International Journal of Artificial Intelligence in Education (2019) 29:568–620
https://doi.org/10.1007/s40593-019-00187-x

ARTICLE

Where's the Reward?

A Review of Reinforcement Learning for Instructional Sequencing

Shayan Doroudi[1,2,3] · Vincent Aleven[4] · Emma Brunskill[2]

Published online: 14 November 2019

Abstract

Since the 1960s, researchers have been trying to optimize the sequencing of instructional activities using the tools of reinforcement learning (RL) and sequential decision making under uncertainty. Many researchers have realized that reinforcement learning provides a natural framework for optimal instructional sequencing given a particular model of student learning, and excitement towards this area of research is as alive now as it was over fifty years ago. But does RL actually help students learn? If so, when and where might we expect it to be most helpful? To help answer these questions, we review the variety of attempts to use RL for instructional sequencing. First, we present a historical narrative of this research area. We identify three waves of research, which gives us a sense of the various communities of researchers that have been interested in this problem and where the field is going. Second, we review all of the empirical research that has compared RL-induced instructional policies to baseline methods of sequencing. We find that over half of the studies found that RL-induced policies significantly outperform baselines. Moreover, we identify five clusters of studies with different characteristics and varying levels of success in using RL to help students learn. We find that reinforcement learning has been most successful in cases where it has been constrained with ideas and theories from cognitive psychology and the learning sciences. However, given that our theories and models are limited, we also find that it has been useful to complement this approach with running more robust offline analyses that do not rely heavily on the assumptions of one particular model. Given that many researchers are turning to deep reinforcement learning and big data to tackle instructional sequencing, we believe keeping these best practices in mind can help guide the way to the reward in using RL for instructional sequencing.

✉ Shayan Doroudi
doroudis@uci.edu

Extended author information available on the last page of the article.

Keywords Reinforcement learning · Instructional sequencing · Adaptive instruction · History of artificial intelligence in education

Introduction

In 1960, a book was published by the name of *Dynamic Programming and Markov Decision Processes* and an article by the name of "Machine-Aided Learning". The former established itself as one of the foundational early texts about Markov decision processes (MDPs), the model that underpins reinforcement learning (RL). The latter is a virtually unknown two-page vision paper suggesting that computers could help individualize the sequence of instruction for each student. Both were written by Ronald Howard, who is one of the pioneers of decision processes and is now considered the "father of decision analysis." These two lines of work are not unrelated; in 1962, Howard's doctoral student Richard Smallwood wrote his dissertation, *A Decision Structure for Teaching Machines*, on the topic of how to use decision processes to adapt instruction in a computerized teaching machine. This is perhaps the first example of using reinforcement learning (broadly conceived) for the purposes of instructional sequencing (i.e., determining how to adaptively sequence various instructional activities to help students learn). Instructional sequencing was thus one of the earliest applications of reinforcement learning.

Over 50 years later, we find that researchers continue to attempt tackling the problem of instructional sequencing with the tools of reinforcement learning in a variety of educational settings (perhaps not always realizing that this problem was first formulated and studied decades ago) and excitement for this area of research is perhaps as alive now as ever. Many researchers were drawn to this area of research, because (1) it is well known that the way in which instruction is sequenced can make a difference on how well students learn (Ritter et al. 2007), and (2) reinforcement learning provides the mathematical machinery to formally optimize the sequence of instruction (Atkinson 1972a). But over the past 50 years, how successful has RL been in discovering useful adaptive instructional policies? More importantly, looking to the future, how might RL best impact instructional sequencing?

In this paper, we seek to address these questions by reviewing the variety of attempts to apply reinforcement learning to the task of instructional sequencing in different settings. We first narrate a brief history of RL applied to instructional sequencing. We identify three waves of research in this area, with the most recent wave pointing to where the field seems to be heading in the future. Second, to assess how successful using RL for instructional sequencing has been in helping students learn, we review all of the empirical research comparing RL-induced instructional policies to baseline instructional policies. We find that over half of the studies found significant effects in favor of RL-induced policies. Moreover, we identify five clusters of studies that vary in the way RL has been used for instructional sequencing and have had varying levels of success in impacting student learning.

We find that reinforcement learning has been most successful in cases where it has been constrained with ideas and theories from cognitive psychology and the learning sciences, which suggest combining theory-driven and data-driven approaches, as opposed to purely relying on black-box data-driven algorithms. However, given that our theories and models are limited, we also find that it has been useful to complement this approach with running more robust offline analyses that do not rely heavily on the assumptions of one particular model.

Recent advances in reinforcement learning and educational technology, such as deep RL (Mnih et al. 2015) and big data, seem to be resulting in growing interest in applying RL to instructional sequencing. Our hope is that this review will productively inform both researchers who are new to the field and researchers who are continuing to explore ways to impact instructional design with the tools of reinforcement learning.

Reinforcement Learning: Towards a "Theory of Instruction"

In 1972, the psychologist Richard Atkinson wrote a paper titled "Ingredients for a Theory of Instruction" (Atkinson 1972b), in which he claims a theory of instruction requires the following four "ingredients":

"1. A model of the learning process.
2. Specification of admissible instructional actions.
3. Specification of instructional objectives
4. A measurement scale that permits costs to be assigned to each of the instructional actions and payoffs to the achievement of instructional objectives."

Atkinson further describes how these ingredients for a theory of instruction map onto the definition of a Markov decision process (MDP). Formally, a finite-horizon MDP (Howard 1960a) is defined as a five tuple (S, A, T, R, H), where

- S is a set of states
- A is a set of actions
- T is a transition function where $T(s'|s, a)$ denotes the probability of transitioning from state s to state s' after taking action a
- R is a reward function where $R(s, a)$ specifies the reward (or the probability distribution over rewards) when action a is taken in state s, and
- H is the horizon, or the number of time steps where actions are taken.

In *reinforcement learning* (RL), the goal is for an *agent* to learn a *policy* π—a mapping from states to actions or probability distributions over actions—that incurs high reward (Sutton and Barto 1998). The policy specifies for each state what action the agent should take. There exist various methods for *planning* in a MDP, such as value iteration (Bellman 1957) or policy iteration (Howard 1960a), which yield the optimal policy for the given MDP. However, RL refers to the task of learning a policy when the parameters of the MDP (the transition function and possibly the reward function) are not known ahead of time.

As Atkinson explained, in the context of instruction, the transition function maps onto a model of the learning process, where the MDP states are the states that the student can be in (such as cognitive states). The set of actions are instructional activities that can change the student's cognitive state. These activities could be problems, problem steps, flashcards, videos, worked examples, game levels in the context of an educational game, etc. Finally, the reward function can be factorized into a cost function for each instructional action (e.g., based on how long each action takes) and a reward based on the cognitive state of the student (e.g., a reward for each skill a student has learned).

We note that this review specifically focuses on applications of reinforcement learning to the sequencing of instructional activities. Reinforcement learning and decision processes have been used in other ways in educational technology that we do not consider here. For example, Barnes and Stamper (2008) have used MDPs to model students' problem solving processes and automatically generate hints for students. Similarly, Rafferty et al. (2015, 2016b) modeled student problem solving as a MDP and used problem solving trajectories to infer the MDP so they could ultimately give feedback to the students about misconceptions they might have. In these papers, the actions of the MDP are problem solving steps *taken by students* in the course of solving a problem, whereas in our paper, we focus on studies where the actions are instructional activities *taken by an RL agent* to optimize a student's learning over the course of many activities.

As we show below, the natural formulation of the instructional process as a decision process and a problem that can be tackled by reinforcement learning drew many researchers, including psychologists like Atkinson, to this problem. In theory, RL could formalize that which was previously an art: instruction. How well it can do so in practice is the subject of investigation of this paper.

Examples of RL for Instructional Sequencing

In order to situate the rest of this paper, it is worth giving some concrete examples of how the techniques of decision processes and RL could be applied to instructional sequencing. We will begin with one of the simplest possible MDPs that could be used in the context of instructional sequencing, and then consider a series of successive refinements to be able to model more authentic phenomena, ending with the model considered by Atkinson (1972b). While there are many more ways of applying RL to instructional sequencing, this section will give us a sense of one concrete way in which it has been done, as well as introduce several of the design decisions that need to be made in modeling how people learn and using such models to induce instructional policies. In the review of empirical studies below, we will discuss a much broader variety of ways in which various researchers have used RL to implement instructional sequencing.

The first model we will consider is a simple MDP that assumes for any given fact, concept, or skill to be learned (which we will refer to as a knowledge component or KC), the student can be in one of two states: the "correct" state or the "incorrect" state. Whenever the student answers a question correctly, the student will transition to the correct state for the associated KC, and whenever the student answers a question incorrectly, the student will transition to the incorrect state for that KC. The complete state can be described with a binary vector of all the individual KC states. The set of actions is the set of items that we can have students practice, where each item is associated with a given KC. For each item, there is a 2-by-2 transition matrix that specifies the probability of its associated KC transitioning from one state to another. (For simplicity, we assume that all items for the same KC have the same probability of transitioning to the correct state.) Suppose our goal is to have the student reach the correct state for as many KCs as possible. Then we can specify a reward function that gives a reward of one whenever the student transitions from the incorrect state to the correct state, a reward of negative one whenever the student transitions from the correct state to the incorrect state, and a reward of zero otherwise. In this case, the optimal instructional policy is trivial: always give an item for the KC that has the highest probability of transitioning to the correct state among all KCs in the incorrect state.

Of course to use this policy in practice, we need to learn the parameters of the MDP using prior data. Given the assumptions we made, the only parameters in this model are the transition probabilities for each KC. In this case, the maximum likelihood transition probability[1] for each KC can be inferred simply by computing how many times students transitioned from the incorrect state to the correct state divided by the number of time steps where the students received an item in the incorrect state.

However, notice that the MDP presented above is likely not very useful, because it assumes our goal is just to have students answer questions correctly. A student may be able to answer questions correctly without displaying proper understanding, for example by guessing or by answering correctly for slightly wrong reasons. In reality, we may assume that students' answers are only noisy signals of their underlying knowledge states. To model the fact that we cannot know a student's true cognitive state, we would need to use a partially observable Markov decision process (POMDP) (Sondik 1971). In a POMDP, the underlying state is inaccessible to the agent, but there is some observation function (O) which maps states to probability distributions of observations. In our example, the observation at each time step is whether the student answers a question correctly or incorrectly, and the probability of answering a question correctly or incorrectly depends on which state the student is in for the current KC that is being taught. Again, we can assume there are two states for each KC, but we will call the states the "learned" state and the "unlearned" state, as they represent whether the student has learned the KC. If we ignore the reward function, this POMDP is equivalent to the Bayesian knowledge tracing model (Corbett and Anderson 1995), which has been used to implement cognitive mastery learning in intelligent tutoring systems (Corbett 2000). Typically BKT is not considered in the RL framework, because a reward function is not *explicitly* specified, although using BKT for mastery learning does *implicitly* follow a reward function. One possible reward function for cognitive mastery learning would be that each time our estimated probability that the student has learned a particular KC exceeds 0.95, then we receive a reward of one, and otherwise we receive a reward of zero. Such a model would then keep giving items for a given KC, until we are 95% confident that the student has learned that KC before moving on. Notice that the optimal policy under this reward

[1] Maximum likeilihood parameters are the parameters that maximize the likelihood of sampling the data that was previously collected.

function (i.e., cognitive mastery learning) can be very different from the optimal policy under other reasonable reward functions (e.g., get a reward of one for each KC that is *actually* in the learned state, which we cannot directly observe).

The parameters of a POMDP like the BKT model are slightly more difficult to infer, because we do not actually know when students are in each state, unlike in the MDP case. However, there are a number of algorithms that could be used to estimate POMDP parameters including expectation maximization (Welch 2003), spectral learning approaches (Hsu et al. 2012; Falakmasir et al. 2013), or simply performing a brute-force grid search over the entire space of parameters (Baker et al. 2010).

We consider one final modification to the model above, namely that which was used by Atkinson (1972b) for teaching German vocabulary words. Note that the BKT model does not account for forgetting. Atkinson (1972b) proposed a POMDP with three states for each word to be learned (or KC, in the general case): an unlearned state, a temporarily learned state, and a permanently learned state. The model allows for some probability of transitioning from either the unlearned or temporarily learned states to the permanently learned state, but one can also transition from the temporarily learned state back to the unlearned state (i.e., forgetting). Moreover, this model assumes that a student will always answer an item correctly unless the student is in the unlearned state, in which case the student will always answer items incorrectly. The reward function in this case gives a reward of one for each word that is permanently learned at the end (as measured via a delayed posttest, where it is assumed that any temporarily learned word will be forgotten). The optimal policy in this case can be difficult to compute because one needs to reason about words that are forgotten over time. Therefore, Atkinson (1972b) used a myopic policy that chooses the best next action as though only one more action will be taken. In this case, the best action is to choose the word that has the highest probability of transitioning to the permanently learned state.

Design Considerations in Reinforcement Learning

Before continuing, it is worthwhile to describe several different settings that are considered in reinforcement learning, and the design considerations that researchers need to make in considering how to apply RL. RL methods are often divided into *model-based* and *model-free* approaches. Model-based RL methods learn the model (transition function and reward function) first and then use MDP planning methods to induce a policy. Model-free methods use data to learn a good policy directly without learning a model first. Most of the studies we review in this paper have used model-based RL. All of the examples described above are model-based—a model is fit to data first and then a policy (either the optimal policy or a myopic policy) is derived using MDP/POMDP planning.

There are two different ways in which RL can be used. In *online* RL, the policy is learned and improved as the agent interacts with the environment. In *offline* RL, a policy is learned on data collected in the past, and is then used in an actual environment. For instance, in the examples we presented above, the models were fit to previously collected data in an offline fashion, which was then used to do model-based RL. While online RL can lead to more quickly and efficiently identifying a good policy, it can be more difficult to use in practice as one must determine and fix the algorithms used before collecting any data.

In online RL, the agent must decide whether to use the current best policy in order to accrue a high reward or to make decisions which it is uncertain about with the hopes of finding a better policy in the future. This is know as the *exploration vs. exploitation trade-off*. Exploration refers to trying new actions to gather data from less known areas of the state and action space, while exploitation refers to using the best policy the agent has identified so far. This trade-off is rarely tackled in the studies we consider below that have applied RL to instructional sequencing, with a few exceptions (Lindsey et al. 2013; Clement et al. 2015; Segal et al. 2018).

As discussed in our examples, since the cognitive state of a student usually cannot be observed, it is common to use a partially observable Markov decision process rather than a (fully observable) MDP. Planning, let alone reinforcement learning, in POMDPs is, in general, intractable, which is why researchers often use approximate methods for planning, such as myopic planning. However, some models of learning (such as the BKT model discussed above) are very restricted POMDPs, making it possible to find an optimal policy.

In model-based RL, our model is generally incorrect, not only because there is not enough data to fit the parameters correctly, but also because the form of the model could be incorrect. As we will see, researchers have proposed various models for student learning, which make rather different assumptions. When the assumptions of the model are not met, we could learn a policy that is not as good as it seems. To mitigate this issue, researchers have considered various methods of *off-policy policy evaluation*, or evaluating a policy offline using data from one or more other policies. Off-policy policy evaluation is important in the context of instructional sequencing, because it would be useful to know how much an instructional policy will help students before testing it on actual students. Ultimately, a policy must be tested on actual students in order to know how well it will do, but blindly testing policies in the real world could be costly and potentially a waste of student time.

From the intelligent tutoring systems literature, we can distinguish between two broad forms of instructional sequencing in terms of the granularity of the instructional activities: *task-loop* (or outer loop) adaptivity and *step-loop* (or inner loop) adaptivity (Vanlehn 2006; VanLehn 2016; Aleven et al. 2016a). In task-loop adaptivity, the RL agent must select distinct tasks or instructional activities. In step-loop adaptivity, the RL agent must choose the exact nature of each step (e.g., how much instructional scaffolding to provide) for a fixed instructional task. For example, an RL agent operating in the step loop might have to decide for all the steps in a problem whether to show the student the solution to the next step or whether to ask the student to solve the next step (Chi et al. 2009). Almost all of the papers we include in this review operate in the task loop. While step-loop adaptivity is a major area of research in adaptive instruction in general (Aleven et al. 2016a), relatively little work has been pursued in this area using RL-based approaches.

A Historical Perspective

The use of reinforcement learning (broadly conceived) for instructional sequencing dates back to the 1960s. We believe at least four factors led to interest in automated instructional sequencing during the 60s and 70s. First, teaching machines (mechanical devices that deliver step-by-step instruction via exercises with feedback) were gaining a lot of interest in the late 50s and 60s, and researchers were interested in implementing adaptive instruction in teaching machines (Lumsdaine 1959). Second, with the development of computers, the field of computer-assisted instruction (CAI) was forming and there was interest in developing computerized teaching machines (Liu 1960). Third, pioneering work on mathematical optimization and dynamic programming (Bellman 1957; Howard 1960a), particularly the development of Markov decision processes, provided a mathematical literature for studying the optimization of instructional sequencing. Finally, the field of mathematical psychology was beginning to formulate mathematical models of learning (Atkinson and Calfee 1963).

