could accurately model student learning. To our knowledge, this is an issue that is not well studied or solved in the education literature. To tackle the problem of mismatched student populations, we can fit our various student models to different subsets of our data corresponding to different student populations (assuming we have data from multiple student sub-populations), and then have these different models (of the same model class) form new rows in our matrix. Interestingly, Clement et al. cast their work as training models on different student populations (characterized by student's with certain knowledge graphs) and seeing how that generalizes to other populations of students (with different knowledge graphs) [4]; their work would be an instance of using REM to explore robustness of policies to different student populations in simulation. Table 3 shows a hypothetical matrix depicting how REM could potentially be used to tackle the various issues of general student model mismatch, varying trajectory lengths, and generalization of student population in tandem.

We wish to highlight that the case studies we have examined were retrospective. We hope that future studies will explore REM's use in a prospective manner, and how it might be leveraged to inform instructional design decisions for later use.

At this point we do not make any universal recommendations for how to use the robust matrix method to determine which instructional policy to use in the future. It is possible that one policy does not consistently do better than all other policies for every row of the matrix, but that it tends to do better, or that on average it does better. In this case, should we be confident in that policy? The answer must be determined on a case-by-case basis. The matrix might help reveal trends that can help the researcher determine whether a policy should be deployed or not. It is not an algorithm that will tell the researcher what to do; it is a heuristic that can help inform the researcher to make better decisions.

	Policy 1	Policy 2	Policy 3	Policy 4
Student Model 1 with Time Model 1 fit to data from Low Performing Students				
Student Model 2 with Time Model 1 fit to data from Low Performing Students				
Student Model 1 with Time Model 1 fit to data from **High Performing Students**				
...				

Table 3. Hypothetical robust evaluation matrix that incorporates both various student model classes, different generative time models of how many problems students will do in a fixed time, and models that are fit to different demographics.

CONCLUSION

We have introduced the robust evaluation matrix, a method to support off-policy policy selection. Interestingly, even though REM cannot enable the user to accurately assess the impact of a policy, it can help a researcher determine when a policy should or should not be deployed. We have shown how REM could have been used before running our own experiment to test new adaptive policies to reduce our confidence that any of the policies we were testing would do better than any other, and perhaps dissuade us from running the experiment until we found a better policy. We additionally showed how prior work [21] has indirectly provided evidence that REM could potentially be used to help gain confidence that a policy will actually improve student performance (beyond baseline policies). This could have implications to the learning at scale community as personalization is one of the most important fronts for learning at scale researchers, and as we have seen, current techniques in policy estimation and policy selection are not sufficient, even at scale. Moreover, this new method could prove promising to the reinforcement learning community, beyond its impact in the domain of education. For ourselves, we have helped turn hindsight into foresight; we hope this foresight will guide future researchers towards more rapidly discovering effective adaptive instructional policies.

ACKNOWLEDGMENTS

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130215 and R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Dept. of Education.

REFERENCES

1. Joseph Beck, Beverly Park Woolf, and Carole R Beal. 2000. ADVISOR: A machine learning architecture for intelligent tutor construction. *AAAI/IAAI* 2000 (2000), 552–557.
2. Joseph Beck and Xiaolu Xiong. 2013. Limits to accuracy: how well can we do at student modeling?. In *Educational Data Mining 2013*.
3. Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. 2011. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction* 21, 1-2 (2011), 137–180.
4. Benjamin Clement, Pierre-Yves Oudeyer, and Manuel Lopes. 2016. A Comparison of Automatic Teaching Strategies for Heterogeneous Student Populations. *International Educational Data Mining Society* (2016).
5. Albert Corbett. 2000. *Cognitive mastery learning in the ACT programming tutor*. Technical Report. AAAI Technical report, SS-00-01.
6. Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.
7. Shayan Doroudi, Kenneth Holstein, Vincent Aleven, and Emma Brunskill. 2015. Towards Understanding How to Leverage Sense-Making, Induction and Refinement, and Fluency to Improve Robust Learning. *International Educational Data Mining Society* (2015).
8. Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* (2011).
9. José P González-Brenes and Yun Huang. 2015. " Your Model Is Predictive–but Is It Useful?" Theoretical and Empirical Considerations of a New Paradigm for Adaptive Tutoring Evaluation. *International Educational Data Mining Society* (2015).
10. Assaf Hallak, COM François Schnitzler, Timothy Mann, and Shie Mannor. 2015. Off-policy Model-based Learning under Unknown Factored Dynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 711–719.
11. Nan Jiang and Lihong Li. 2015. Doubly Robust Off-policy Evaluation for Reinforcement Learning. *arXiv preprint arXiv:1511.03722* (2015).
12. Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. 2003. The expertise reversal effect. *Educational psychologist* 38, 1 (2003), 23–31.
13. Jung In Lee and Emma Brunskill. 2012. The Impact on Individualizing Student Models on Necessary Practice Opportunities. *International Educational Data Mining Society* (2012).
14. Robert V Lindsey, Mohammad Khajah, and Michael C Mozer. 2014. Automatic discovery of cognitive skills to improve the prediction of student learning. In *Advances in neural information processing systems*. 1386–1394.
15. Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. 2014. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014*

international conference on Autonomous agents and multi-agent systems. International Foundation for Autonomous Agents and Multiagent Systems, 1077–1084.

16. Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. 2007. Bias and variance approximation in value function estimates. *Management Science* 53, 2 (2007), 308–322.
17. Christopher M Mitchell, Kristy Elizabeth Boyer, and James C Lester. Evaluating State Representations for Reinforcement Learning of Turn-Taking Policies in Tutorial Dialogue. In *Proceedings of the Fourteenth Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL-2013)*. 339–343.
18. Philip I Pavlik Jr, Hao Cen, and Kenneth R Koedinger. 2009. Performance Factors Analysis–A New Alternative to Knowledge Tracing. *Online Submission* (2009).
19. Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*. 505–513.
20. Doina Precup. 2000. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* (2000), 80.
21. Anna N Rafferty, Emma Brunskill, Thomas L Griffiths, and Patrick Shafto. 2015. Faster Teaching via POMDP Planning. *Cognitive Science* (2015).
22. Martina A Rau, Vincent Aleven, and Nikol Rummel. 2013. Complementary effects of sense-making and fluency-building support for connection making: A matter of sequence?. In *International Conference on Artificial Intelligence in Education*. Springer, 329–338.
23. Joseph Rollinson and Emma Brunskill. 2015. From Predictive Models to Instructional Policies. *International Educational Data Mining Society* (2015).
24. Jonathan P Rowe and James C Lester. 2015. Improving student problem solving in narrative-centered learning environments: A modular reinforcement learning framework. In *International Conference on Artificial Intelligence in Education*. Springer, 419–428.
25. Jonathan P Rowe, Bradford W Mott, and James C Lester. 2014. Optimizing Player Experience in Interactive Narrative Planning: A Modular Reinforcement Learning Approach.. In *AIIDE*.
26. Philip S Thomas and Emma Brunskill. 2016. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. *arXiv preprint arXiv:1604.00923* (2016).
27. Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 2015. High-Confidence Off-Policy Evaluation.. In *AAAI*. 3000–3006.
28. Michael Yudelson and Steve Ritter. 2015. Small Improvements for the Model Accuracy — Big Improvements for the Student. In *International Conference on Artificial Intelligence in Education*. Springer, 903–905.
29. Li Zhou and Emma Brunskill. 2016. Latent Contextual Bandits and their Application to Personalized Recommendations for New Users. *arXiv preprint arXiv:1604.06743* (2016).