As mentioned earlier, Ronald Howard, one of the pioneers of Markov decision processes, was interested in using decision processes to personalize instruction (Howard 1960b). In 1962, Howard's PhD student, Richard Smallwood, wrote his dissertation, *A Decision Structure for Teaching Machines* (Smallwood 1962), which presented what is to our knowledge the first time an RL-induced instructional policy was tested on actual students. Even though the field of reinforcement learning had not yet developed, Smallwood was particularly interested in what we now call online reinforcement learning, where the system could improve over time as it interacts with more students. In fact, he provided preliminary evidence in his dissertation that the policy developed for his computerized teaching machine did in fact change with the accumulation of more data. Smallwood's PhD student Edward Sondik's dissertation, *The Optimal Control of Partially Observable Markov Decision Processes*, was seemingly the first text that formally studied planning in partially observable Markov decision processes (POMDPs). Sondik wrote in his dissertation, "The results obtained by Smallwood [on the special case of determining optimum teaching strategies] prompted this research into the general problem" (Sondik 1971). Thus, the analysis of POMDPs, an important area of research in optimal control, artificial intelligence, and reinforcement learning, was prompted by its application to instructional sequencing.

Around the same time, a group of mathematical psychologists at Stanford, including Richard Atkinson and Patrick Suppes, were developing models of learning from a psychological perspective and were interested in optimizing instruction according to these models, using the tools of dynamic programming developed by Howard and his colleagues. Atkinson and his students tested several instructional policies that optimized various models of learning (Dear et al. 1967; Laubsch 1969; Atkinson and Lorton 1969; Atkinson 1972b; Chiang 1974).

Curiously, there is almost no work on deriving optimal policies from the mid-70s to the early 2000s. While we cannot definitively say why, there seem to be a number of contributing factors. Researchers from the mathematical optimization community (including Howard and his students) stopped working on this problem after a few years and continued to work in their home disciplines. On the other hand, Atkinson's career in psychology research ended in 1975 when he left for the National Science Foundation (Atkinson 2014), and presumably the field of mathematical psychology lost interest in optimizing instructional policies over time. Research in automated instructional sequencing re-emerged at the turn of the twenty-first century for seemingly three reasons that completely parallel the trends that existed in the 60s. First, there was growing interest in intelligent tutoring systems, a natural testbed for adaptive instructional policies, paralleling the interest in teaching machines and computer-assisted instruction in the 60s. Second, the field of reinforcement learning formally formed in the late 1980s and early 1990s (Sutton and Barto 1998), combining machine learning with the tools of Markov decision processes and dynamic programming built in the 60s. Finally, the field of Artificial Intelligence in Education (AIED) and, later, educational data mining (EDM) were interested in developing statistical models of learning, paralleling mathematical psychologists' interest in models of learning several decades earlier.

Even though there has been no void of research on instructional sequencing since the early 2000s, there seems to be a third wave of research appearing in this area in recent years. This is due to certain shifting trends in the research landscape that might be attracting a new set of researchers to the problem of data-driven instructional sequencing. First, there is a new "automated" medium of instruction, like the teaching machines, CAI, and ITSs of previous decades: MOOCs and other large-scale online education providers.[2] And with MOOCs comes the promise of big data. Second, the field of deep reinforcement learning has formed, leading to significantly more interest in the promise of reinforcement learning as a field. Indeed, there were around 35% more papers and books mentioning reinforcement learning in 2017 than in 2016 (as per the number of Google Scholar search hits). While initial advances in deep reinforcement learning have been focused largely on playing games such as Atari (Mnih et al. 2015) and Go (Silver at el. 2016, 2017), we have recently seen researchers applying deep reinforcement learning to the problem of instructional sequencing (Piech et al. 2015; Chaplot et al. 2016; Reddy et al. 2017; Wang et al. 2017a; Upadhyay et al. 2018; Shen et al. 2018a). Finally, in tandem with the use of deep reinforcement learning, there is also a growing movement within the AIED and EDM communities to use deep machine learning models to model human learning (Piech et al. 2015; Chaplot et al. 2016); this is a significantly different approach from the previous trends to use models that were more interpretable in the 1990s and models that were more driven by psychological principles in the 1960s.

Table 1 summarizes the trends that we believe have been responsible for the "three waves" of interest in applying reinforcement learning and decision processes to instructional sequencing. We find that there is a general trend that the methods of instructional sequencing have become more data-driven over time and the

[2] Although many researchers are still testing RL-induced policies in ITSs and other platforms, there is reason to believe that MOOCs and other online instructional platforms such as Khan Academy and Duolingo have attracted many new researchers to this area. This can be witnessed by the emergence of Learning@Scale as a new conference that emerged as a result of MOOCs. Indeed, one of us (Doroudi) was drawn to AIED as a result of the development of MOOCs.

Table 1 Trends in the three waves of interest in applying reinforcement learning to instructional sequencing

	First wave (1960s–1970s)	Second wave (1990s–2010s)	Third wave (2010s–)
Instructional technology	Teaching machines/CAI	ITSs	MOOCs
Optimization methods	MDP planning	RL	Deep RL
Models of learning	Mathematical psychology	EDM/AIED	Deep learning

The "Instructional Tecnology" row shows technologies that were being developed or saw a lot of hype in the associated time period, even though older technologies were still used during the later time periods. The "Optimization Methods" row shows the form that RL research took in each time period; notice that the field of "reinforcement learning" was formally introduced in the late 1980s, but earlier work in MDP planning used with data-driven models in the 60s would still be considered RL. The "Models of Learning" row shows the research communities where new types of models of learning were emerging from during each time period

media for delivering instruction have become generally more data-generating. Perhaps researchers are inclined to believe that more computational power, more data, and better reinforcement learning algorithms makes this a time where RL can have a demonstrable impact on instruction. However, we do not think these factors are sufficient for RL to leave its mark; we believe there are insights to gain about how RL can be impactful from the literature, which is where we will look to next. Based on the growth of interest in reinforcement learning in general and deep reinforcement learning in particular, we anticipate many more researchers will be interested in tackling instructional sequencing in the coming years. We hope this history and the review of empirical literature that follows will be informative to these researchers.

Review of Empirical Studies

To understand how successful RL has been in impacting instructional sequencing, we conduct a broad review of the empirical literature in this area. In particular we are interested in any studies that run a controlled experiment comparing one or more instructional policies, at least one of which is induced by an RL-based approach. We are interested in seeing how often studies find a significant difference between RL-induced policies and baseline policies, and what factors might affect whether or not an RL-induced policy is successful in helping students learn beyond a baseline policy. This review of the empirical literature was prompted by two experiments on RL-induced instructional sequencing that we ran in a fractions intelligent tutoring system. Both experiments resulted in no significant differences among the policies we tested. We were interested in identifying reasons why our experiments led to null results, and how our findings compared to other studies that tested RL-induced policies. Our own experiments have also given us insights into the challenges of applying RL to instructional sequencing, which have informed the discussion following the literature review below; details from one of our experiments along with some of the insights it provided are given in Appendix B.

Inclusion Criteria: Scope of the Review

One challenge of conducting this systematic review is determining what counts as an "RL-induced policy." First of all, not all studies (especially ones from the 60s and 70s) use the term reinforcement learning, but they are clearly doing some form of RL or at least applying Markov decision processes to the task of instructional sequencing. Second, some studies are not clearly conducting some form of RL, but still have the "flavor" of using RL in that they find instructional policies in a data-driven way or they use related techniques such as multi-armed bandits (Gittins 1979; Auer et al. 2002) or Bayesian optimization (Mockus 1994; Brochu et al. 2010). On the other hand, some studies that *do* use the language of RL rely on heuristics or approximations in trying to find an instructional policy (such as myopic planning). We included all studies that had the "flavor" of using RL-induced instructional policies, even when the language of RL or related optimization techniques were not used.

There are two components to reinforcement learning: (1) optimization (e.g., MDP planning in the model-based setting) and (2) learning from data (e.g., learning the MDP in the model-based setting). For a study to be considered as using RL for instructional sequencing, it should use some form of optimization and data to find instructional policies. More formally, we included any studies where:

- The study acknowledges (at least implicitly) that there is a model governing student learning and giving different instructional actions to a student might probabilistically change the state of a student according to the model.
- There is an instructional policy that maps past observations from a student (e.g., responses to questions) to instructional actions.
- Data collected from students (e.g., correct or incorrect responses to previous questions), either in the past (offline) or over the course of the study (online), are used to learn either:
    - a statistical model of student learning, and/or
    - an instructional policy.
- If a statistical model of student learning is fit to data, the instructional policy is designed to approximately optimize that model according to some reward function, which may be implicitly specified.

Notice that this means we consider any studies that might learn a model from prior data and then use a heuristic to find the instructional policy (such as myopic planning rather than long-horizon planning). This also means we *did not* include any studies that applied planning to a pre-specified MDP or POMDP (e.g., a BKT model with hand-set parameters), since learning is a critical component of reinforcement *learning*.

Searching for all papers that match our inclusion criteria is challenging as not all papers use the same language to discuss data-driven instructional sequencing. Therefore, to conduct our search, we began with an initial set of papers that we knew matched our inclusion criteria in addition to any papers that we became aware of over time. We iteratively added more papers by performing one-step forward and backward citation tracing on the growing pool of papers. That is, for every paper that we included in our review, we looked through the papers that it cited as well all papers that cited it—as identified by Google Scholar as of December 2018—to see if any of those papers also matched our inclusion criteria. This means if we have missed any relevant studies, they are disconnected (in terms of direct citations) from the studies that we have identified. We found relevant papers coming from a diversity of different research communities including mathematical psychology, cognitive science, optimal control, AIED, educational data mining, machine learning, machine teaching, and human-robot interaction.

Results

We found 34 papers containing 41 studies that matched our criteria, including a previously unpublished study that we ran on our fractions tutoring system, which is described in Appendix B. Before discussing these studies in depth, we briefly mention the kinds of papers that did not match our inclusion criteria, but are still related to studying RL for instructional sequencing. Among these papers, we found 19 papers that learned policies on offline data but did not evaluate the performance of these policies on actual students.[3] At least an additional 23 papers learned (and compared) policies using only simulated data (i.e., no data from real learners were used).[4] At least eight papers simply proposed using RL for instructional sequencing or proposed an algorithm for doing so in a particular setting without using simulated or real data.[5] We also found at least fourteen papers that did study instructional policies with real students, but did not match our inclusion criteria for various reasons, including not being experimental, varying more than just the instructional policy across conditions, or using hand-set model parameters.[6] For example, Corbett and Anderson (1995) compare using a model (BKT) to determine how many remediation exercises should

[3] These papers include: Chi et al. (2008), Theocharous et al. (2010), Mitchell et al. (2013a), Mitchell et al. (2013b), Mota et al. (2015), Piech et al. (2015), Rollinson and Brunskill (2015), Lan and Baraniuk (2016), Hoiles and Schaar (2016), Antonova et al. (2016), Käser et al. (2016), Chaplot et al. (2016), Lin and Chi (2016), Shen and Chi (2016a), Wang et al. (2016), Wang et al. (2017a), Sawyer et al. (2017), and Tabibian et al. (2017), and Fenza et al. (2017).

[4] These papers include: Chant and Atkinson (1973), Iglesias et al. (2003), Martin and Arroyo (2004), Sarma and Ravindran (2007), Iglesias et al. (2009), Theocharous et al. (2009), Folsom-Kovarik et al. (2010), Kujala et al. (2010), Champaign and Cohen (2010), Malpani et al. (2011), Pietquin et al. (2011), Daubigney et al. (2013), Dorça et al. (2013), Schatten et al. (2014), Andersen et al. (2016), Clement et al. (2016), Reddy et al. (2017), Goel et al. (2017), Wang et al. (2017b), and Zaidi et al. (2017), and Mu et al. (2018), Upadhyay et al. (2018), and Lakhani (2018).

[5] These papers include: Beck (1997), Bennane et al. (2002), Legaspi and Sison (2002), Almond (2007), Brunskill and Russell (2011), Ramachandran and Scassellati (2014), and Mejía-Lavalle et al. (2016), and Spaulding and Breazeal (2017).

[6] These papers include: Smallwood (1962), Corbett and Anderson (1995), Joseph et al. (2004), Iglesias et al. (2006), Pavlik et al. (2008), Van Rijn et al. (2009), Nijboer (2011), Folsom-Kovarik (2012), Lomas et al. (2012), Wang (2014), Leyzberg et al. (2014), Settles and Meeder (2016), and Sense (2017), and Hunziker et al. (2018).

be given for each KC, but they compare this to providing no remediation, not another way of sequencing remediation exercises. Finally, many papers have mathematically studied deriving optimal instructional policies for various models of learning, especially during the first wave of optimizing instructional sequencing (e.g. Karush and Dear, 1967; Smallwood, 1968, 1971). The sheer number of papers that study RL-induced policies in one form or another shows that there is broad interest in applying RL to instructional sequencing, especially as these papers come from a variety of different research communities.

For the studies that met our inclusion criteria, the first row of Table 2 shows how the studies are divided in terms of varying "levels of significance." Twenty-one of the 36 studies found that at least one RL-induced policy was statistically significantly better than all baseline policies for some outcome variable, which is typically performance on a posttest or time to mastery. Four studies found no significant difference overall, but a significant aptitude-treatment interaction (ATI) favoring low-performing students (i.e., finding that the RL-induced policy performed significantly better than the baselines for lower performing students but no significant difference was detected for high performing students). Four studies found mixed results, namely that an RL-induced instructional policy outperformed at least one baseline policy but not all the baselines. Ten studies found no significant difference between adaptive policies and baseline policies. Only one study found a baseline policy outperformed an RL-induced policy.[7]

Thus, over half of the studies found that adaptive policies outperform all baselines that were tested. Moreover, the studies that found a significant difference, as well as those that demonstrated an aptitude-treatment interaction, often found a Cohen's *d* effect size of at least 0.8, which is regarded as a large effect (Cohen 1988). While this is a positive finding in favor of using RL-induced policies, it does not tell us *why* some studies were successful in showing that RL-induced policies can help students learn beyond a baseline policy, and why others were less successful. To do so, we qualitatively cluster the studies into five different groups based on how they have applied RL. The clusters generally vary in terms of the types of instructional actions considered and how they relate to each other. In paired-associate learning tasks, each action specifies the content presented to the student, but each piece of content is assumed to be independent of the rest. In the concept learning tasks cluster, actions are interrelated insofar as they give different bits of information about a particular concept. In the sequencing interdependent content cluster, the various pieces of content are assumed to be interdependent, but not in the restricted form present in concept learning tasks. In the sequencing activity types cluster, the order of content is fixed, and each potential action specifies the type of instructional activity for the fixed content. In all of these studies, the goal is to maximize how much students learn or how quickly they learn a prespecified set of material. The final cluster contains two studies that maximize objectives other than learning gains/speed.

[7] As we discuss later, the "baseline" in this case was actually an adaptive policy that the authors developed—Adaptive Response-Time-based sequencing (ARTS) (Mettler et al. 2011). However, since the ARTS model is not fit to data, it does not satisfy our criteria for an RL-induced adaptive policy, whereas the policy they compared to was based on the model from Atkinson (1972b) and was fit to data.

Table 2 Comparison of clusters of studies based on the “significance level” of the studies in each cluster: **Sig** indicates that at least one RL-induced policy significantly outperformed all baseline policies, **ATI** indicates an aptitude-treatment interaction, **Mixed** indicates the RL-induced policy significantly outperformed some but not all baselines, **Not sig** indicates that there were no significant differences between policies, **Sig worse** indicates that the RL-induced policy was significantly worse than the baseline policy (which for the only such case was an adaptive policy)

	Sig	ATI	Mixed	Not Sig	Sig Worse
All Studies	21	4	4	11	1
Paired-associate learning tasks	11	0	0	2	1
Concept learning tasks	4	0	2	1	0
Sequencing interdependent content	0	0	2	6	0
Sequencing activity types	4	4	0	2	0
Maximizing other objectives	2	0	0	0	0

There are many other ways in which we could have chosen to cluster the studies, including distinctions between the types of RL algorithms used (e.g., model-based vs. model-free, online vs. offline, MDP vs. POMDP), the form of instructional media that content was delivered in (e.g., CAI vs. ITSs vs. online platforms vs. educational games), and the types of baseline policies used. We chose to cluster studies based on the types of instructional actions, because we found the type of MDP or POMDP that underlies each of these clusters differs drastically from one another. In paired-associate learning tasks, the transition dynamics of the MDP can be factored into separate dynamics for each piece of content, and the key consideration becomes how people learn and forget individual pieces of content over time. If we assume content is interdependent, then the dynamics must capture the dependencies between the pieces of content. If we are trying to sequence activity types, then the dynamics must capture some relationship between activity types and what the student knows. Moreover, as we completed this literature review, it became clear that these differences play a role in the difficulty of sequencing instruction—and relatedly, the empirical success of applying RL to instructional sequencing. Table 2 shows for each cluster, the number of studies in each “significance level” that we identified above (e.g., whether the study showed a significant effect in favor of RL-induced policies, an aptitude-treatment interaction etc.). The table clearly shows that the different types of studies had very varying levels of success. Therefore, a qualitative understanding of each cluster will help us understand when and where RL can be most useful for instructional sequencing.

In what follows we describe the five clusters in more depth. For each cluster, we provide a table that gives a summary of all of the studies in that cluster and we describe some of the key commonalities and differences among the studies. In doing so, we will (a) demonstrate the variety of ways in which RL can be used to sequence instructional activities, and (b) set the stage for obtaining a better understanding of the conditions under which RL has been successful in sequencing instruction for students, which we discuss in the next section. Appendix A—including Tables 8 and

9—provides more technical details about the particulars of all the studies, including the types of models and instructional policies used in these studies.

Paired-Associate Learning Tasks

The studies in this cluster are listed in Table 3. All of the studies that were run in the first wave of instructional sequencing (1960s-70s) belong to this cluster. A paired-associate learning task is one where the student must learn a set of pairs of associations, such as learning a set of vocabulary words in a foreign language. In such tasks, a stimulus (e.g., a foreign word) is presented to the student, and the student must attempt to provide the translation of the word. The student will then see the correct translation. Such tasks may also be referred to as flashcard learning tasks, because the student is essentially reviewing a set of words or concepts using "flashcards." A key assumption in any paired-associate learning task is that the stimuli are

Table 3 Summary of all empirical studies in the paired-associate learning tasks cluster

Paper(s)	Domain	Population	Setting	Baseline Policies	Num of Subjects	Effect
Laubsch (1969)	Swahili-English	Uni	Lab	R	24	Sig
Atkinson and Lorton (1969)	English Spelling	Gr 4-6	Lab	R/S	42	Sig
Atkinson (1972b)	German-English	Uni	Lab	R	30	Sig
Chiang (1974) Exp 1	Chinese-English	Uni	Lab	R	12	Sig
Chiang (1974) Exp 2	Chinese-English	Uni	Lab	RL^+	12	Sig
Katsikopoulos et al. (2001) Exp 1	String-Num Mapping	Adults	Lab	R	16	Sig
Pavlik and Anderson (2008)	Japanese Words	Adults	Lab	RL^+/H	20	Sig
Lindsey et al. (2014)	English-Spanish	Gr 8	Class	H	57	Sig
Lindsey (2014)	English-Spanish	Gr 8	Class	R/H	56	Sig
Papoušek et al. (2016)	Geography	Online	Online	R	≈5000	Sig
Leyzberg et al. (2018)	Spanish-English	Gr 1	Lab	R	9	Sig
Dear et al. (1967)	Num-Num Mapping	Uni	Lab	R	40	Not sig
Katsikopoulos et al. (2001) Exp 2	String-Num Mapping	Adults	Lab	R	12	Not sig
Mettler et al. (2011)	African Countries	Uni	Class	H	50	Sig worse

The baseline policies column denotes whether each baseline policy involves random sequencing (R), a heuristic/expert-designed policy (H), student choice (S), an RL policy that was demonstrated to be effective in another study (RL^+), or an RL policy that was specificially designed to be ineffective by minimizing rewards (RL^-). More details on the baseline policies used are provided in Tables 8 and 9 and Appendix A.2.3. The number of subjects column reports the number of subjects in the condition with the least subjects independent of one another. For example, if one learns how to say “chair” in Spanish, it is assumed that it does not help or hinder one’s ability to learn how to say “table” in Spanish.[8] Because of this assumption, we may also think of this cluster as “sequencing independent content,” which clearly contrasts it with some of the later clusters.

The key goal in sequencing instruction for paired-associate learning tasks is to balance between (1) teaching stimuli that the student may have not learned yet, and (2) reviewing stimuli that the student may have forgotten or be on the verge of forgetting. The psychology literature has shown that sequencing instruction in such tasks is important due to the presence of the spacing effect (Ebbinghaus 1885), whereby repetitions of an item or flashcard should be spaced apart in time. Thus, a key component of many of the instructional policies developed for paired-associate learning tasks is using a model of forgetting to predict the optimal amount of spacing for each item. Early models used to sequence instruction such as the One-Element Model (OEM) ignored forgetting and only sequenced items based on predictions of whether students had learned the items or not (Bower 1961; Dear et al. 1967). Atkinson (1972b) later developed the Markov model we described in Section “Examples of RL for Instructional Sequencing”, which accounted for forgetting, and he showed that it could be used to successfully sequence words in a German to English word translation task. More recently, researchers have developed more sophisticated psychological models that account for forgetting such as the Adaptive Control of Thought—Rational (ACT-R) model (Anderson 1993; Pavlik and Anderson 2008), the Adaptive Response Time based Sequencing (ARTS) model (Mettler et al. 2011), and the DASH model (Lindsey et al. 2014). See Appendix A.1 for a brief description of these models. In some of the studies, policies using these more sophisticated models were shown to outperform RL-induced policies that used Atkinson’s original memory model (Pavlik and Anderson 2008; Mettler et al. 2011).

Thus, aside from the type of task itself, a key feature of the studies in this cluster is their use of statistical psychological models of human learning. As shown, in Table 2, in 11 out of 14 studies, RL-induced policies outperformed baseline policies. In the two studies where there were no significant differences between policies (Dear et al. 1967; Katsikopoulos et al. 2001), the model that was used was the OEM model—a simple model that does not account for forgetting and hence cannot space instruction of paired-associates over time. Similarly, Laubsch (1969) compared two RL-induced policies to a random baseline policy, and found that the policy using the OEM model did not do significantly better than the baseline while the policy based on the more sophisticated Random-Trial Increment (RTI) model did better. Finally, the only study that showed a baseline policy significantly outperformed an RL-induced policy, was the comparison of a policy based on the ARTS model with a policy based on the model used by Atkinson (1972b). The ARTS model was actually a more sophisticated psychological model than Atkinson’s, but the parameters of the model were

[8] Of course, this assumption will almost certainly not hold when learning words that might share the same root, but may still be a reasonable approximation in some cases.

not learned from data, and therefore, we considered the policy based on ARTS to technically be a non-RL-induced "baseline" policy.

Concept Learning Tasks

Concept learning is another type of task where several researchers have applied RL-based instructional sequencing. The studies in this cluster are shown in Table 4. Concept learning tasks are typically artificially-designed tasks that can be used to study various aspects of human cognition, and as such are commonly studied in the cognitive science literature. In a concept learning task, a student is presented with examples that either belong or do not belong to an initially unknown concept, and the goal is to learn what constitutes that concept (i.e., how to distinguish between positive examples that fit the concept and negative examples that do not). For example, (Rafferty et al. 2016a) used a POMDP to sequence instructional activities for two types of concept tasks, one of which is called the Number Game (Tenenbaum 2000), where students see numbers that either belong or do not belong to some category of numbers such as multiples of seven or numbers between 64 and 83. While such tasks are of little direct educational value, the authors' goal was to show that models of memory and concept learning from cognitive psychology could be combined with a POMDP framework to teach people concepts quickly, which they succeeded in doing. Whitehill and Movellan (2017) extended this idea to teaching a concept learning task that is more authentic: learning foreign language vocabulary words via images. Whitehill and Movellan (2017) call this a "Rosetta Stone" language learning task, as it was inspired by the way the popular language learning software, Rosetta Stone, teaches foreign language words via images. Notice that this task differs from teaching vocabulary as a paired-associate learning task, because there are multiple images that might convey the meaning of a foreign word (i.e., the concept), and the the goal is to find a policy that can determine at any given time both what foreign vocabulary word to teach and what image to present to convey the meaning of that word. Sen et al. (2018) also used instructional policies in an educationally-relevant concept learning task, namely perceptual fluency in identifying if two chemical molecules shown in different representations are the same. In this task, the student must learn features of the representations that help identify which chemical molecule is being shown.

Unlike paired-associate learning tasks, the various pieces of content that can be presented in a concept learning task are mutually interdependent, but in a very particular way. That is, seeing different (positive or negative) examples for a concept help refine one's idea of a concept over time.[9] For example, in the Number Game, knowing that 3, 7, and 11 are in a concept might lead one to think the concept is likely odd numbers, while also knowing that 9 is not a member of the concept might lead

[9] Despite this difference between concept learning tasks and paired-associate learning tasks, one of the concept learning tasks used by Rafferty et al. (2011, 2016a), Alphabet Arithmetic, is actually quite similar to paired-associate learning tasks, in that the goal is for students to learn a mapping of letters (A-F) to numbers (1-6), which is the concept to be learned. What distinguishes the task from paired-associate learning tasks is that the examples shown to the learner aren't mappings but rather arithmetic equations (e.g., A + B = 5). Therefore, each example gives some information about two letter-number mappings.

Table 4 Summary of all empirical studies in the concept learning tasks cluster. Details are as described in the caption of Table 3

Paper(s)	Domain	Population	Setting	Baseline Policies	Num of Subjects	Effect
(Rafferty et al. 2011)	Alphabet Arithmetic	Online	Online	R	20	Sig
(Rafferty et al. 2016a) Exp 1	Alphabet Arithmetic	Uni	Lab	R	20	Sig
(Rafferty et al. 2016a) Exp 2	Number Game	Uni	Lab	R	20	Sig
(Sen et al. 2018)	Chemical Molecules	AMT	Online	H/R	100	Sig
(Lindsey et al. 2013)	Concept Learning	AMT	Online	H	50	Mixed
(Whitehill and Movellan 2017)[a]	Picture-Word Mapping	AMT	Online	R/H	26	Mixed
(Geana 2015)	Concept Learning	Uni	Lab	RL^-	25	Not sig

[a] This experiment was reported earlier in a dissertation (Whitehill 2012)

one to believing the concept is likely prime numbers. The exact sequence of examples presented can have a large influence on a student's guess as to what the correct concept might be. Therefore, determining the exact sequence of examples to present is critical for how to most quickly teach a given concept. Moreover, in these tasks, it is often beneficial to use information-gathering activities (e.g., giving a quiz to test what concept the student finds most likely), to determine what examples the student needs to refine their understanding.

As with the paired-associate learning task studies, one common feature among the studies in this cluster is that they have typically used psychologically inspired models of learning coming from the concept learning literature and computational cognitive science literature. For example, Rafferty et al. (2016a) considered three different psychological models of human learning of varying degrees of complexity. The simplest of these models—based on a model from the concept learning literature (Restle 1962)—assumes that students have a (known) prior distribution over concepts and at any given time they posit a concept as the correct one. When presented with an example, they change their concept to be consistent with the example presented, picking a random concept with probability proportional to their prior. In more complex models, students might have some memory of previous examples shown or might maintain a distribution over concepts at any given time. While the dynamics of such models are mostly prespecified by the structure of the model, there are certain model parameters (e.g., the probability of answering a question accurately according to one's concept) that could be fit to data, as done by Rafferty et al. (2016a).

As seen in Table 4, the majority of studies in this cluster have been successful in showing that RL-induced policies outperformed some baselines. The studies in this cluster often had several baseline policies, including decent heuristic policies, so they set a higher bar for RL-induced policies. This could explain why two studies found mixed results where RL-induced policies outperformed some but not all baselines. Moreover, Rafferty et al. (2016a) compared the same policies on multiple concept learning tasks, and while their POMDP policies were generally better than the random baseline policies, there was no one POMDP policy that outperformed baseline policies *for all* concept learning tasks. This study indicates that even though RL-induced policies may be effective, the same model may not be optimal for all tasks.

Sequencing Interdependent Content

This cluster focuses on sequencing content, under the assumption that different areas of content are interdependent. The studies in this cluster are shown in Table 5. The sequencing task here is closest to traditional "curriculum sequencing," or ordering various content areas for a given topic. However, unlike traditional curriculum sequencing, the ordering of content can be personalized and adaptive, for example based on how well students have mastered various pieces of content. While concept learning tasks also have interdependent content, the goal in concept learning tasks is to teach a single underlying concept. In this cluster, the goal is to teach a broader scope of content under the assumption that how the content is sequenced affects students ability to learn future content. An instructional policy in this context must implicitly answer questions like the following: When teaching students how to make a fraction from the number line, when should we move on to the next topic and what should that topic be? Should the next topic depend on how well the student answered questions about the number line? If the student is struggling with the next topic, should we go back and teach some prerequisites that the student might have missed? When should we review a content area that we have given the student previously?

For these studies, typically a network specifying the relationship between different content areas or KCs (such as a prerequisite graph) must either be prespecified or automatically inferred from data. Appendix B describes one of our studies performed in a fractions tutoring system where the relationships between different KCs were automatically inferred from data. As we see from Table 2, the studies in this cluster have been the least successful, with all of them resulting in either a mixed result or

Table 5 Summary of all empirical studies in the sequencing interdependent content cluster. Details are as described in the caption of Table 3

Paper(s)	Domain	Population	Setting	Baseline Policies	Num of Subjects	Effect
Green et al. (2011) Exp 1	Finite Field Arithmetic	Uni	Lab	R/H	26	Mixed
Green et al. (2011) Exp 2	Artificial Language	Uni	Lab	R/H	5	Mixed
Clement et al. (2015)	Arithmetic (with Coins)	7-8yo	Class	H	133	Not sig
David et al. (2016)	Basic Math	K-12	Class	H	35	Not sig
Schatten (2017)	Basic Math	K-12	Class	H	49	Not sig
Doroudi et al. (2017a)	Fractions	Gr 4-5	Class	H	69	Not sig
Appendix B	Fractions	Gr 4-5	Class	H	100	Not sig
Segal et al. (2018)	Math	Gr 7	Class	H	9	Not sig

no significant difference between policies. We analyze why this might be in the next section.

Sequencing Activity Types

While the previous three clusters of studies were based on the way various pieces of content did or did not depend on each other—this cluster is about how to sequence the types of activities students engage with rather than the content itself. The studies in the sequencing activity types cluster are shown in Table 6. These studies used RL to determine what activity type to give at any given time for a fixed piece of content, based on the content being taught and the work that the student has done so far. For example, Shen and Chi (2016b), Zhou et al. (2017), and Shen et al. (2018a), and Shen et al. (2018b) all consider how to sequence worked examples and problem solving tasks. Similarly, Chi et al. (2009, 2010a) consider, for each step, whether the student should be told the solution or whether the student should be asked to provide the solution, and, in either case, whether the student should be asked to justify the solution. Notice that Chi et al. (2009, 2010a) consider using RL for step-loop adaptivity as opposed to task-loop adaptivity, which all of the other studies reported in this review consider.

Table 6 Summary of all empirical studies in the sequencing activity types cluster. Details are as described in the caption of Table 3

Paper(s)	Domain	Population	Setting	Baseline Policies	Num of Subjects	Effect
Chi et al. (2010a)	Physics	Uni	Lab	RL^-	29	Sig
Lin et al. (2015) Exp 1	Linear Algebra	Uni	Lab	RL^-	13	Sig
Lin et al. (2015) Exp 2[a]	Linear Algebra	Uni	Lab	RL^-	12	Sig
Zhou et al. (2017)	Probability	Uni	Class	R	77	Sig
Shen and Chi (2016b)	Logic	Uni	Class	R	33	ATI
Shen et al. (2018a) Exp 1	Logic	Uni	Class	R	37	ATI
Shen et al. (2018a) Exp 2	Logic	Uni	Class	R	34	ATI
Shen et al. (2018b)	Logic	Uni	Class	R	39	ATI
Chi et al. (2009)	Physics	Uni	Lab	R	37	Not sig
Rowe and Lester (2015)[b]	Microbiology	Gr 8	Class	R	28	Not sig

[a] The distinguishing factor between Exp 2 and Exp 1 by Lin et al. (2015), is that in Exp 2, the subjects did not have prior knowledge in the domain while in Exp 1, they did have prior knowledge

[b] This experiment was reported earlier in a dissertation (Rowe 2013)

Table 7 Summary of all empirical studies in the maximizing other objectives cluster. Details are as described in the caption of Table 3

Paper(s)	Domain	Population	Setting	Baseline Policies	Num of Subjects	Effect
Beck et al. (2000)	Arithmetic	Gr 6	Class	H	39	Sig
Mandel et al. (2014)	Fractions	Kids	Game	H/R	500	Sig

For the studies that use RL to sequence worked examples and problem solving tasks, we note the existence of an expertise-reversal effect (Kalyuga et al. 2003), where novices benefit more from reviewing worked examples while experts benefit more from problem solving tasks. This suggests an ordering where worked examples are given prior to problem solving tasks (for learners who are initially novice). Renkl et al. (2000) have further shown that fading steps of worked examples over time, such that students have to fill-in incomplete steps of worked examples until they solve problems on their own, is more beneficial than simply pairing worked examples with problem solving tasks. Thus, in this setting, we know that the sequence of instructional activities can make a difference, which could help explain the relative empirical success of studies in this cluster.

In general, most of the studies in this cluster found either that RL-induced policies significantly outperformed baseline policies (four out of ten) or that there was an aptitude-treatment interaction favoring the RL-induced policy (four out of ten). However, the studies in this cluster often compared to a policy that randomly sequenced tasks. Thus, it is not known if the RL-induced adaptive policies explored in this cluster would do better than a more reasonable heuristic (e.g., as suggested by the expertise-reversal effect). Future work in this area is needed to determine whether RL is useful in inducing adaptive policies for sequencing activity types beyond heuristic techniques, or if RL can simply help find one of many decent policies that can outperform randomly sequencing activity types.

Maximizing Other Objectives

There are two studies that do not fit into any of the previous four clusters, because they do not optimize for how much or how fast students learn (see Table 7). Beck et al. (2000) sequence instructional activities in an intelligent tutoring system with the goal of minimizing the time spent per problem, which their resulting policy achieved. While minimizing the time per problem could result in teaching students faster, it could also lead to the policy choosing instructional activities that are less time consuming (but not necessarily beneficial for student learning). Mandel et al. (2014) try to maximize the number of levels completed in an educational game, and their RL policy does significantly increase the number of levels completed over both a random policy and an expert-designed baseline policy. While interesting, these two papers do not shed light on whether RL can be used to significantly improve student learning over strong baseline policies.

Discussion: Where's the Reward?

We now turn to analyzing what the results of this review tell us about how impactful RL has been in the domain of instructional sequencing, and when and where it might be most impactful. We discuss a few factors which we believe have played a role in determining the success of RL-based approaches.

Leveraging Psychological Learning Theory

Our results suggest that RL has seemingly been more successful in more constrained and limited settings. For example, the cluster where RL has been most successful is paired-associate learning tasks, which treats pieces of content as independent of one another. RL has also been relatively successful in sequencing for concept learning tasks, typically constrained tasks designed for understanding aspects of cognition in lab studies rather than authentic tasks in traditional classroom settings. Moreover, RL has been relatively successful in sequencing activity types, where the agent must typically only choose between one of two or three actions. However, when it comes to sequencing interdependent content, there is not yet evidence that RL can induce instructional policies that are significantly better than reasonable baselines. This could be in part due to the fact that under the assumption that content is interrelated, the student's state may be a complicated function of the history of activities done so far and estimating the parameters of such a model may require an inordinate amount of data.

We believe the relative success of RL in some of these clusters over others could, at least in part, be explained by the ability to draw on psychological learning theory. As mentioned earlier, for both paired-associate learning and concept learning tasks, the models that were used were informed by the psychology literature. On the other hand, for sequencing activity types and interdependent content, the models used were solely data-driven. Moreover, in the case of paired-associate learning tasks, we noted that as psychological models got more sophisticated over time, the result of using them to induce instructional policies also got more successful, to the point that policies from more sophisticated psychological models sometimes outperformed policies from simpler models (see Section "Paired-Associate Learning Tasks" for more details). We also noted that an instructional policy derived from the ARTS model (a psychological model that was not fit to data) outperformed an instructional policy derived from the data-driven model developed by Atkinson (1972b). Thus, in some cases, a good psychological theory might be more useful for finding good instructional policies than a data-driven model that is less psychologically plausible.

In addition, for paired-associate learning tasks and sequencing activity types, there are well-known results from psychology and the learning sciences that shows sequencing matters: the spacing effect (Ebbinghaus 1885) and the expertise-reversal effect (Kalyuga et al. 2003) respectively. On the other hand, for sequencing interdependent content, we do not yet have domain-general principles from the learning sciences that tell us whether and how sequencing matters.

Thus, psychology and the learning sciences can give us insights for both how to make RL more likely to succeed in finding good instructional policies as well as when to hypothesize the precise sequencing of instructional activities might matter. Settings which have been more extensively studied by psychologists—and hence where we have better theories and principles to rely upon—are often more constrained, because such settings are easier for psychologists to tackle. But this does not mean RL should only be used in simple, unrealistic settings. Rather, it suggests that we should leverage existing theories and principles when using RL, rather than simply taking a data-driven approach. We explore this idea further in Section “Planning for the Future”.

Prior Knowledge

RL may have more room for impact in instructional settings where students are learning material for the first time, because students have more room to learn and because there is less variance in students’ experiences. Almost all of the paired-associate learning tasks are in domains where students have never learned the material before, such as foreign language learning. In many of these studies, researchers specifically recruited students who did not have expertise in the foreign language. The same holds for concept learning tasks, where students are learning a concept that is artificially devised, and as such, new to the student. Moreover, many of the studies in the sequencing activity types cluster were also teaching content to students for the first time. For example, Chi et al. (2009, 2010a) explicitly recruited students that had taken high school algebra but not college physics (which is what their dialogue-based ITS covered). Zhou et al. (2017) and Shen and Chi (2016b), and Shen et al. (2018a, b) all ran experiments in a university course on discrete mathematics, where the ITS was actually used to teach course content to the students. This could also possibly explain why many of these studies found an aptitude-treatment interaction in favor of low-performing students: students who have more room to improve can benefit more from a better instructional policy than students who have more prior knowledge. On the other hand, almost all of the studies in the sequencing interdependent content cluster were on basic K-12 math skills, where the student was also presumably learning the content outside of using the systems in the studies. The only exceptions to this were the lab studies run by Green et al. (2011) with university students, which actually showed that RL-induced policies did outperform random policies but not expert hand-crafted or heuristic baselines.

When students are learning material for the first time, there is also less variance in terms of students’ intial starting state, which makes RL-based approaches more likely to find policies that work for many students from prior data. Furthermore, in many of these cases, students are only being exposed to the content via the RL policy, often in a single lab session, rather than learning content through other materials. This again reduces the variance in the effect of an RL policy and makes it easier to estimate a student’s state. Indeed, only three out of 15 studies that were run in classroom settings found an RL-induced policy was significantly better than baselines and four found aptitude-treatment interactions.

Baselines

Another factor that might affect why some studies were more likely to obtain significant results could be the choice of baseline policies. Among the 24 studies that found a significant effect or aptitude-treatment interaction, 17 of them (71%) compared adaptive RL-induced policies to a random baseline policy and/or other RL-induced policies that have not been shown to perform well, rather than comparing to state-of-the-art baselines. On the other hand, among the studies that did not find a significant effect, only 6 of them (35%) only compared to random or RL-induced baseline policies. This suggests that while the ordering of instructional activities matters, it does not give us insight into whether RL-based policies lead to substantially better instructional sequences than relying on learning theories and experts for sequencing. Indeed, in some studies, researchers intentionally compared to baseline policies designed to perform poorly (e.g., by *minimizing* rewards according to a MDP), in order to determine if instructional sequencing has any effect on student learning whatsoever (Chi et al. 2010a; Lin et al. 2015; Geana 2015).

Of course, it is important to note that random sequencing is not always unreasonable. In some cases, a random baseline may actually be a fairly decent policy. For instance, when the policy must decide whether to assign worked examples or problem solving tasks, both actions have been shown to be beneficial in general, and hence a policy that sequences them randomly is thought to be reasonable (Zhou et al. 2017; Shen et al. 2018a). Moreover, in paired-associate learning tasks, random policies may be reasonable because they happen to space problems fairly evenly. However, given that we now have better heuristics for potentially sequencing worked-examples and problem solving tasks (Kalyuga et al. 2003; Kalyuga and Sweller 2005) as well as paired-associate learning tasks (Pavlik and Anderson 2008; Lindsey et al. 2014), it would be useful to compare RL-induced policies to these more advanced baselines.

The most successful cases of demonstrating that RL-induced policies can outperform reasonable baselines are in the context of paired-associate learning tasks. Lindsey et al. (2014) compared their policy against both a policy that spaces units of vocabulary words over time and a policy that blocked units of vocabulary words. Pavlik and Anderson (2008) compared their policy against a heuristic that learners might naturally use when learning with flashcards. However, even in this context, there are more sophisticated (but not data-driven) algorithms that are commonly used in flashcard software such as Leitner system (Leitner 1972) and SuperMemo (Wozniak 1990). Future work should consider comparing to some of these state-of-the-art baseline to determine if RL-induced policies can improve upon current educational practice.

Robust Evaluations

Several of the studies that have been successful in using RL performed some kind of robust evaluation to try to evaluate in advance of the study if the proposed policy was likely to yield benefits, given some uncertainty over how students learn. Lindsey et al. (2014) justified their use of a greedy heuristic policy by some simulations they ran in prior work (Khajah et al. 2014) that showed the heuristic policy can be approximately as good as the optimal policy according to two different cognitive models (ACT-R and MCM). Rafferty et al. (2016a) also ran simulations to evaluate how well various policies would be under three different models of concept learning. Although they actually tested all policies that they ran in their simulations on actual students (for a better understanding of how effective various models and policies are), the kind of robust evaluation they did could have informed which policy to use if they did not want to test all policies. These techniques are specific instances of a method we proposed in prior work called the robust evaluation matrix (REM), which involves simulating each instructional policy of interest using multiple plausible models of student learning that were fit to previously collected data (Doroudi et al. 2017a). Mandel et al. (2014) used importance sampling, a technique that can give an unbiased estimate of the value of a policy without assuming any particular model is true, to choose a policy to run in their experiment. On the other hand, several of the studies that did not show a significant difference between adaptive policies and baseline policies, including one of our own, only used a single model to simulate how well the policies would do, and that model overestimated the performance of the adaptive policy (Chi et al. 2010a; Rowe et al. 2014; Doroudi et al. 2017a).

Of course, even robust evaluations are limited by the models considered when doing the evaluation. For example, in our second experiment reported in Appendix B, we used REM to identify a simple instructional policy that was expected to outperform a baseline according to several different models. However, our experiment showed no significant difference between the adaptive policy and the baseline. Post-hoc analyses helped us identify two factors that we had not adequately accounted for in our robust evaluations: (1) the student population in this experiment was quite different from the population in our past data that we used to fit the models, and (2) the order in which problems were presented was quite different than the order in our prior experiments. Despite the null experimental result, these evaluations led to insights about what aspects our models were not adequately considering, which could inform future studies and the development of better models of student learning.

Summary

In short, it appears that reinforcement learning has yielded more benefits to students when one or more of the following things held:

- the sequencing problem was constrained in one or more ways (e.g., simple learning task with restricted state space or restricted set of actions),
- statistical models of student learning were inspired by psychological theory,
- principles from psychology or the learning sciences suggested the importance of sequencing in that setting,
- students had fairly little prior knowledge coming in (but enough prior knowledge such that they could learn from the software they were interacting with),
- RL-induced policies were compared to relatively weak baselines (such as randomly presenting actions or policies that were not expected to perform well), and

– policies were tested in more robust and principled ways before being deployed on students.

This gives us a sense of the various factors that may influence the success of RL in instructional sequencing. Some of these factors suggest best practices which we believe might lead to more successfully using RL in future work. Others suggest practices that are actually best to avoid—such as using weak baseline policies when stronger baselines are available—in order to truly determine if RL-induced policies are beneficial for students. We now turn to how we can leverage some of these best practices in future work.

Planning for the Future

Our review of the empirical literature suggests that one exciting potential direction is to further combine data-driven approaches with psychological theories and principles from the learning sciences. Theories and principles can help guide (1) our choice of models, (2) the action space under consideration, and (3) our choice of policies. We briefly discuss the prospects of each of these in turn.

Psychological theory could help inform the use of reasonable models for particular domains as has been done in the case of paired-associate learning tasks and concept learning tasks in the literature. These models can then be learned and optimized using data-driven RL techniques. Researchers should consider how psychological models can be developed for educationally relevant domains beyond just paired-associate and concept learning tasks. Indeed such efforts could hopefully be productive both in terms of improving student learning outcomes in particular settings, as well as in testing and contextualizing existing or newly-developed theories.

Our results also suggest focusing on settings where the set of actions is restricted but still meaningful. For example, several of the studies described above consider the problem of sequencing worked examples and problem solving tasks, which meaningfully restricts the decision problem to two actions in an area where we know the sequence of tasks makes a difference (Kalyuga et al. 2003).

Finally, learning sciences principles can potentially help constrain the space of policies as well. For example, given that the expertise-reversal effect suggests that worked examples should precede problem solving tasks and that it is best to slowly fade away worked example steps over time, one could consider using RL to search over the space of policies that follow such a structure. This could mean rather than deciding at each time step what activity type to give to the student, the agent would simply need to decide when to switch to the next activity type. The expertise-reversal effect also suggests such switches should be based on the cognitive load on the student, which in turn can guide the representation used for the state space. Such policies have been implemented in a heuristic fashion in the literature on faded worked examples (Kalyuga and Sweller 2005; Salden et al. 2010; Najar et al. 2016), but researchers have not yet explored using RL to automatically find policies in this constrained space. Related to this, the learning sciences literature could suggest stronger baseline policies with which to compare RL-induced policies, as discussed in Section “Baselines”.

As the psychology and learning sciences literature identify more principles and theories of sequencing, such ideas can be integrated with data-driven approaches to guide the use of RL in instructional sequencing. Given that deep reinforcement learning has been gaining lots of traction in the past few years and will likely be increasingly applied to the problem of instructional sequencing, it seems especially important to find new ways of meaningfully constraining these approaches with psychological theory and learning sciences principles. A similar argument was made by Lindsey and Mozer (2016) when discussing their successful attempts of using a data-driven psychological model for instructional sequencing: "despite the power of big data, psychological theory provides essential constraints on models, and … despite the success of psychological theory in providing a qualitative understanding of phenomena, big data enables quantitative, individualized predictions of learning and performance."

However, given that finding a single plausible psychological model might be difficult in more complex settings, a complementary approach is to explicitly reason about robustness with respect to the choice of the model. Of course, such robust evaluations are not silver bullets and they can make inaccurate predictions, but even if the results do not match the predictions, this can help prompt new research directions in understanding the limitations of the models and/or instructional policies used.

Beyond these promising directions and suggestions, we note that the vast majority of the work we have reviewed consists of *system-controlled* methods of sequencing instruction that target *cognitive* changes. However, for data-driven instructional sequencing to have impact, we may need to consider broader ways of using instructional sequencing. The following are meant to be thought-provoking suggestions for consideration that build on current lines of research in the artificial intelligence in education community. In line with our recommendation to combine data-driven and theory-driven approaches, a common theme in many of these ideas is to combine machine intelligence with human intelligence, whether in the form of psychological theories, student choice, or teacher input.

Learner Control

In this review, we have only considered approaches where an automated instructional policy determines all decisions about what a learner should do. However, allowing for student choice could make students more motivated to engage with an instructional system (Fry 1972; Kinzie and Sullivan 1989) and may benefit from the learner's own knowledge of their current state. Among the studies reported in our empirical review, only (Atkinson 1972b) compared an RL-induced policy to a fully learner-controlled policy, and he found that while the learner-controlled policy was 53% better than random, it was not as good as the RL-induced policy (108% better than random). While this result was taken in favor of system-controlled policies, Atkinson (1972a) suggested that while the learner should not have complete control over the sequencing of activities, there is still "a place for the learner's judgments in making instructional decisions."

There are a number of ways in which a machine's instructional decisions could be combined with student choice. One is for the agent to make recommendations about what actions the student should take, but ultimately leave the choice up to the student. This type of shared control has been shown to succesfully improve learning beyond system control in some settings (Corbalan et al. 2008). Green et al. (2011) found that expert policies do better than random policies, regardless of whether either policy made all decisions or gave the student a choice of three actions to take. Cumming and Self (1991) also describe such a form of shared control in their vision of "intelligent educational systems," where the system is a collaborator to the student rather than an instructor. A related approach is to give students the freedom to select problems, but have the system provide feedback on students' problem-selection decisions, which Long and Aleven (2016) showed can lead to higher learning gains than system control. Another approach would be for the agent to make decisions where it is confident its action will help the student, and leave decisions that it is less confident about up to the student. RL-induced policies could also take learner decisions and judgements as inputs to consider during decision making (e.g., as part of the state space). For instance, Nelson et al. (1994) showed that learners can effectively make "judgments of learning" in paired-associate learning tasks, and remarked that judgments of learning could be used by MDPs to make instructional decisions for students. Such a form of shared control has recently been considered in the RL framework for performance support (Javdani et al. 2018; Reddy et al. 2018; Bragg and Brunskill 2019), but has not been considered in the context of instructional sequencing to our knowledge.

Teacher Control

Building on the previous point, sometimes when an instructional policy does not know what to do, it could inform the teacher and have the teacher give guidance to the student. For example, Beck and Gong (2013) have shown that mastery learning policies could lead to "wheel-spinning" where students cannot learn a particular skill, perhaps because the policy cannot give problems that help the student learn. Detectors have been designed to detect when students are wheel-spinning (Gong and Beck 2015; Matsuda et al. 2016). These detectors could then relay information back to teachers, for example through a teacher dashboard (Aleven et al. 2016b) or augmented reality analytics software (Holstein et al. 2018), so that teachers know to intervene. In these cases, an RL agent could encourage the teacher to pick the best activity for the student to work on (or a recommend a set of activities that the student could choose from). Finding the right balance between learner-control, teacher-control, and system-control is an open and important area of research in instructional sequencing.

Beyond the Cognitive

Almost all of the empirical studies we have reviewed used cognitive models of learning that were designed to lead to cognitive improvements in learning (e.g., how much students learned or how fast they learned). However, RL could also take into account affective, motivational, and metacognitive features in the state space and could also be used in interventions that target these non-cognitive aspects of student learning by incorporating them into reward functions. For example, could a policy be derived to help students develop a growth mindset or to help students develop stronger metacognitive abilities? While detecting affective states is a growing area of research in educational data mining and AIED (Calvo and D'Mello 2010; Baker et al. 2012), only a few studies have considered using affective states and motivational features to adaptively sequence activities for students (Aleven et al. 2016a). For example, Baker et al. (2006) used a detector that predicts when a student is gaming the system in order to assign students supplementary exercises when they exhibit gaming behavior and Mazziotti et al. (2015) used measures of both the student's cognitive state and affective state to determine the next activity to give the student. There has also been work on adaptive learning technologies that improve students' self-regulatory behaviors, but this work has not aimed to improve self-regulation via instructional sequencing per se (Aleven et al. 2016a). While there is a risk that modeling metacognition or affect may be even harder than modeling students' cognitive states in a reinforcement learning framework, there may be certain places where we can do so effectively, and the impact of such interventions might be larger than solely cognitive interventions.

Conclusion

We have shown that over half of the empirical studies reviewed found that RL-induced policies outperformed baseline methods of instructional sequencing. However, we have also shown that the impact of RL on instructional sequencing seems to vary depending on what is being sequenced. For example, for paired-associate learning and concept learning tasks, RL has been fairly successful in identifying good instructional policies, perhaps because for these domains, psychological theory has informed the choice of statistical models of student learning. Moreover, when determining the sequence of activity types, RL-induced policies have been shown to outperform randomly choosing activity types, especially for lower performing students. But for sequencing interdependent content, we have yet to see if a data-driven approach can drastically improve upon other ways of sequencing such as expert-designed non-adaptive curricula. While the order of content almost certainly matters for domains with interconnected content (like mathematics), it can be difficult to identify good ways to adaptively sequence content with typical amounts of data.

Even in the cases where RL has been successful, one caveat is that the baseline policies are often naïve (e.g., randomly sequencing activities) and may not represent current best practices in instructional sequencing. For this reason, it does not seem like RL-based instructional policies have significantly impacted educational practice to date. Some studies have shown that RL-induced policies can outperform more sophisticated baselines, but more work is needed in this area.

One of the key recommendations we have drawn from this review is that instructional sequencing can perhaps benefit most by discovering more ways to combine psychological theory with data-driven RL. More generally, we suggested a number of ways in which instructional sequencing might benefit by combining machine intelligence with human intelligence, whether in the form of theories from domain experts and psychologists, a teacher's guidance, or the students's own metacognition.

We conclude by noting that the process of using reinforcement learning for instructional sequencing has been beneficial beyond its impact on student learning. Perhaps the biggest success of framing instructional sequencing as a reinforcement learning problem has actually been its impact on the fields of artificial intelligence, operations research, and student modeling. As mentioned in our historical review, investigations in optimizing instruction have helped lead to the formal development of partially observable Markov decision processes (Sondik 1971; Smallwood and Sondik 1973), an important area of study in operations research and artificial intelligence. More recently, in some of our own work, the challenge of estimating the performance of different instructional policies has led to advancements in general statistical estimation techniques (Doroudi et al. 2017b) that are relevant to treatment estimation in healthcare, advertisement selection, and many other areas. Finally, in the area of student modeling, the robust evaluation matrix (Doroudi et al. 2017a) can help researchers not only find good policies but also discover the limitations of the models when a policy under-delivers. Not only should we use theories of learning to improve instructional sequencing, but also by trying to improve instructional sequencing, perhaps we can gain new insights about how people learn.

Acknowledgements The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130215 and R305B150008 to Carnegie Mellon University. This research was also supported in part by a Google grant and a Microsoft Research Faculty Fellowship. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences, the U.S. Dept. of Education, Google, or Microsoft. The authors thank the reviewers and editor for their thoughtful reviews and for suggesting changes that have led to a significantly improved paper.

Appendix A: Details of Studies in Empirical Review

For the studies that found a significant difference, Table 8 gives more details about the technical aspects of each study, such as the form of RL (online vs. offline), the models that were used, the nature of the RL-induced and baseline policies, the outcome variables of interest, and the effect sizes. Table 9 reports the same information for the remainder of the studies. The outcome variables used in the studies include posttest score, learning gains (i.e., posttest - pretest), normalized learning gains (NLG), time (i.e., how long it takes to reach some desired level of completion), performance (i.e., how well the student performs on some tasks *during* the intervention), and time per problem (i.e., how long students spend on the assigned problems on average). We note that many studies report on more than one outcome variables; in such cases we chose to only report one outcome variable, tending to favor the outcome variable closest to what the instructional policies were directly optimizing. Descriptions of the various models used in these studies (as well as acronyms used in Tables 8 and 9) are given in Appendix A.1; similarly a description of the variety of RL-induced and baseline policies is given in Appendix A.2. Appendix A.3 describes how some studies performed model selection or policy selection (i.e., how they chose which models or policies to use). There is of course a lot of relevant detail for each study that could

Table 8 Technical details and effect sizes of the experiments performed by all empirical studies where adaptive policies significantly outperformed all baselines

Paper(s)	RL Setting	Num of Actions	Model	Adaptive Policies	Baseline Policies	Outcome Variable	Effect Size
Laubsch (1969)	Online	$\binom{140}{25}$	1. RTI	1. Myopic-1	Cycle	Posttest	1: $d = 1.02$
			2. OEM	2. Optimal			2. Not sig.
Atkinson and Lorton (1969)	Online	24	OEM	Myopic-1	Cycle	Posttest	$d = 0.96$
Atkinson (1972b)	Offline	12	1. 3SM-Hom	Myopic-1	A. Cycle	Posttest	1vB: 36% increase
			2. 3SM-Het		B. Student Choice	Posttest	2vB: Not sig
Chiang (1974) Exp 1	Online	$\binom{84}{21}$	1. 3SM-Hom	Myopic-1	Cycle	Posttest	1: ?
			2. 3SM-Het				2: ?
Chiang (1974) Exp 2	Mixed[a]	$\binom{84}{21}$	3SM-Het	Myopic-1	3SM-Hom	Response Latency[b]	?
Beck et al. (2000)	Offline	12	Model-Free	TD(0)-based	Heuristic	Time Per Problem	30% reduction
Katsikopoulos et al. (2001) Exp 1	Offline	12	4SM-Hom	Myopic-1	Cycle	Posttest	$d = 1.04$
Pavlik and Anderson (2008)	Offline	180	ACT-R	Myopic-1	A. Cycle-till-Correct	Posttest	B: $d = 0.796$
					B. 3SM-Het		
Chi et al. (2010a)[c]	Offline	2	Feat-MDP	Optimal	Inverse	Adjusted Posttest	$d = 0.86$

Table 8 (continued)

Paper(s)	RL Setting	Num of Actions	Model	Adaptive Policies	Baseline Policies	Outcome Variable	Effect Size
Lindsey et al. (2014)	Online	221	DASH	Threshold	A. Massed B. Spaced	Posttest	B: $d = 1.05$
Lindsey (2014)	Mixed	221	DASH	Threshold	A. Random B. Spaced	Posttest	A: $d = 0.18$[d]
Mandel et al. (2014)	Offline	7	Feat-POMDP	1. QMDP Optimal 2. Fixed	A. Expert B. Random	# of Levels Completed	1vB: 32% increase 2vB: Not sig
Lin et al. (2015) Exp 1	Online	3	Model-Free	Genetic Algorithm	Inverse	Posttest	$d = 1.29$
Lin et al. (2015) Exp 2	Online	3	Model-Free	Genetic Algorithm	Inverse	Posttest	$d = 1.48$
Rafferty et al. (2011)	Offline	45	POMDP	Myopic-2[e]	Random	Time	46% faster
Rafferty et al. (2016a) Exp 1	Offline	45	POMDP	1. Myopic-2 2. Max Info Gain	Random	Time	1. 54% faster 2. 57% faster
Rafferty et al. (2016a) Exp 2	Offline	300	POMDP	1. Myopic-2 2. Max Info Gain	Random	Time	1. Sig faster[f] 2. Sig worse
Papoušek et al. (2016)	Offline	?	PFA + Elo	Threshold[g]	Random	Test Items	?
Zhou et al. (2017)	Offline	2	Feat-MDP	Optimal	Random	Adjusted Posttest	$d = 0.43$

Table 8 (continued)

Paper(s)	RL Setting	Num of Actions	Model	Adaptive Policies	Baseline Policies	Outcome Variable	Effect Size
Leyzberg et al. (2018)	Online	4	3SM	Myopic-1	Random	Posttest	$d = 2.47$
Sen et al. (2018)	Offline	71	ANN	Fixed	A. Expert	Posttest	A. $d = 0.16$

Descriptions of the models and policies are provided in the main text. Some experiments use multiple adaptive policies, where either the model differs or type of policy differs. In such cases the different models/policies are designated by a number. Similarly, when multiple baseline policies are used, the various baseline policies are designated by a letter. The effect size column shows the effect of each adaptive policy compared to the baseline policy that was found to perform best (e.g., 1vB indicates adaptive policy 1 compared to baseline policy B, where B was the best baseline policy). Cohen's d effect sizes are presented when they were given or could be derived; in other cases the percentage improvement over the baseline is presented. A "?" indicates something that we could not deduce from the paper

[a] In this experiment, data from Chiang (1974) Exp 1 were used to initialize the model parameters, which was expected to help the 3SM-Het model, since this model has individual parameters for each item. Additionally, a different form of the 3SM model was used in Exp 2, determined by comparing model fits to data from Exp 1

[b] Chiang (1974) Exp 2 actually used posttest scores as an outcome variable, but no significant difference was found between policies. However, since they found a significant difference in response latency, we included this study among the ones that found a significant difference

[c] In addition to this study, Chi et al. (2010b) perform a quasi-experimental study where they compare the adaptive policy from Chi et al. (2010a) to both the adaptive and baseline policies from Chi et al. (2009), and find that the new adaptive policy is statistically significantly better. Note that the policy from Chi et al. (2010a) was fit to data from the other two policies, which were tested in previous years

[d] The effect is likely deflated as students found a way to skip review words, which was the content that was being sequenced

[e] Rafferty et al. (2011) actually used three POMDP policies that leveraged different assumptions about how people learn concepts as described in Section "Concept Learning Tasks". We report the significance for the best of these. Rafferty et al. (2016a) also used the same three POMDP policies in addition to a different way of deriving a policy from one of the POMDPs (maximum information gain); they also compared against two random baseline policies that differed in terms of which actions they could randomly select. We report results for the best POMDP policy compared to the best random baseline policy

[f] Rafferty et al. (2016a) Exp 2 compared policies on three different Number Game tasks, and found that averaging over all tasks, the POMDP policies outperformed two random baseline policies. However, it appears there is no one POMDP policy that was statistically significantly better than one of the baseline policies *for all* tasks

[g] Papoušek et al. (2016) actually tested three different adaptive policies: one that adaptively chooses the question and randomly chooses the set of distractors, one that randomly chooses the question and adaptively chooses the distractors, and one that adaptively chooses both. Interestingly, they found that only adaptively choosing the distractors to meet a threshold level of difficulty seemed to have a significant difference

Table 9 Technical details and effect sizes of the experiments performed by all empirical studies where adaptive policies *did not* significantly outperform baselines

Paper(s)	RL Setting	Num of Actions	Model	Adaptive Policies	Baseline Policies	Outcome Variable	Effect Size
Shen and Chi (2016b)	Offline	2	Feat-MDP	Optimal	Random	NLG	d = 1.01 (ATI)
Shen et al. (2018a) Exp 1	Offline	2	POMDP	Optimal?	Random	NLG	d = 0.82 (ATI)
Shen et al. (2018a) Exp 2	Offline	2	1. POMDP	1. Optimal?	Random	1. Time	1. d = 1.01 (ATI)
			2. Model-Free	2. DQN		2. NLG	2. Not sig
Shen et al. (2018b)	Offline	2	1. Feat-MDP	1. ?	Random	Posttest	1: d = 0.83 (ATI)
			2. Feat-POMDP	2. ?		Posttest	2. Not sig
Lindsey et al. (2013)	Online	256	Model-Free	BO	A. Max Fade[a]	Posttest	A. 12.6% increase
					B. Max Fade + Block		B. Not sig
					C. Intermediate		C. Not sig
Whitehill and Movellan (2017)	Offline	16997	POMDP	Policy Gradient	A. Random	Time	A. 27% faster
					B. Heuristic		B. Not sig
Green et al. (2011) Exp 1	Offline	?	FMDP	Optimal	A. Random	Posttest	A. $d \approx 3.9$
					B. Heuristic		B. Not sig
Green et al. (2011) Exp 2	Offline	?	FMDP	Optimal	A. Random	Posttest	A. ?
					B. Expert[b]		B. Not sig
Dear et al. (1967)	Offline	16	OEM	Optimal	Cycle	Posttest	Not sig
Katsikopoulos et al. (2001) Exp 2	Offline	24	OEM	1. Myopic-1	Cycle	Posttest	1. Not sig
				2. Myopic-1 + Lag			2. Not sig
Chi et al. (2009)[c]	Offline	2	Feat-MDP	Optimal	Random	Posttest	Not sig
Rowe and Lester (2015)	Offline	2-6[d]	Feat-MDP	Optimal	Random	Posttest	Not sig

Table 9 (continued)

Paper(s)	RL Setting	Num of Actions	Model	Adaptive Policies	Baseline Policies	Outcome Variable	Effect Size
Clement et al. (2015)	Online	≥ 28	Model-Free	MAB	Inc Difficulty	Difficulty Level	Not sig
Geana (2015)	Offline	216	RL Agent	1. Myopic	A. Inverse	Posttest	1vA: d = 0.45
							1vB: Not sig?[e]
			Bayesian Learner	2. Myopic	B. Inverse		2vB: Not sig
David et al. (2016)	Offline	?	BKT-based	Threshold	Inc Difficulty	Performance	Not sig
Schatten (2017)	Offline	?	Matrix Factorization	Threshold	Expert	Posttest	Not sig
Doroudi et al. (2017a)	Offline	155	1. FMDP	1. Mastery	Spiral Difficulty	Posttest	Not sig
			2. FMDP	2. Myopic-2			
			3. BKT	3. Mastery			
Appendix B	Offline	155	BKT	Inc Time	Spiral Difficulty	Posttest	Not sig
Segal et al. (2018)	Mixed	?	1. Model-Free	1. MAB	Inc Difficulty	Posttest	Not sig
			2. BKT	2. Threshold			Not sig
Mettler et al. (2011)	Offline	24	3SM-Het	Myopic-1	ARTS	Posttest	d = -0.68

A policy is labeled as “Optimal?” when we could not verify that it was the optimal policy. For policies with a mixed effect, the effect of comparing the adaptive policy vs. each baseline is shown (rather than just the best baseline). Other details are as described in the caption of Table 8

[a] Lindsey et al. (2013) parameterize policies by the degree of fading and degree of blocking. The three baseline policies were maximum fading and no blocking, maximum fading and blocking, and intermediate fading and blocking. See the paper for more details

[b] Green et al. (2011) also compared their FMDP policy to both expert and random baseline policies that gave students a choice of three problems to choose from. The overall finding was that regardless of whether students had a choice, the DBN and expert policies outperformed random

[c] Chi et al. (2009) actually performs a quasi-experimental study, where the baseline policy was executed earlier and data from the baseline policy was actually used to induce the adaptive policy

[d] Rowe and Lester (2015) used different MDPs for different decision points in their game, and the number of actions considered for each MDP ranged from two to six

[e] Geana (2015) did not actually report the significance of this comparison, but it appears to be not significant

not be encapsulated in these tables; we refer the reader to the individual papers for a better understanding of the experiments conducted.

A.1 Models of Learning

Several models of learning have been used to derive instructional policies. Some of these models are based on psychological theories, whereas others were devised from a more machine learning or data-driven perspective. Here we summarize the variety of models used in the empirical studies along with the acronyms used for the models in Tables 8 and 9. We begin by describing the models used for paired-associate learning tasks in the first wave of RL applied to instructional sequencing:

- One-Element Model (**OEM**): This is a simple model that describes learning in all-or-none fashion: you either know something or you don't (Bower 1961). The OEM supposes that for each item, the student is either in a latent learned state or unlearned state. If the student is in the learned state, they will always answer questions on the item correctly, but if they are in the unlearned state, they have a probability of guessing, but will otherwise answer incorrectly. With each item presentation, the student also has some probability of learning the item.
- Single-Operator Linear Model (**SOL**): In opposition to the OEM, this model assumes learning occurs incrementally. In particular, with each item presentation, the probability of a student answering incorrectly decreases by a constant multiplier.
- Random-Trial Increment (**RTI**): This model combines OEM and SOL; with each item-presentation there is some probability that the the student will incrementally learn the skill (i.e., that the probability of answering incorrectly decreases by a constant multiplier).
- Three-State Markov Model (**3SM**): This describes a variety of models like the final model discussed in Section "Examples of RL for Instructional Sequencing". There are three latent knowledge states (instead of two as in OEM), and when the student is in the intermediary state, there is some probability that they will forget the item. Forgetting can occur with each presentation of any other item; therefore, unlike all the previous models, a student's knowledge of an item can change even when the student is practicing other items. These models allow for spacing effects. We will differentiate between two cases of this model: one where all items are assumed to be homogeneous or have the same parameters (**3SM-Hom**), and one where the items can be heterogeneous or each item is allowed to have different parameters (**3SM-Het**). Many years after the experiments performed by Atkinson and colleagues, Katsikopoulos et al. (2001) builds on this work by using a similar four-state Markov model (**4SM**).

Although these models are relatively simple and all assume that items are independent of one another, several interesting properties emerge when considering using these models to derive instructional policies. First, we note that if we consider the problem of which item to present to a student, then all of the models above can be described as factored POMDPs (where the state can be factored into individual items, and there is a separate transition and observation matrix for each item), provided that we give an appropriate reward function, such as a delayed reward proportional to the number of items learned in the end. The SOL model in particular can be described as a factored MDP, since we can determine the state of each item uniquely by how many times it was presented to the student. Moreover, SOL is a deterministic MDP. Thus, the optimal policy for SOL is a non-adaptive policy (or response insensitive strategy, as it is referred to in the literature). In particular, if we assume homogeneous parameters, the optimal policy is to repeatedly cycle through all items in any order to maximize the minimum number of presentations for each item. Even though this is technically the optimal policy for a model, it is such a simple instructional policy that it is commonly used as a baseline in many of the empirical studies.

On the other hand, for OEM, the optimal strategy is adaptive (or response sensitive). In particular, as shown by Karush and Dear (1967), the optimal policy for OEM (assuming homogeneous parameters) is simply to give the item that has had the shortest consecutive streak of correct responses. Interestingly, as with SOL, (1) the myopic policy is optimal, and (2) this optimal policy does not depend on the parameters of the OEM model, so one does not need to learn the parameters of the model in order to execute its optimal policy. Even though no learning is necessary, we still consider the optimal policy for the OEM model to be an RL-induced policy, albeit a simple one.

We now consider models that have been used in more recent papers:

- Bayesian Knowledge Tracing (**BKT**): This model is identical to OEM, except that it also allows for a probability of slipping when the item or skill is in the learned state. In addition, it is typically used in a different context than OEM. In particular, an instructional action (such as a problem in an ITS) may have several different skills on it. When a student works on a problem, their knowledge state of several skills may change. Thus an optimal policy for BKT may be much more complicated than for OEM; however, a heuristic policy is typically used for BKT where problems are given until a student is believed to be in the learned state with high probability (e.g., at least 95%). David et al. (2016) used a modified version of BKT to choose problems believed to be in the student's zone of proximal development.
- Performance Factors Analysis (**PFA**): This is a model that was proposed as an alternative to BKT in the educational data mining literature (Pavlik et al. 2009). It models the probability of answering and item correctly as a logistic function that depends on the number of times the item was answered correctly in the past and the number of times it was answered incorrectly. Although it is commonly used to predict learning, Papoušek et al. (2016) are the only ones who have evaluated its efficacy in instructional sequencing to our knowledge. Papoušek et al. (2016) use a modified version of the PFA algorithm by combining it with the Elo rating system (Pelánek et al. 2017).
- Featurized (PO)MDP (**Feat-(PO)MDP**): This type of model has been popular in recent papers that consider using RL for instructional sequencing. The idea is to

create a MDP (or POMDP) that uses a set of features collected from the software to encapsulate the state of the MDP (or observation space of the POMDP). Features could include, for example, whether the student answered questions of various types correctly in the past, actions taken by the student in an ITS, and aspects of the history of instructional actions such as how many actions of type X the policy has given so far. Since there are many features one could feasibly consider, some papers have looked at various methods of doing feature reduction (Chi et al. 2009; Zhou et al. 2017) and feature compression (Mandel et al. 2014). For instance, Mandel et al. (2014) used features to represent the observation space of a POMDP; however, to make the problem tractable, they did feature compression on thousands of features, which resulted in only two features.

- Factored MDP (**FMDP**): This is a particular type of featurized MDP, where the state space is factored into a set of features and each feature's transition dynamics only depends on some (ideally small) subset of features. This is a useful way to represent how students learn a set of interrelated skills, where the ability to learn one skill might depend on some, but not all, of the remaining skills. The relationship between skills can be determined by a domain expert (e.g., through a prerequisite graph) (Green et al. 2011) or determined automatically (Doroudi et al. 2017a).
- **POMDP**: All of the models above could be described as a particular type of POMDP, but some papers explicitly use the POMDP formulation to describe their models of learning. Rafferty et al. (2016a) and Whitehill and Movellan (2017) use POMDPs to naturally describe how learners perform concept learning tasks based on cognitive science theories. For example, Rafferty et al. (2016a) use models where the state space consists of hypotheses over concepts, and when the student is presented information that goes against their hypothesis, they randomly transition to a new hypothesis that is in line with the evidence presented.
- **ACT-R**: ACT-R is a cognitive architecture that generally describes human cognition and how people acquire procedural and declarative knowledge (Anderson 1993). Pavlik and Anderson (2008) extended ACT-R and used it to derive an instructional policy for sequencing vocabulary items.
- **DASH**: DASH combines a data-driven logistic regression model with psychological theories of memory that capture "difficulty, ability, and study history" (Lindsey et al. 2014).
- **Machine Learning Models**: Some papers have assumed that the learner learns according to a particular type of machine learning model. For example, (Sen et al. 2018) assume the learner learns according to an artificial neural network (**ANN**) and (Geana 2015) assumes that learners are either a reinforcement learning agent or a Bayesian learner. Notice that assuming the student is an RL agent is different from assuming that the cognitive state of the student changes according to a MDP. Rather, it means that the student is an RL agent that is trying to learn the parameters of a MDP.

A.2 Instructional Policies

A.2.1 Model-Based Policies

Given a model, there are many ways to derive an instructional policy. In this section, we will mention some of the most common types of policies used for model-based methods:

- **Optimal**: This refers to the optimal policy given a model. Optimal policies for MDPs can be derived using dynamic programming methods such as value iteration (Bellman 1957) and policy iteration (Howard 1960a). As mentioned earlier, for simple models such as SOL and OEM, the optimal policy takes a very simple form that does not actually depend on the models. Optimal policies are typically only used for these simple models as well as for feature-based MDPs. For more complex models such as POMDPs and FMDPs with large state spaces, solving for the optimal policy can be intractable. An approximation can be made for POMDPs by solving for the optimal policy in the equivalent QMDP, which is a POMDP where the state is assumed to be known after one action (**QMDP Optimal**).
- **Myopic**: These policies optimize the reward function by only considering the consequences of the next action (**Myopic-1**) or the next two actions (**Myopic-2**), rather than considering the entire horizon. In some simple cases the myopic strategy might be optimal, and in other cases it might be close to optimal (Matheson 1964).
- **Threshold**: Given a model that predicts the probability that a student will answer any item correctly, a commonly used policy is to pick the item closest to some threshold. Interestingly, such policies are motivated in terms of educational theories such as desirable difficulty (Lindsey et al. 2014) and the zone of proximal development (Schatten 2017). Khajah et al. (2014) showed in simulation that a threshold policy could be nearly optimal according to two different models of student learning for tasks where there are a set of independent items like paired-associate learning tasks.
- **Fixed**: A fixed policy is a non-adaptive policy that is meant to be near optimal according to a given model. Mandel et al. (2014) used the best fixed policy according to their model, while Sen et al. (2018) used a hill-climbing approach to find a good fixed policy that led to the least error for their neural network model (although the policy found could have achieved a local optimum).

A.2.2 Model-Free Policies

A minority of the papers that derived RL-induced policies for instructional sequencing did so using model-free methods. In some cases, authors used RL methods such as temporal difference learning (**TD(0)**) (Sutton and Barto 1998) and the Deep Q-Network (**DQN**) (Mnih et al. 2015). In other cases, related methods such as Bayesian optimization (**BO**) (Mockus 1994; Brochu et al. 2010), multi-armed bandits (**MABs**) (Gittins 1979; Auer et al. 2002), and genetic algorithms were used. While MABs are typically used to find the best static decision (assuming no change in state), researchers have used them in novel ways to account for student learning, such as by progressing students through a knowledge graph and using the multi-armed bandit to find the best actions among the set that the student is ready for Clement et al. (2015) or by modifying which actions are optimal by adjusting the weights of actions based on the level of difficulty student can currently tackle (Segal et al. 2018). Like the model-based threshold policies, these policies also motivate the idea of using a set of appropriately challenging problems at any given time using educational theory, namely the zone of proximal development (Clement et al. 2015; Segal et al. 2018).

A.2.3 Baseline Policies

RL-induced policies have been compared to a variety of baseline policies. Some of the common baseline policies include:

- **Random**: This policy presents content in a random order. Although random sequencing will often be a weak baseline, it can be reasonable in cases where the content being sequenced does not have strong dependencies, especially given that interleaving content has been shown to be effective.
- **Cycle**: This policy randomly cycles through items, such as words in a paired-associate learning task, in a randomly-determined fixed order. As mentioned earlier, this is actually the optimal policy under the simple SOL model. However, because it is one of the simplest policies that could be considered and a non-adaptive policy, we consider it as a baseline policy whenever it is used.
- **Inverse**: This policy takes the reward function of an MDP and minimizes it instead of maximizing it. Therefore, it is a policy that is made to intentionally perform poorly. The idea behind using such a policy is to show that the ordering of instructional activities actually makes a difference. However, this baseline cannot be used to discern if RL-induced policies are effective ways of teaching students beyond reasonable methods for instructional sequencing.
- **Inc Difficulty**: This general type of policy orders content in order of increasing difficulty or complexity as determined by domain experts. In our experiments, we used a modified version of this general policy referred to as **Spiral Difficulty** where three broad topics were ordered in terms of difficulty, but once students finished twelve randomly chosen problems per topic, they would be returned to the first topic and so on, loosely motivated by the idea of a spiral curriculum (Bruner 1960; Harden 1999).

A.3 Model/Policy Selection

In Tables 8 and 9, we report the models ultimately used in each of the studies; however, many of the researchers considered a variety of models and policies before settling on one in particular. These researchers used a variety of model selection (or policy selection) criteria to choose which policy to use in the experiments. For a number of different studies, Min Chi and colleagues (Chi et al. 2009, 2010a; Shen and Chi, 2016b; Zhou et al., 2017; Shen et al., 2018a) fit several MDPs on different featurized representations of the state space and used the expected cumulative reward (ECR) of the optimal policies under each model to determine which policy was expected to perform best. As shown by Mandel et al. (2014), ECR is a biased and inconsistent estimator of the true value of a policy. Instead, Mandel et al. (2014) used importance sampling to evaluate policies for several different types of models to settle on a final policy. While the importance sampling estimator can give an unbiased estimate of the value of a policy, it is only data-efficient when sequencing a few instructional activities and can lead to biased decisions when used for policy selection (Doroudi et al. 2017b).

To provide a more robust estimator than ECR while mitigating the data inefficiency of importance sampling, we proposed the robust evaluation matrix method to perform policy selection (Doroudi et al. 2017a). REM involves simulating each instructional policy of interest using multiple plausible models of student learning that were fit to previously collected data. If a policy robustly outperforms other policies according to multiple different models, we can have increased confidence that it will actually be a better instructional policy rather than "overfitting" to a particular model (which ECR is susceptible to). We used REM to select a policy for our second experiment as described in Appendix B. Several other studies used similar robust evaluation methods (Rafferty et al. 2016a; Lindsey et al. 2014).

Appendix B: Case Study: Fractions Tutor Experiment

Here we report on a case study of our own work in applying RL to instructional sequencing in a fractions intelligent tutoring system (ITS), as this experiment is unpublished and was a primary motivation behind writing the current paper. We ran two experiments in our ITS, both of which resulted in no significant difference in posttest performance between any of the conditions. We have reported on the first experiment in prior work (Doroudi et al. 2017a). Given that our experiments were for a very particular use of RL in a particular domain and tutoring system, we questioned the generalizability of our null results and were initially hesitant about how informative they would be to the broader research community. However, by situating our particular null results in the broader review we present in this paper, we believe we can gain some insights from this case study.

B.1 Fractions Tutor

We extended the Fractions Tutor, a web-based intelligent tutoring systems for fourth and fifth grade fractions learning (Rau et al. 2012, 2013). Our version of the ITS included activities for three topics: making and naming fractions, fraction equivalence and comparison, and fraction addition. Our ITS was also designed to include activities that support the three learning mechanisms posited by the knowledge-learning-instruction (KLI) framework (Koedinger et al. 2012): sense-making, induction and refinement, and fluency building. Our goal was to find an adaptive policy that could discover how to optimally sequence content in these three topics as well as how to sequence different types of activities to support learning mechanisms at the appropriate time. KLI does not give strong recommendations about how to optimally order activities that support each of the learning mechanisms, therefore we were hoping our data-driven approach would not only lead to a good adaptive policy for our ITS but would also inform the theory behind KLI learning mechanisms. However, our policies did not make decisions at the level of topic or learning mechanism. Rather we took the more ambitious route of trying to pick a specific activity (one of 155) at each time step. These activities included individual problems, groups of problems, and videos followed by conceptual questions, but for simplicity we use the terms activities and problems interchangeably.

B.2 Data Collection

We initially collected data from over 1000 students working on our ITS in a number of different schools. Students took a pretest, used the tutor for several sessions, and then took a posttest that was identical to the pretest. Students were free to work at their own pace and hence completed varying numbers of problems. We presented activities in a semi-randomized order as a compromise between two potentially competing objectives. The first objective was to enhance student learning for the students that participated in this initial data collection. This objective would push us towards selecting an activity order that draws upon existing research on effective sequencing, and satisfies commonly assumed topic orderings (e.g. obtaining a basic understanding of fractions before doing fraction addition). Our second objective was to find a good instructional policy. To find a good instructional policy, RL methods require that many states were explored in the initial dataset and various actions were taken from each state. Since a student's state could potentially depend on the actions taken thus far, to ensure we explore the state-action space sufficiently, we would ideally give random actions. Our semi-randomized order enforced that the first 26 activities a student saw consisted of activities that we believed were important for the student to encounter before moving on to more advanced topics. Moreover, among those 26 activities, the sequence of activities was restricted to follow a prerequisite graph. When there were multiple possible activities that could have been presented, the activity was chosen randomly. After the first 26 activities, activities were chosen uniformly at random without replacement from a pool of 130 activities.

In addition to this data, we collected data from over 300 students in our first experiment. In this experiment, students were assigned problems according to one of five instructional policies; see Doroudi et al. (2017a) for more details. We used data from both experiments to fit models for our robust evaluation matrix. Since these two datasets consisted of different policies and different student populations and were collected in different years, we fit the same models to each dataset in order to improve the robustness of REM.

B.3 Policy Selection

The null results of our first experiment led to the development of the robust evaluation matrix as a tool for off-policy policy selection (Doroudi et al. 2017a). A natural next step was to see if REM could be used to actually discover a good instructional policy for our next experiment. We wanted to compare a new instructional policy to one of the baseline policies from our first experiment. This baseline policy only gave problems that supported the induction and refinement mechanism (as these problems are standard in many ITSs) and spiraled through the curriculum by first giving several problems on making and naming fractions, then on fraction equivalence and comparison, and finally on fraction addition, and then giving problems in that order again. Problems were randomly chosen within each topic. In what follows, we will refer to this policy as “the baseline policy”.

Before coming up with new instructional policies, we wanted to include the time spent per problem in our REM analyses, as that is something we ignored when devising policies for our first experiment. Namely, the off-policy estimation we did prior to our first experiment assumed that 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 had considered 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 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 the baseline policy only assigns 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 the baseline did as well as the other policies in our experiment; these students completed more problems, which could make up for the lack of diversity or adaptivity in problem selection. To tackle this issue, each of the student models we used in REM assumed that the time per problem was sampled from how long students took in our prior data, and to increase robustness, we experimented with sampling times from different student populations that we had data for. However, as we will soon demonstrate, such a simple model for predicting time per problem was not sufficient.

To see how important the time spent per problem might be, we tested a simple policy that sequenced problems in increasing order of average time students spent in our previous experiment (i.e., students would first get the problem that took the least amount of time on average). REM predicted that this policy would be better than the baseline policy under a variety of (but not all) student models. To make this policy adaptive, we augmented this policy with a simple rule to skip any problem where all skills taught in that problem were already believed to have been mastered, using a Bayesian Knowledge Tracing model with a mastery threshold of 0.9. We thought this might help avoid over-practice, especially because assigning problems in order of increasing time often meant giving similar problems multiple times in sequence. Indeed, this new adaptive policy was predicted by REM to be considerably better than the baseline according to many student models, including ones that predicted the non-adaptive version would be worse than the baseline. Models predicted the improvement of this new policy over the baseline would be between 0.31 and 2.23 points on the posttest (out of 16 points), with most models predicting an improvement of at least one point on the posttest. Thus we chose to use this policy in our next experiment.

B.4 Experiment and Results

We ran an experiment with 220 4th and 5th grade students to see if our new data-driven adaptive policy could outperform the baseline induction and refinement policy. The experimental design was the same as for our initial data collection and first experiment; only the instructional policies were changed. Despite our REM predictions, when we ran our experiment, we found that students assigned the baseline policy had a mean posttest score of 8.12 (out of 16) and students assigned the new adaptive policy had a mean posttest score of 7.97, indicating the new policy was no better than the baseline. In terms of learning gains (posttest minus pretest score), the baseline had a mean score of 1.32, while the new adaptive policy had a mean scores of 1.55. While there was a positive difference in learning gains, it was not significant.

B.5 Discussion

So one might ask, why did the new policy fare no better than the baseline, when REM predicted otherwise? There are two factors that we did not adequately account for in our REM analyses: (1) the student population in this experiment was quite different from the population in our past data that we used to fit the models, and (2) the order in which problems were presented was quite different than in our prior experiments. To account for the first issue, we had done REM analyses by fitting models to subpopulations of our prior data, but we had still predicted that the new adaptive policy would do better. We did more extensive analyses after the experiment, and we found that the predicted difference between the two policies was much smaller for students from a particular school district. Developing models and instructional policies that can generalize to new student populations is a big open question in the literature (Baker 2019). While REM can help with this by seeing how different policies might interact with different populations of students we have collected data from, it cannot definitively tell us how the policy will effect new students.

The second issue may have had an even greater effect on our results. All of the models that we used in REM assumed that the time per problem was sampled according to our prior data. Our new adaptive policy gave problems that took the least amount of time first, but it ignores the fact that students in our previous experiments had typically done those problems after having completed many other problems, which could be why they worked through those problems quickly. Indeed, in our experiment we found that problems given early on were taking students much longer than those same problems took for students in our first experiment or in the baseline condition. Our experiment highlights the importance of not only modeling how students answer problems over time, but also how long they spend on problems, especially when we want to use time spent as a variable to determine how to adaptively assign problems to students. We believe future researchers can build on this insight in one of two ways: (1) developing more sophisticated ways of predicting how long students will spend on problems to use in offline analyses (such as REM analyses), or (2) developing policies that can be robust to how long students actually spend on problems by taking into account data collected from the student online (e.g., if a student appears to be slower than average on a certain type of problem, use that information in deciding what problem to give the student next).

While both of our experiments had null results, and we did not successfully demonstrate how using REM could lead to improved instructional policies, this process has revealed a number of challenges that can affect the process of using RL to induce instructional policies. Moreover, by identifying some of the limitations of *how we used* REM, we now have some insights that can lead to the development of more robust instructional policies, for example by taking the student population and time per problem into account. As we demonstrate in this paper, we were not the only researchers to have faced challenges in demonstrating how RL could be used to derive impactful instructional policies. We hope that the retrospective insights we developed about REM as well as the insights drawn from the review presented in this paper can help researchers mitigate some of these challenges in the future.

References

Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R. (2016a). Instruction based on adaptive learning technologies. In Mayer, R.E., & Alexander, P.A. (Eds.) *Handbook of research on learning and instruction.* chapter 24. 2nd edn. (pp. 522–559): Routledge.

Aleven, V., Xhakaj, F., Holstein, K., McLaren, B.M. (2016b). Developing a teacher dashboard for use with intelligent tutoring systems. In *IWTA@EC-TEL* (pp. 15–23).

Almond, R.G. (2007). An illustration of the use of Markov decision processes to represent student growth (learning). *ETS Research Report Series*, 2007(2).

Andersen, P.-A., Kråkevik, C., Goodwin, M., Yazidi, A. (2016). Adaptive task assignment in online learning environments. In *Proceedings of the 6th international conference on web intelligence, mining and semantics*: ACM.

Anderson, J.R. (1993). Rules of the Mind. Lawrence Erlbaum Associates.

Antonova, R., Runde, J., Lee, M.H., Brunskill, E. (2016). Automatically learning to teach to the learning objectives. In *Proceedings of the third (2016) ACM conference on learning@ scale* (pp. 317–320): ACM.

Atkinson, R.C. (1972a). Ingredients for a theory of instruction. *American Psychologist*, *27*(10), 921.

Atkinson, R.C. (1972b). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, *96*(1), 124.

Atkinson, R.C. (2014). Computer assisted instruction: Optimizing the learning process. In Annual Convention of the Association for Psychological Science.

Atkinson, R.C., & Calfee, R.C. (1963). Mathematical learning theory. Technical Report 50, Institute of Mathematical Studies in the Social Sciences.

Atkinson, R.C., & Lorton, P. Jr.. (1969). Computer-based instruction in spelling: an investigation of optimal strategies for presenting instructional material. Final report. Technical report, U.S. Department of Health, Education, and Welfare.

Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, *32*(1), 48–77.

Baker, R.S. (2019). Some challenges for the next 18 years of learning analytics (Keynote at the 9th International Conference on Learning Analytics & Knowledge).

Baker, R.S., Corbett, A.T., Gowda, S.M., Wagner, A.Z., MacLaren, B.A., Kauffman, L.R., Mitchell, A.P., Giguere, S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. In *International conference on user modeling, adaptation, and personalization* (pp. 52–63): Springer.

Baker, R.S., Corbett, A.T., Koedinger, K.R., Evenson, S., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J.E. (2006). Adapting to when students game an intelligent tutoring system. In *International conference on intelligent tutoring systems* (pp. 392–401): Springer.

Baker, R.S., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J., Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra. In *Proceedings of the 5th international conference on educational data mining* (pp. 126–133). International Educational Data Mining Society.

Barnes, T., & Stamper, J. (2008). Toward automatic hint generation for logic proof tutoring using historical student data. In *International conference on intelligent tutoring systems* (pp. 373–382): Springer.

Beck, J., Woolf, B.P., Beal, C.R. (2000). Advisor: a machine learning architecture for intelligent tutor construction. In *Proceedings of the seventeenth national conference on artificial intelligence* (pp. 552–557): AAAI Press.

Beck, J.E. (1997). Modeling the student with reinforcement learning. In *Machine learning for user modeling workshop at the sixth international conference on user modeling*.

Beck, J.E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (Eds.) *Artificial intelligence in education* (pp. 431–440). Berlin: Springer.

Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, 679–684.

Bennane, A., D'Hondt, T., Manderick, B. (2002). An approach of reinforcement learning use in tutoring systems. In *Proceedings of the 1st international conference on machine learning and applications* (p. 993).

Bower, G.H. (1961). Application of a model to paired-associate learning. *Psychometrika*, *26*(3), 255–280.

Bragg, J., & Brunskill, E. (2019). Fake it till you make it: Learning-compatible performance support. In *Uncertainty in artificial intelligence*. Association for uncertainty in artificial intelligence.

Brochu, E., Cora, V.M., De Freitas, N. (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

Bruner, J.S. (1960). *The process of education*. Cambridge: Harvard University Press.

Brunskill, E., & Russell, S. (2011). Partially observable sequential decision making for problem selection in an intelligent tutoring system. In *Proceedings of the 4th international conference on educational data mining* (pp. 327–328). International Educational Data Mining Society.

Calvo, R.A., & D'Mello, S. (2010). Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37.

Champaign, J., & Cohen, R. (2010). A model for content sequencing in intelligent tutoring systems based on the ecological approach and its validation through simulated students. In *Proceedings of the twenty-third international florida artificial intelligence research society conference (FLAIRS 2010)* (pp. 486–491).

Chant, V.G., & Atkinson, R.C. (1973). Optimal allocation of instructional effort to interrelated learning strands. *Journal of Mathematical Psychology*, *10*(1), 1–25.

Chaplot, D.S., Rhim, E., Kim, J. (2016). Personalized adaptive learning using neural networks. In *Proceedings of the third (2016) ACM conference on learning@ scale* (pp. 165–168): ACM.

Chi, M., Jordan, P., VanLehn, K., Hall, M. (2008). Reinforcement learning based feature selection for developing pedagogically effective tutorial dialogue tactics. In *Proceedings of the 1st international conference on educational data mining* (pp. 258–265). International Educational Data Mining Society.

Chi, M., Jordan, P.W., Vanlehn, K., Litman, D.J. (2009). To elicit or to tell: Does it matter? In *Proceedings of the 2009 conference on artificial intelligence in education* (pp. 197–204). Amsterdam: IOS Press.

Chi, M., VanLehn, K., Litman, D. (2010a). Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In *International conference on intelligent tutoring systems* (pp. 224–234): Springer.

Chi, M., VanLehn, K., Litman, D., Jordan, P. (2010b). Inducing effective pedagogical strategies using learning context features. In *International conference on user modeling, adaptation, and personalization* (pp. 147–158): Springer.

Chiang, A. (1974). Instructional algorithms derived from mathematical learning models: An application in computer assisted instruction of pairedassociated items. PhD thesis, City University of New York.

Clement, B., Oudeyer, P.-Y., Lopes, M. (2016). A comparison of automatic teaching strategies for heterogeneous student populations. In *Proceedings of the 9th international conference on educational data mining* (pp. 330–335). International educational data mining society.

Clement, B., Roy, D., Oudeyer, P.-Y., Lopes, M. (2015). Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining (JEDM)*, *7*(2), 20–48.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd edn.

Corbalan, G., Kester, L., Van Merriënboer, J.J. (2008). Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology*, *33*(4), 733–756.

Corbett, A. (2000). Cognitive mastery learning in the act programming tutor. In *Papers from the AAAI spring symposium*: AAAI Press.

Corbett, A.T., & Anderson, J.R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Cumming, G.D., & Self, J. (1991). Learner models in collaborative intelligent educational systems. In Goodyear, P. (Ed.) *Teaching knowledge and intelligent tutoring* (pp. 85–104): Ablex Publishing Corporation.

Daubigney, L., Geist, M., Pietquin, O. (2013). Model-free pomdp optimisation of tutoring systems with echo-state networks. In *SIGDIAL conference* (pp. 102–106).

David, Y.B., Segal, A., Gal, Y.K. (2016). Sequencing educational content in classrooms using bayesian knowledge tracing. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 354–363): ACM.

Dear, R.E., Silberman, H.F., Estavan, D.P., Atkinson, R.C. (1967). An optimal strategy for the presentation of paired-associate items. *Systems Research and Behavioral Science*, *12*(1), 1–13.

Dorça, F.A., Lima, L.V., Fernandes, M.A., Lopes, C.R. (2013). Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: an experimental analysis. *Expert Systems with Applications*, *40*(6), 2092–2101.

Doroudi, S., Aleven, V., Brunskill, E. (2017a). Robust evaluation matrix: Towards a more principled offline exploration of instructional policies. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 3–12): ACM.

Doroudi, S., Thomas, P.S., Brunskill, E. (2017b). Importance sampling for fair policy selection. In *Uncertainity in artificial intelligence*. Association for uncertainty in artificial intelligence.

Ebbinghaus, H. (1885). *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Berlin: Duncker & Humblot.

Falakmasir, M.H., Pardos, Z.A., Gordon, G.J., Brusilovsky, P. (2013). A spectral learning approach to knowledge tracing. In *Proceedings of the 6th international conference on educational data mining* (pp. 360–363). International educational data mining society.

Fenza, G., Orciuoli, F., Sampson, D.G. (2017). Building adaptive tutoring model using artificial neural networks and reinforcement learning. In *2017 IEEE 17th international conference on advanced learning technologies (ICALT)* (pp. 460–462): IEEE.

Folsom-Kovarik, J., Sukthankar, G., Schatz, S., Nicholson, D. (2010). Scalable POMDPs for diagnosis and planning in intelligent tutoring systems. In *Proactive assistant agents: papers from the AAAI fall symposium*: AAAI Press.

Folsom-Kovarik, J.T. (2012). Leveraging help requests in POMDP intelligent tutoring systems. PhD thesis, University of Central Florida.

Fry, J.P. (1972). Interactive relationship between inquisitiveness and student control of instruction. *Journal of Educational Psychology*, *63*(5), 459.

Geana, A. (2015). Information sampling, learning and exploration. PhD thesis, Princeton University.

Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164.

Goel, K., Dann, C., Brunskill, E. (2017). Sample efficient policy search for optimal stopping domains. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 1711–1717). International joint conferences on artificial intelligence.

Gong, Y., & Beck, J.E. (2015). Towards detecting wheel-spinning: Future failure in mastery learning. In *Proceedings of the second (2015) ACM conference on learning@ scale* (pp. 67–74): ACM.

Green, D.T., Walsh, T.J., Cohen, P.R., Chang, Y.-H. (2011). Learning a skill-teaching curriculum with dynamic bayes nets. In *Proceedings of the twenty-third innovative applications of artificial intelligence conference* (pp. 1648–1654): AAAI Press.

Harden, R.M. (1999). What is a spiral curriculum? *Medical Teacher*, *21*(2), 141–143.

Hoiles, W., & Schaar, M. (2016). Bounded off-policy evaluation with missing data for course recommendation and curriculum design. In *International conference on machine learning* (pp. 1596–1604).

Holstein, K., McLaren, B.M., Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (Eds.) *Artificial intelligence in education* (pp. 154–168). Cham: Springer International Publishing.

Howard, R.A. (1960a). *Dynamic programming and Markov processes*. Oxford: Wiley.

Howard, R.A. (1960b). Machine-aided learning. *High speed computer system research: quarterly progress report*, *9*, 19–20.

Hsu, D., Kakade, S.M., Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, *78*(5), 1460–1480.

Hunziker, A., Chen, Y., Mac Aodha, O., Rodriguez, M.G., Krause, A., Perona, P., Yue, Y., Singla, A. (2018). Teaching multiple concepts to forgetful learners. arXiv:1805.08322.

Iglesias, A., Martínez, P., Aler, R., Fernández, F. (2006). Learning pedagogical policies from few training data. In *Proceedings of the 17th European conference on artificial intelligence workshop on planning, learning and monitoring with uncertainty and dynamic worlds*.

Iglesias, A., Martínez, P., Aler, R., Fernández, F. (2009). Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Applied Intelligence*, *31*(1), 89–106.

Iglesias, A., Martinez, P., Fernández, F. (2003). An experience applying reinforcement learning in a web-based adaptive and intelligent educational system. *Informatics in Education*, *2*, 223–240.

Javdani, S., Admoni, H., Pellegrinelli, S., Srinivasa, S.S., Bagnell, J.A. (2018). Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research*, 717–742.

Joseph, S.R., Lewis, A.S., Joseph, M.H. (2004). Adaptive vocabulary instruction. In *IEEE international conference on advanced learning technologies, 2004. Proceedings* (pp. 141–145): IEEE.

Kalyuga, S., Ayres, P., Chandler, P., Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31.

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development*, *53*(3), 83–93.

Karush, W., & Dear, R. (1967). Optimal strategy for item presentation in a learning process. *Management Science*, *13*(11), 773–785.

Käser, T., Klingler, S., Gross, M. (2016). When to stop?: towards universal instructional policies. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 289–298): ACM.

Katsikopoulos, K.V., Fisher, D.L., Duffy, S.A. (2001). Experimental evaluation of policies for sequencing the presentation of associations. *IEEE Transactions on Systems Man, and Cybernetics-Part A: Systems and Humans*, *31*(1), 55–59.

Khajah, M.M., Lindsey, R.V., Mozer, M.C. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. *Topics in Cognitive Science*, *6*(1), 157–169.

Kinzie, M.B., & Sullivan, H.J. (1989). Continuing motivation, learner control, and cai. *Educational Technology Research and Development*, *37*(2), 5–14.

Koedinger, K.R., Corbett, A.T., Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798.

Kujala, J.V., Richardson, U., Lyytinen, H. (2010). A bayesian-optimal principle for learner-friendly adaptation in learning games. *Journal of Mathematical Psychology*, *54*(2), 247–255.

Lakhani, A. (2018). Adaptive teaching: learning to teach. Master's thesis, University of Victoria.

Lan, A.S., & Baraniuk, R.G. (2016). A contextual bandits framework for personalized learning action selection. In *Proceedings of the 9th international conference on educational data mining* (pp. 424–429). International Educational Data Mining Society.

Laubsch, J.H. (1969). An adaptive teaching system for optimal item allocation. PhD thesis, Stanford University.

Legaspi, R.S., & Sison, R.C. (2002). A machine learning framework for an expert tutor construction. In *International conference on computers in education, 2002. Proceedings* (pp. 670–674): IEEE.

Leitner, S. (1972). So lernt man lernen: angewandte Lernpsychologie–ein Weg zum Erfolg. Herder.

Leyzberg, D., Ramachandran, A., Scassellati, B. (2018). The effect of personalization in longer-term robot tutoring. *ACM Transactions on Human-Robot Interaction (THRI)*, *7*(3), 19.

Leyzberg, D., Spaulding, S., Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction* (pp. 423–430): ACM.

Lin, C., & Chi, M. (2016). Intervention-BKT: incorporating instructional interventions into Bayesian knowledge tracing. In *International conference on intelligent tutoring systems* (pp. 208–218): Springer.

Lin, H.-T., Lee, P.-M., Hsiao, T.-C. (2015). Online pedagogical tutorial tactics optimization using genetic-based reinforcement learning. The Scientific World Journal.

Lindsey, R. (2014). Probabilistic models of student learning and forgetting. PhD thesis, University of Colorado at Boulder.

Lindsey, R.V., & Mozer, M.C. (2016). Predicting and improving memory retention: Psychological theory matters in the big data era. In *Big data in cognitive science* (pp. 43–73): Psychology Press.

Lindsey, R.V., Mozer, M.C., Huggins, W.J., Pashler, H. (2013). Optimizing instructional policies. In *Advances in neural information processing systems* (pp. 2778–2786).

Lindsey, R.V., Shroyer, J.D., Pashler, H., Mozer, M.C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–647.

Liu, C.L. (1960). A study in machine-aided learning. PhD thesis, Massachusetts Institute of Technology.

Lomas, D., Stamper, J., Muller, R., Patel, K., Koedinger, K.R. (2012). The effects of adaptive sequencing algorithms on player engagement within an online game. In *International conference on intelligent tutoring systems* (pp. 588–590): Springer.

Long, Y., & Aleven, V. (2016). Mastery-oriented shared student/system control over problem selection in a linear equation tutor. In *International conference on intelligent tutoring systems* (pp. 90–100): Springer.

Lumsdaine, A. (1959). Teaching machines and self-instructional materials. *Audiovisual Communication Review*, *7*(3), 163–181.

Malpani, A., Ravindran, B., Murthy, H. (2011). Personalized intelligent tutoring system using reinforcement learning. In *Proceedings of the twenty-fourth international Florida artificial intelligence research society conference* (pp. 561–562): AAAI Press.

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., Popovic, Z. (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* (pp. 1077–1084). International foundation for autonomous agents and multiagent systems.

Martin, K.N., & Arroyo, I. (2004). AgentX: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems. In *Intelligent tutoring systems* (pp. 564–572): Springer.

Matheson, J.E. (1964). Optimum teaching procedures derived from mathematical learning models. PhD thesis, Stanford University.

Matsuda, N., Chandrasekaran, S., Stamper, J.C. (2016). How quickly can wheel spinning be detected? In *International educational data mining society* (pp. 607–608).

Mazziotti, C., Holmes, W., Wiedmann, M., Loibl, K., Rummel, N., Mavrikis, M., Hansen, A., Grawemeyer, B. (2015). Robust student knowledge: Adapting to individual student needs as they explore the concepts and practice the procedures of fractions. In *Workshop on intelligent support in exploratory and open-ended learning environments learning analytics for project based and experiential learning scenarios at the 17th international conference on artificial intelligence in education (AIED 2015)* (pp. 32-40).

Mejía-Lavalle, M., Victorio, H., Martínez, A., Sidorov, G., Sucar, E., Pichardo-Lagunas, O. (2016). Toward optimal pedagogical action patterns by means of partially observable Markov decision process. In *Mexican international conference on artificial intelligence* (pp. 473–480): Springer.

Mettler, E., Massey, C.M., Kellman, P.J. (2011). Improving adaptive learning technology through the use of response times. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2532–2537). Cognitive Science Society.

Mitchell, C.M., Boyer, K.E., Lester, J.C. (2013a). Evaluating state representations for reinforcement learning of turn-taking policies in tutorial dialogue. In *SIGDIAL conference* (pp. 339–343).

Mitchell, C.M., Boyer, K.E., Lester, J.C. (2013b). A Markov decision process model of tutorial intervention in task-oriented dialogue. In Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (Eds.) *Artificial intelligence in education* (pp. 828–831). Berlin: Springer.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Mockus, J. (1994). Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, *4*(4), 347–365.

Mota, P., Melo, F., Coheur, L. (2015). Modeling students self-studies behaviors. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems* (pp. 1521–1528). International foundation for autonomous agents and multiagent systems.

Mu, T., Wang, S., Andersen, E., Brunskill, E. (2018). Combining adaptivity with progression ordering for intelligent tutoring systems. In *Proceedings of the fifth annual ACM conference on learning at scale*: ACM.

Najar, A.S., Mitrovic, A., McLaren, B.M. (2016). Learning with intelligent tutors and worked examples: selecting learning activities adaptively leads to better learning outcomes than a fixed curriculum. *User Modeling and User-Adapted Interaction*, *26*(5), 459–491.

Nelson, T.O., Dunlosky, J., Graf, A., Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*(4), 207–213.

Nijboer, M. (2011). Optimal fact learning: Applying presentation scheduling to realistic conditions. Master’s thesis, University of Groningen.

Papoušek, J., Stanislav, V., Pelánek, R. (2016). Evaluation of an adaptive practice system for learning geography facts. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 134–142): ACM.

Pavlik, P., Bolster, T., Wu, S.-M., Koedinger, K., Macwhinney, B. (2008). Using optimally selected drill practice to train basic facts. In *International conference on intelligent tutoring systems* (pp. 593–602): Springer.

Pavlik, P.I., & Anderson, J.R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101.

Pavlik, P.I., Cen, H., Koedinger, K.R. (2009). Performance factors analysis–a new alternative to knowledge tracing. In *Proceedings of the 2009 conference on artificial intelligence in education* (pp. 531–538): IOS Press.

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, *27*(1), 89–118.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in neural information processing systems* (pp. 505–513).

Pietquin, O., Daubigney, L., Geist, M. (2011). Optimization of a tutoring system from a fixed set of data. In *SLaTE 2011* (pp. 1–4).

Rafferty, A.N., Brunskill, E., Griffiths, T.L., Shafto, P. (2011). Faster teaching by pomdp planning. In Biswas, G., Bull, S., Kay, J., Mitrovic, A. (Eds.) *Artificial intelligence in education* (pp. 280–287). Berlin: Springer.

Rafferty, A.N., Brunskill, E., Griffiths, T.L., Shafto, P. (2016a). Faster teaching via POMDP planning. *Cognitive Science*, *40*(6), 1290–1332.

Rafferty, A.N., Jansen, R., Griffiths, T.L. (2016b). Using inverse planning for personalized feedback. In *Proceedings of the 9th international conference on educational data mining* (pp. 472–477). International educational data mining society.

Rafferty, A.N., LaMar, M.M., Griffiths, T.L. (2015). Inferring learners’ knowledge from their actions. *Cognitive Science*, *39*(3), 584–618.

Ramachandran, A., & Scassellati, B. (2014). Adapting difficulty levels in personalized robot-child tutoring interactions. In *Papers from the 2014 AAAI workshop*: AAAI Press.

Rau, M.A., Aleven, V., Rummel, N., Rohrbach, S. (2012). Sense making alone doesn’t do it: Fluency matters too! its support for robust learning with multiple representations. In *International conference on intelligent tutoring systems* (pp. 174–184): Springer.

Rau, M.A., Scheines, R., Aleven, V., Rummel, N. (2013). Does representational understanding enhance fluency–or vice versa? Searching for mediation models. In *Proceedings of the 6th international conference on educational data mining* (pp. 161–168). International educational data mining society.

Reddy, S., Levine, S., Dragan, A. (2017). Accelerating human learning with deep reinforcement learning. In *NIPS workshop: teaching machines, robots, and humans*.

Reddy, S., Levine, S., Dragan, A. (2018). Shared autonomy via deep reinforcement learning. arXiv:1802.01744.

Renkl, A., Atkinson, R.K., Maier, U.H. (2000). From studying examples to solving problem: Fading worked-out solution steps helps learning. In *Proceedings of the 22nd annual conference of the cognitive science society* (pp. 393–398). Cognitive Science Society.

Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review*, *69*(4), 329.

Ritter, F.E., Nerb, J., Lehtinen, E., O'Shea, T.M. (2007). *In order to learn: How the sequence of topics influences learning*. Oxford: Oxford University Press.

Rollinson, J., & Brunskill, E. (2015). From predictive models to instructional policies. In *Proceedings of the 8th international conference on educational data mining* (pp. 179–186). International educational data mining society.

Rowe, J.P. (2013). Narrative-centered tutorial planning with concurrent Markov decision processes. PhD thesis, North Carolina State University.

Rowe, J.P., & Lester, J.C. (2015). Improving student problem solving in narrative-centered learning environments: a modular reinforcement learning framework. In Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (Eds.) *Artificial intelligence in education* (pp. 419–428). Cham: Springer International Publishing.

Rowe, J.P., Mott, B.W., Lester, J.C. (2014). Optimizing player experience in interactive narrative planning: a modular reinforcement learning approach. In *Proceedings of the tenth AAAI conference on artificial intelligence and interactive digital entertainment (AIIDE-14)* (pp. 160–166): AAAI Press.

Salden, R.J., Aleven, V., Schwonke, R., Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, *38*(3), 289–307.

Sarma, B.S., & Ravindran, B. (2007). Intelligent tutoring systems using reinforcement learning to teach autistic students. In *Home informatics and telematics: ICT for the next billion* (pp. 65–78): Springer.

Sawyer, R., Rowe, J., Lester, J. (2017). Balancing learning and engagement in game-based learning environments with multi-objective reinforcement learning. In André, E., Baker, R.S., Hu, X., Rodrigo, M.M.T., du Boulay, B. (Eds.) *Artificial intelligence in education* (pp. 323–334). Cham: Springer International Publishing.

Schatten, C. (2017). Intelligent Tutoring Systems based on online learning Recommenders. PhD thesis, University of Hildesheim, Germany.

Schatten, C., Janning, R., Schmidt-Thieme, L. (2014). Vygotsky based sequencing without domain information: a matrix factorization approach. In *International conference on computer supported education* (pp. 35–51): Springer.

Segal, A., David, Y.B., Williams, J.J., Gal, K., Shalom, Y. (2018). Combining difficulty ranking with multi-armed bandits to sequence educational content. arXiv:1804.05212.

Sen, A., Patel, P., Rau, M.A., Mason, B., Nowak, R., Rogers, T.T., Zhu, X. (2018). Machine beats human at sequencing visuals for perceptual-fluency practice. In *Proceedings of the 11th international conference on educational data mining* (pp. 137–146). International educational data mining society.

Sense, F. (2017). Making the Most of Human Memory: Studies on Personalized Fact-learning and Visual Working Memory. PhD thesis, University of Groningen.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers)*, (Vol. 1 pp. 1848–1858).

Shen, S., Ausin, M.S., Mostafavi, B., Chi, M. (2018a). Improving learning & reducing time: a constrained action-based reinforcement learning approach. In *Proceedings of the 2018 conference on user modeling adaptation and personalization*: ACM.

Shen, S., & Chi, M. (2016a). Aim low: Correlation-based feature selection for model-based reinforcement learning. In *Proceedings of the 9th international conference on educational data mining* (pp. 507–512). International educational data mining society.

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

Shen, S., Mostafavi, B., Lynch, C., Barnes, T., Chi, M. (2018b). Empirically evaluating the effectiveness of pomdp vs. mdp towards the pedagogical strategies induction. In Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (Eds.) *Artificial intelligence in education* (pp. 327–331). Cham: Springer International Publishing.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354.

Smallwood, R.D. (1962). *A decision structure for teaching machines*. Cambridge: MIT Press.

Smallwood, R.D. (1968). Optimum policy regions for computer-directed teaching systems. Technical report, U.S. Department of Health, Education, and Welfare.

Smallwood, R.D. (1971). The analysis of economic teaching strategies for a simple learning model. *Journal of Mathematical Psychology*, *8*(2), 285–301.

Smallwood, R.D., & Sondik, E.J. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, *21*(5), 1071–1088.

Sondik, E.J. (1971). The optimal control of partially observable Markov decision processes. PhD thesis, Stanford University.

Spaulding, S., & Breazeal, C. (2017). Learning behavior policies for interactive educational play.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement learning: an introduction*. Cambridge: MIT Press.

Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schoelkopf, B., Gomez-Rodriguez, M. (2017). Optimizing human learning. arXiv:1712.01856.

Tenenbaum, J.B. (2000). Rules and similarity in concept learning. In *Advances in neural information processing systems* (pp. 59–65).

Theocharous, G., Beckwith, R., Butko, N., Philipose, M. (2009). Tractable pomdp planning algorithms for optimal teaching in "spais" In *IJCAI PAIR workshop*.

Theocharous, G., Butko, N., Philipose, M. (2010). Designing a mathematical manipulatives tutoring system using POMDPs. In *Proceedings of the POMDP practitioners workshop on solving real-world POMDP problems at the 20th international conference on automated planning and scheduling* (pp. 12–16): Citeseer.

Upadhyay, U., De, A., Gomez-Rodriguez, M. (2018). Deep reinforcement learning of marked temporal point processes. arXiv:1805.09360.

Van Rijn, H., van Maanen, L., van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th international conference of cognitive modeling* (pp. 110–115).

Vanlehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.

VanLehn, K. (2016). Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education*, *26*(1), 107–112.

Wang, F. (2014). Learning teaching in teaching: online reinforcement learning for intelligent tutoring. In *Future information technology* (pp. 191–196): Springer.

Wang, P., Rowe, J., Min, W., Mott, B., Lester, J. (2017a). Interactive narrative personalization with deep reinforcement learning. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 3852–3858). International joint conferences on artificial intelligence.

Wang, P., Rowe, J., Min, W., Mott, B., Lester, J. (2017b). Simulating player behavior for data-driven interactive narrative personalization. In *Proceedings of the thirteenth AAAI conference on artificial intelligence and interactive digital entertainment (AIIDE-17)* (pp. 255–261): AAAI Press.

Wang, P., Rowe, J., Mott, B., Lester, J. (2016). Decomposing drama management in educational interactive narrative: a modular reinforcement learning approach. In *Interactive storytelling: 9th international conference on interactive digital storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9* (pp. 270–282): Springer.

Welch, L.R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, *53*(4), 10–13.

Whitehill, J., & Movellan, J. (2017). Approximately optimal teaching of approximately optimal learners. IEEE Transactions on Learning Technologies.

Whitehill, J.R. (2012). stochastic optimal control perspective on affect-sensitive teaching. PhD thesis, University of California, San Diego.

Wozniak, P. (1990). Optimization of learning. Master's thesis, University of Technology in Poznan.

Zaidi, A.H., Moore, R., Briscoe, T. (2017). Curriculum Q-learning for visual vocabulary acquisition. In *NIPS workshop: visually grounded interaction and language*.

Zhou, G., Wang, J., Lynch, C.F., Chi, M. (2017). Towards closing the loop: Bridging machine-induced pedagogical policies to learning theories. In *Proceedings of the 10th international conference on educational data mining* (pp. 112–119). International educational data mining society.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Shayan Doroudi[1,2,3] · **Vincent Aleven**[4] · **Emma Brunskill**[2]

[1] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2] Computer Science Department, Stanford University, Stanford, CA 94305, USA

[3] School of Education, University of California, Irvine, CA 92697, USA

[4] Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA