Special Issue Article

JDMS

Journal of Defense Modeling and Simulation: Applications, Methodology, Technology
2022, Vol. 19(2) 173–193

DOI: 10.1177/15485129211028317
journals.sagepub.com/home/dms
SAGE

A reinforcement learning approach to adaptive remediation in online training

Randall Spain[1], Jonathan Rowe[1], Andy Smith[1], Benjamin Goldberg[2], Robert Pokorny[3], Bradford Mott[1] and James Lester[1]

Abstract

Advances in artificial intelligence (AI) and machine learning can be leveraged to tailor training based on the goals, learning needs, and preferences of learners. A key component of adaptive training systems is tutorial planning, which controls how scaffolding is structured and delivered to learners to create dynamically personalized learning experiences. The goal of this study was to induce data-driven policies for tutorial planning using reinforcement learning (RL) to provide adaptive scaffolding based on the Interactive, Constructive, Active, Passive framework for cognitive engagement. We describe a dataset that was collected to induce RL-based scaffolding policies, and we present the results of our policy analyses. Results showed that the best performing policies optimized learning gains by inducing an adaptive fading approach in which learners received less cognitively engaging forms of remediation as they advanced through the training course. This policy was consistent with preliminary analyses that showed constructive remediation became less effective as learners progressed through the training session. Results also showed that learners' prior knowledge impacted the type of scaffold that was recommended, thus showing evidence of an aptitude–treatment interaction. We conclude with a discussion of how AI-based training can be leveraged to enhance training effectiveness as well as directions for future research.

Keywords

Tutorial planning, adaptive remediation, reinforcement learning, adaptive instructional systems

1. Introduction

Artificial intelligence (AI) will play a central role in modernized training systems for the military. The transformative potential of AI-enhanced training and education is made possible by advancements in a wide array of capabilities, including natural language processing,[1] multimodal learning analytics,[2] automated scenario generation,[3] and adaptive decision-making,[4] that can be leveraged to create engaging and effective training experiences that are tailored to individual learners and teams. Numerous studies have shown that when designed effectively, adaptive instructional systems (AISs), such as intelligent tutoring systems, can be as effective as a skilled human tutor at supporting student learning.[5] Still, determining what, how, and when to adapt instruction is a central question facing AIS developers. This challenge stems in part from the wide range of pedagogical strategies and tactics that can be implemented in AISs, as well as a lack of empirically grounded guidance about the relative contribution of different adaptive interventions on learning outcomes.[6] Another challenge is the lack of tools that AISs can use to develop theory-driven frameworks and workflows to

[1]Center for Educational Informatics, North Carolina State University, USA
[2]U.S. Army Combat Capabilities Development Center – Soldier Center, USA
[3]Intelligent Automation Inc., USA

Corresponding author:
Randall Spain, Department of Computer Science, Center for Educational Informatics, North Carolina State University, Campus Box 8206, 890 Oval Drive, Raleigh, NC 27695-8206, USA.
Email: rdspain@ncsu.edu

explore the effectiveness of different pedagogical plans and decisions.

Tutorial planning is a critical component of AISs, controlling how instructional feedback and support are structured and delivered to learners. Tutorial planners utilize a set of formalized pedagogical rules to drive decisions about how an AIS sequences instruction and delivers scaffolding. These rules are typically based on learning theories and provide general guidance on when to scaffold learners, what type of scaffolding to provide, and how scaffolding should be carried out within an AIS. Despite their importance, however, tutorial planners often suffer from three important limitations. Firstly, creating tutorial planners is expensive, requiring labor-intensive knowledge engineering processes that involve close collaboration between subject matter experts, education experts, and software developers.[7] Secondly, once a tutorial planner has been created, it typically remains fixed; it does not improve or change over time unless manually updated by an expert. Thirdly, tutorial planners that utilize production rules to make decisions about adaptive instruction are poorly suited for reasoning about the inherent uncertainty in learning, such as determining how learners respond to different types of tutorial strategies or perform on assessments.

Recent advances in machine learning (ML) have introduced opportunities to reduce the authoring burden of AISs by inducing data-driven models of tutorial planning directly from data generated by student interactions with an AIS.[4,8,9] The induced models are designed to automatically control how pedagogical support is structured and delivered to learners at run-time to create personalized learning experiences. Leveraging decision-theoretic frameworks, such as Markov decision processes (MDPs), these models explicitly account for the inherent uncertainty in how learners respond to different types of tutorial strategies and tactics, and automatically generate and refine tutorial planning policies that seek to optimize learning outcomes.

This paper presents results from work to devise and investigate data-driven tutorial planning policies using reinforcement learning (RL) to provide learners with adaptive remediation. Toward this goal, we describe a human subjects' study that collected training interaction data from over 500 learners who completed a 90-minute adaptive online training course that taught doctrinal concepts associated with counterinsurgency (COIN) and stability operations while receiving adaptive remediation. We describe the online adaptive training course, which was authored using the Generalized Intelligent Framework for Tutoring (GIFT), an open-source framework for developing and evaluating AISs that includes a suite of tools for devising and testing theory-driven pedagogical interventions. The remediation activities were based on Chi's Interactive, Constructive, Active, Passive (ICAP) framework of cognitive engagement, which distinguishes between different levels of cognitive engagement that a learner may have with instructional material.[10] We describe preliminary results examining the effectiveness of the ICAP-inspired remediation and the learner interaction data. Following this, we discuss the results of a second study that induced RL policies using the data collected from the human subjects' study and review how the policies can be used to scaffold learning by providing students with tailored remediation in an online training course. Our findings point toward the significant promise of using data-driven techniques to model adaptive scaffolding in AISs to enhance learning.

1.1. Adaptive instructional systems

AISs guide student learning experiences by tailoring instruction based on the individual goals, needs, and preferences of learners in the context of domain learning objectives.[11] Adapting instructional content and support to the needs of learners can be accomplished in many different ways in AISs.[12] For example, AISs can present learners with different types of instructional feedback,[13] hints,[14] and faded worked examples[15] based on a learner's demonstrated level of mastery in order to improve learning outcomes and learning experiences. AISs can also present new learning content that is tailored to a learner's current skill in order to keep them motivated and engaged.[16]

A significant challenge in the design of AISs is determining how pedagogical interventions should be structured and delivered to learners. Decisions about what to adapt, how to adapt, and when to adapt are often formalized as a set of production rules within an AIS's pedagogical model. Although these rules are often rooted in learning theory, the guidance they provide may be too coarse to provide effective remediation for many learners. Further, because the rules are manually encoded, they can be burdensome to update, adding to AIS development time and sustainment costs.[17,18]

Data-driven tutorial planning offers a method for automatically formulating how pedagogical support and scaffolding are delivered to learners to create personalized learning experiences.[8,19,20] Data-driven tutorial planners use observations about student actions and their impact on training performance to determine when and how to provide pedagogical support. This bottom-up, data-driven approach complements traditional top-down approaches to the design of tutorial planners, in that the system can observe which actions and pedagogical decisions are most effective and update its production rules accordingly. In particular, RL techniques have shown promise for automatically inducing tutorial planning rules that optimize student learning outcomes and do not require pedagogical

rules to be manually programmed or demonstrated by expert tutors.

1.2. Reinforcement learning

RL is a family of ML techniques that center on creating software agents that perform actions in a stochastic environment to optimize numerical reward.[21] In classical RL, an agent seeks to learn a policy for selecting actions in an uncertain environment in order to accomplish a goal. The environment is characterized by a set of states and a probabilistic model describing transitions between those states. The agent is capable of observing the environment's state and using its observations to guide decisions about which actions to perform. In contrast to supervised ML, RL agents are not provided with external instruction about which actions to take. Instead, the environment produces rewards that provide positive or negative feedback about the agent's actions. The agent's task is to utilize the reward signal in order to learn a policy that maps observed states to actions and maximizes its total accumulated reward. RL problems are often formalized using MDPs. MDPs provide a principled mathematical framework for modeling stochastic control problems, such as tutorial planning, which involve sequential decision-making under uncertainty.

Over the past decade, RL- and MDP-based techniques have been the subject of growing interest in the AIS community.[4,8,22–26] This work has emphasized probabilistic models of behavior, as opposed to explicit models of cognitive states, in order to analyze student learning. For example, Chi et al.[23] used MDPs to model tutorial dialogues, devising pedagogical tactics directly from student data in the Cordillera physics tutor. Rowe and Lester[8] utilized modular RL to induce policies for narrative-centered tutorial planning in an educational game for middle school microbiology education. More recently, Ausin et al.[22] investigated deep RL techniques to model tutorial decision-making in an intelligent tutoring system for undergraduate logic proofs. Complementary work investigated partially observable MDPs to model tutorial planning, yielding novel approaches for compactly representing MDP state representations.[27,28]

The driving motivation of the current study was to demonstrate how RL could be used to create a set of tutorial planning policies that could be implemented in an AIS to provide learners with adaptive remediation as they completed an online training course. To address this objective, we engaged in two key activities. The first activity involved conducting a human subjects' study to collect a rich dataset of learning interaction log data from users who completed an online adaptive training course about COIN and stability operations implemented with the GIFT framework. During the course, learners were presented with feedback and remedial content that was designed to elicit different levels of cognitive engagement. The goal of the second activity was to utilize RL to create a set of tutorial planning policies trained on the learner interaction log data collected from the human subjects' study, which could be used to provide learners with adaptive remediation in an AIS. We refer to these two activities as Study 1 and Study 2, respectively, throughout the remainder of the paper.

2. Study 1 – human subjects' study to develop a training dataset

RL techniques are data-intensive, so in order to collect sufficient data to induce RL-based tutorial policies we devised a study to collect learning outcome and learner interaction log data from a sample of users who completed an online adaptive training course. The online training course was designed to meet three objectives: (a) it contained numerous opportunities for learners to receive instructional remediation; (b) it could be deployed through online crowdsourcing platforms to facilitate broad distribution to many learners; and (c) the course initially enacted an exploratory (i.e., random) remediation policy in order to broadly sample the space of possible pedagogical decisions to produce a dataset for inducing RL-based policies for adaptive tutorial planning.

The instructional remediation activities presented in the adaptive training course were modeled after the ICAP framework of cognitive engagement, which identifies four levels of learner engagement activities: Interactive, Constructive, Active, and Passive, each of which has associated learning behaviors and cognitive processes.[10] *Interactive* activities involve learning that centers on back-and-forth discourse between a student and teacher, student and machine, or student and peer. *Constructive* activities refer to learning that involves the creation of novel artifacts (e.g., written summaries, concept maps) to reify one's understanding of a subject or topic, producing outputs that go beyond content that was previously presented. The *active* category refers to learning that involves pointing, highlighting, note taking, or other forms of physical engagement that exceed passive learning in its impact. *Passive* learning refers to learning that involves listening or viewing direct instruction on a topic.

The ICAP model predicts that as students are more actively engaged with learning material, their learning will increase (i.e., passive < active < constructive < interactive).[29] We aimed to investigate whether our results would support the predictions of the ICAP model or whether trade-offs in terms of remediation effectiveness would emerge as learners advanced through a 90-minute online training course. If trade-offs in remediation effectiveness are observed during the training course, then the results

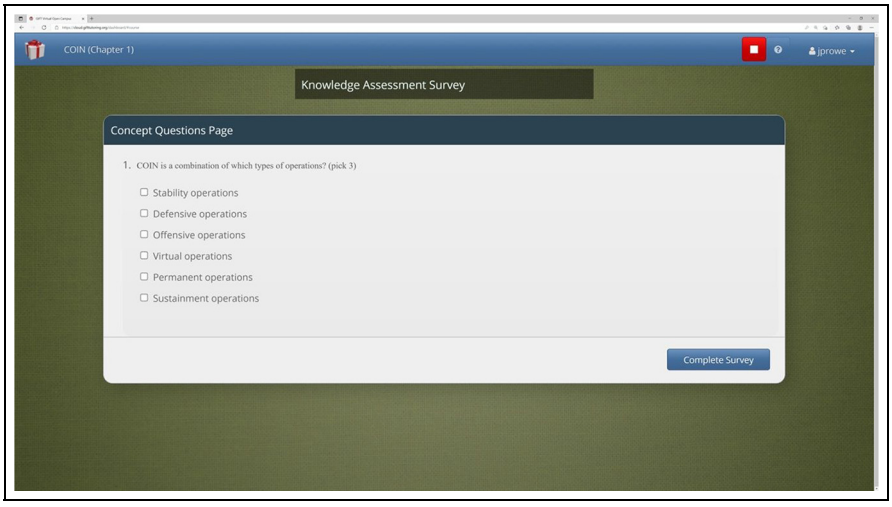

Figure 1. Single-concept recall quiz question from the online training course.

would highlight the potential for using data-driven techniques, such as RL, to develop pedagogical policies that provide tailored forms of remediation. We also aimed to investigate the feasibility of using the dataset to devise data-driven tutorial planning policies. Specifically, the first study was guided by the following research questions.

(1) Did participants demonstrate positive learning gains as a result of completing the adaptive training course?
(2) Did the dataset contain a sufficiently large number of remediation instances to support data-driven tutorial planning? How many instances of remediation, on average, did learners receive while completing the course?
(3) Which forms of ICAP-inspired remediation were most effective for helping learners overcome an impasse? Are there any trade-offs in terms of remediation effectiveness across successive remediation attempts?

2.1. Method

2.1.1. Adaptive training course. The online training course was created and deployed using GIFT, an open-source software framework for designing, deploying, and evaluating adaptive training systems.[30,31] The course's training content built upon materials from the UrbanSim Primer, a self-paced hypermedia training course that provides direct instruction on the themes, terms, and principles associated with leading COIN and stability operations.[32] The course was organized into four chapters, each of which included a series of short videos, multiple-choice quiz questions, remedial training activities, and glossary terms that aimed to teach learners about the tenets and principles of leading COIN operations.[33] The instructional videos were each approximately 90 seconds in length and covered topics such as ''Identifying the center of gravity in COIN operations,'' ''Defining intelligence preparation for the battlefield,'' and ''Understanding lines of effort in COIN operations.'' The multiple-choice quiz questions, which were presented to learners after viewing an instructional video, consisted of single- or multi-concept review questions that aligned with the course's learning objectives. Single-concept review questions required learners to recall and apply concepts presented within the video lesson they had just viewed (see Figure 1). Multi-concept review questions required learners to demonstrate a deeper understanding of the course material by integrating concepts from multiple lesson videos.

The remediation interventions included in the course were structured according to the ICAP framework of cognitive engagement and presented learners with either passive, active, or constructive remediation activities if they incorrectly answered a quiz question. *Passive* remediation required learners to passively read the narrated content in text format that was just presented in the lesson video. After they finished reading the content, they returned to the previously missed quiz question and attempted to answer it correctly. *Active* remediation required learners to read the narrated content and actively highlight the portion of text that answered the quiz question that was just missed. *Constructive* remediation content required learners to read the narrated content and constructively summarize the answer to the quiz question that had just been missed in their own words (Figure 2). The active and constructive

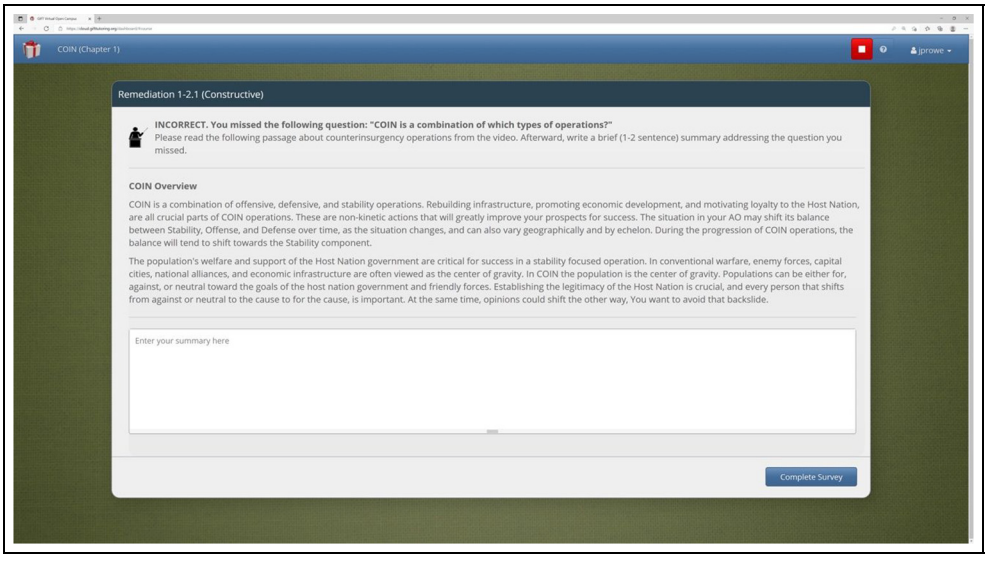

Figure 2. Constructive remediation activity.

remediation prompts also included expert highlighting/summaries and asked students to rate the similarity of their responses to the expert responses using a five-point Likert scale (from 1 – not similar to 5 – very similar). The system could also provide no remediation after a missed quiz question, in which case learners would receive a simple feedback message stating they incorrectly answered the question. Interactive remediation activities, which typically consist of tutorial dialogue between a learner and an AIS, were not included in this study because the feature had not yet been integrated with GIFT's remediation framework.

The adaptive training course presented remediation activities to learners whenever they answered a recall quiz question incorrectly. Upon completing the remediation activity, participants attempted to correctly answer the previously missed quiz question again. The order of the answer choices was randomized upon each successive question attempt. Students continued to receive remediation until they demonstrated concept mastery (i.e., correctly answered the quiz question; Figure 3). The course utilized a random policy to determine the type of remediation participants received after each missed multiple-choice quiz question. Thus, participants were not assigned to specific remediation conditions (i.e., only passive, only active, only constructive) but instead received a random combination of all types of remediation throughout the course.

In total, 12 instructional videos and 39 multiple-choice quiz questions were distributed throughout the four-chapter training course. The online training course also included a set of web-based surveys designed to collect information about the participants' age, education, interest in COIN operations and military science topics, and goal orientation, as well as parallel forms of a 12-item pre- and posttest that measured knowledge of COIN topics, terminology, and principles.[33]

2.1.2. Participants. To address the goal of collecting data from a large sample of learners, we recruited participants through Amazon's Mechanical Turk (MTurk) platform. We collected completed training data from 533 participants (42% female, ages ranged from 18 to 65). To be eligible, participants had to be at least 18 years of age, reside in the USA, and have completed at least 95% of the tasks through MTurk that they previously accepted to complete (e.g., 5% dropout rate). Participants were compensated US$8 for completing the training course, which took on average 90 minutes to complete. Analysis of pretest scores revealed that participants answered approximately one third of the pretest questions correctly (M = .35, SD =.18), suggesting they had low prior knowledge of the concepts covered in the training course prior to taking the course.[33]

2.1.3. Procedure. A description of the study was posted on the MTurk website where participants were able to read a short description of the study. Participants who were interested in completing the study were directed via a hyperlink to read and electronically sign an informed consent. Afterwards, participants proceeded to the training course, which was hosted on the cloud-based instance of GIFT. The course began with a general welcome message. Following this introduction, participants completed a

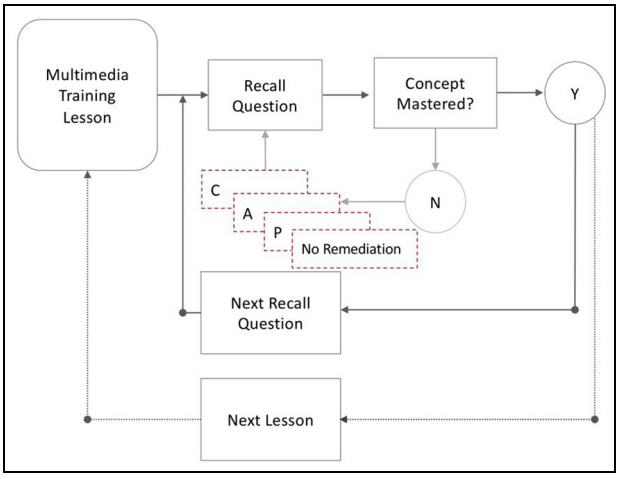

Figure 3. Instructional workflow for the online course and remediation delivery.

demographic questionnaire that gathered information about their age, years of education, and familiarity with COIN topics and concepts. Then, they completed a goal orientation questionnaire that measured task-based and intrinsic motivation to learn,[34] followed by a 12-item pretest that measured prior knowledge of COIN principles and terminology.[33]

After completing the pre-training surveys, participants began the adaptive online COIN training course. Participants watched a series of narrated videos that covered lesson topics such as the importance of population support, processes for intelligence gathering, and issues in successful COIN operations. After each video, participants answered a series of multiple-choice quiz questions that consisted of single- or multi-concept review items that aligned with the content covered in the video. An incorrect response to a quiz question resulted in participants receiving an ICAP-inspired remediation activity that required them to either passively, actively, or constructively engage with the training content, unless the system selected to provide no remediation. After completing the remediation exercise, participants were gated back to the previously attempted quiz question. Learners continued to receive remediation until they correctly answered the multiple-choice quiz question. This meant that on some attempts, participants received multiple rounds of remediation before they advanced to the next question. The type of remediation students received varied randomly across attempts.

Upon finishing the final video lesson and quiz question, participants completed a series of post-training surveys that included a multiple-choice posttest to measure retention of the concepts and principles presented in the training and a short questionnaire to collect opinions about the training experience. After completing these activities, participants received a unique completion code that could be used to verify course completion through the MTurk website and were thanked for their participation.

2.2. Results

To evaluate participants' interaction behaviors with the course content, the pretest, posttest, survey, and learner interaction log data were recorded and analyzed. The learner interaction log data consisted of a timestamped record of learner actions, course states, and pedagogical requests made by GIFT as well as information regarding how many times learners received remediation, how long they spent interacting with the different forms of remediation, correctness of responses, and remediation helpfulness ratings. We conducted a set of preliminary analyses to identify how well participants performed on the pre- and posttest assessments, how often learners received remediation, and which forms of remediation were most effective at helping learners overcome an impasse. The results presented below summarize and expand analyses that are presented in previous research.[35,36]

2.2.1. Learning gains. Participants' pretest and posttest scores were analyzed to determine if the course was effective in promoting participants' knowledge of COIN concepts, terminology, and principles. Pretest and posttest scores were calculated by summing the total number of correct responses on the 12-item tests. Results showed that posttest scores (M = 8.68, SD = 2.50) were significantly higher than pretest scores (M = 4.35, SD = 2.25), $F(1, 482) = 1590.88$, $p < .001$, suggesting that the course was successful in meeting its instructional objectives. In addition to examining differences in pre- and posttest scores, we

Table 1. Total count of remediations presented over the duration of the course and for remediation attempts.

Remediation type	Total count of remediations	1st attempt	2nd attempt	3rd attempt
Constructive	2098	1529	304	88
Active	2147	1544	323	105
Passive	465	315	78	30
None	487	365	50	30
Total	5197	3753	755	253

also examined participants' normalized learning gains (NLGs), which were calculated to account for participants' pretest performance. *NLGs* reflect an individuals' relative learning in a course. These scores are derived by calculating the ratio of actual improvement from pre- to posttest over the maximum possible improvement.[37] NLG values range from –1 to +1 with, values below 0 indicating learning losses (i.e., students performed worse on the posttest than pretest), 0 indicating no gains, and positive values indicating higher learning gains. NLG allows for fair comparison among learners who scored high on the pretest with learners who scored low on the pretest by standardizing gains. Results showed participants made significant learning gains from completing the course, improving their posttest scores by more than 57% of the total possible gains available.

2.2.2. Remediation statistics. Because RL techniques are data-intensive, our second set of analyses aimed to determine whether the dataset contained a sufficiently large number of remediation instances to support data-driven tutorial planning. Results showed the dataset included a total of 5197 instances of remediation. On average, learners received 10 instances of remediation while completing the online course (SD = 12.60; range 1–113). Given the wide range of remediation instances, we conducted a closer examination of the remediation distribution data and found that 90% of participants received fewer than 20 instances of remediation while completing the training course. Analyses also showed that although the course was designed to implement a uniform random control policy, a software error resulted in approximately 40% of all remediation interventions being constructive interventions, 40% being active, 10% passive, and 10% no remediation. Table 1 presents the total count of each remediation type as well as the how many instances corresponded to the first three remediation instances.

2.2.3. Remediation effectiveness. Next, a set of exploratory analyses were conducted to identify which form of remediation was most effective at helping learners overcome an impasse on a missed recall question. The ICAP model predicts that constructive remediation should be more effective than active remediation at helping students overcome an impasse, and that active remediation should be more effective than passive remediation. However, there could be trade-offs between these different forms, because higher levels of cognitive processing require additional time and effort on the part of learners, particularly as learners progress through a training course. For this set of analyses, *remediation effectiveness* was operationally defined as the proportion of cases in which participants correctly answered a recall question after receiving a given type of remediation (constructive, active, passive, none). Remediation effectiveness was calculated for the first, second, and third remediation instances delivered following missed attempts on a given recall question. By examining remediation effectiveness over successive attempts, we aimed to identify trade-offs in remediation effectiveness that may have occurred as learners transitioned from one unsuccessful remediation attempt to another. A series of z-tests were computed to examine the effectiveness of the remediation types (Table 2).

Results showed learners were more likely to correctly answer a previously missed quiz question after the first remediation if they received constructive remediation (.84) compared to active remediation (.80; $z = 2.45$, $p < .05$, two tailed), that active remediation (.84) was more effective than passive (.63; $z = 6.57$, $p < .01$, two tailed), and passive remediation (.63) was more effective than no remediation (.53; $z = 2.71$, $p < .01$, two tailed; Figure 4). For cases in which participants received two rounds of remediation before correctly answering a recall question, results showed that constructive was not more effective than active remediation ($z = 1.10$, $p = .27$, two tailed), but that active remediation was more effective than passive remediation ($z = 2.77$, $p < .01$, two tailed). Interestingly, presenting no remediation appeared to be more effective than presenting passive remediation; however, this observed effect did not reach statistical significance ($z = 1.47$, $p = .14$, two tailed). Finally, for cases in which participants correctly answered a recall question after the third remediation attempt, active remediation appeared to be the most effective form of remediation, followed by constructive remediation. Results showed no difference in remediation effectiveness between constructive and active remediation ($z = 3.04$, $p < .05$).

2.3. Discussion

The goal of the first study was to capture learning outcomes as well as learning trace-data from a sample of learners who completed an online adaptive training course that could be used for evaluating student learning behaviors and developing RL-based tutorial policies. Learner

Table 2. Remediation effectiveness comparisons among remediation types.

First remediation attempt							
	Constructive	Active	Passive	None	Comparison	*z*	*p*
Effective on 1st attempt	1278	1238	199	193	C vs A	2.45	0.014
Total instances provided	1529	1544	315	365	A vs P	6.57	0.000
Effectiveness ratio	0.84	0.80	0.63	0.53	P vs N	2.71	0.007
Second remediation attempt							
	Constructive	Active	Passive	None	Comparison	*z*	*p*
Effective on 2nd attempt	212	212	38	31	C v A	1.10	0.273
Total instances provided	304	323	78	50	A vs P	2.77	0.006
Effectiveness ratio	0.70	0.66	0.49	0.62	P vs N	1.47	0.141
Third remediation attempt							
	Constructive	Active	Passive	None	Comparison	*z*	*p*
Effective on 3rd attempt	44	61	8	8	C vs A	1.12	0.261
Total instances provided	88	105	30	30	A vs P	3.04	0.002
Effectiveness ratio	0.50	0.58	0.27	0.27	P vs N	0	1

C: constructive; A: active; P: passive; N: none.

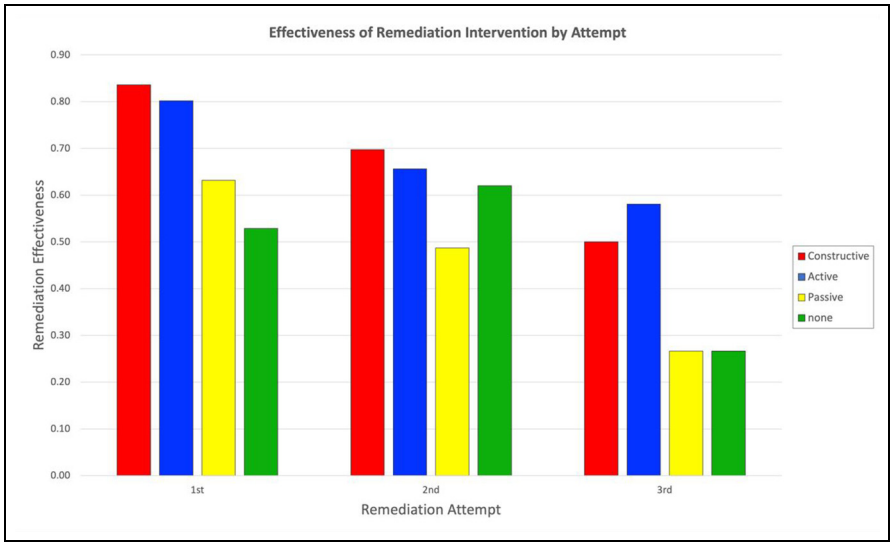

Figure 4. Remediation effectiveness across remediation attempts.

interaction data were collected from 500 participants. Analyses indicated the dataset contained over 5000 instances of remediation and that participants demonstrated learning gains as a result of completing the course. Results also showed that the ICAP-inspired remediation presented to learners broadly follows trends predicted by the ICAP model concerning instructional effectiveness and student cognitive engagement. However, results also suggest that the effectiveness of ICAP-inspired remediation may change over time and under different conditions, pointing toward the need for adaptive tutorial policies to control how and when different forms of remediation are

delivered to learners. In particular, the effects observed in Study 1 suggest that providing remediation that requires less cognitive engagement (e.g., active rather than constructive remediation, or passive rather than constructive) could be an effective instructional approach, particularly when a learner requires multiple instances of remediation to master a specific learning objective or knowledge component. It is important to note that the present study only examined remediation effectiveness. If AIS designers were interested in maximizing multiple rewards, such as learning and engagement, then additional features about the learner such as their pre-existing knowledge or performance and interaction patterns with previous remediation activities could impact which form of remediation would be most effective. These findings set the stage for investigating the application of RL techniques to automatically induce tutorial policies for controlling how and when ICAP-inspired remediation is delivered to learners.

3. Study 2 – inducing Markov decision process-based tutorial policies with reinforcement learning

Building on the preliminary analysis and the dataset gathered from Study 1, we next utilized offline RL techniques to investigate the creation of data-driven tutorial policies for controlling ICAP-inspired remediation within the online training course. The purpose of the investigation was to address three primary research questions.

(1) To what extent does RL-based tutorial planning reproduce the pairwise ordering predicted by the ICAP framework, that is, constructive remediation interventions are preferred to active remediation interventions, which are in turn preferred to passive remediation interventions?
(2) How do alternative representations for encoding MDP states and rewards in RL-based tutorial planners impact induced tutorial policies to control ICAP-inspired remediation within an adaptive online course for COIN training?
(3) Under what circumstances do RL-based tutorial policies for ICAP-inspired remediation select constructive remediation? Active remediation? Passive remediation? No remediation?

3.1. Method

To formalize tutorial planning as a RL task, we adopted a *tabular policy representation*, which utilizes discrete state and action representations to encode the model in terms of a MDP. We utilized a tabular representation due to the relatively small size of the training dataset, which is common in applications of RL with educational data.[8,9,23] In addition, tabular representations increase the simplicity of integration with run-time AIS platforms like GIFT, which was an eventual objective of the research. We investigated several alternate formulations of the MDP model to investigate how different combinations of state features and reward influence tutorial policies induced from the training dataset. For each MDP formulation, a shared action set representation was utilized that consisted of four actions: (a) constructive remediation; (b) active remediation; (c) passive remediation; and (d) no remediation. Tutorial planning was modeled as an episodic task, where each student log corresponded to a single RL episode. Each decision point for the MDP coincided with the occurrence of a missed recall question in the course, which triggered a tutorial decision about how to deliver feedback and remediation to the learner. MDP states were computed at each of these decision points, and state transitions corresponded to changes in state between two successive decision points, or between a decision point and the terminal state of the session. Rewards were computed at the conclusion of a session based on the participant's performance on the COIN content pretest and posttest assessments.

To induce tutorial policies, we utilized a certainty equivalent approach, a simple form of model-based RL that has been used widely in research on intelligent tutoring systems.[8,9,23,38] In this approach, the MDP state transition model P and reward model R are estimated directly from the training dataset. After estimating the state transition and reward models, we applied the *value iteration* algorithm to estimate the value-function that quantified the estimated reward associated with each possible combination of state–action pairs. Next, we devised an action-selection policy that chose actions according to a greedy strategy with respect to the induced value-function, thus producing an ''optimal'' policy for controlling ICAP-inspired remediation decisions.[21] (Value iteration yields a theoretically optimal policy for given a state transition model and reward model if the MDP observes the Markov property. The Markov property holds for a decision process when future states are independent of past states and actions.)

Throughout all of the analyses, a discount rate $\gamma = 0.9$ was utilized. All policies were induced using the PyMDP software toolkit. Developed at North Carolina State University, PyMDP is a simple Python library for implementing MDP-based models of tutorial planning that can be induced from learner interaction log data and pre–post assessment data.[39] For small state representations, such as those described in this study, policies took less than 1 minute to be induced. Policies were encoded as a direct mapping between MDP states and tutorial actions.

3.1.1. MDP state representation. In devising the RL models, we investigated four different state representations and two different reward models. Each state representation consisted of a unique combination of the following four discrete state features.

- **Pretest_Score_Level:** a binary indicator of a learner's performance on the COIN knowledge pretest. Pretest_Score_Level was calculated by performing a median split on all learner pretest scores and assigning learners with less than median score to a value of 0, and learners with equal or greater than median score to a value of 1.
- **Current_Chapter:** an ordinal variable that denoted which chapter of the course that the learner was in during a decision point about delivering ICAP-inspired remediation. This feature could have four possible values: 1, 2, 3, or 4.
- **Remediation_Count_High_Low:** a binary indicator denoting whether a learner received a high or low amount of remediation thus far based on average rates of remediation estimated from all participants in the MTurk study. Specifically, we estimated that learners received approximately two instances of remediation per chapter on average by taking the mean number of remediation instances per student and dividing by the total number of chapters in the online course. Thus, this feature was calculated by evaluating whether the total number of remediation instances the learner had received thus far was less than 2 * Current_Chapter. If so, it had a value of 0. Otherwise, it had a value of 1.
- **Previous_Remediation_Type:** a nominal variable describing what type of ICAP-inspired remediation was most recently delivered prior to the current remediation decision point. This feature could have four possible values, each representing a different type of remediation: Constructive, Active, Passive, or None.

The four state representations were additive; they built upon one another by concatenating additional state features to enrich the information utilized to determine what form of remediation to deliver at a particular decision point in the training course. The simplest state representation consisted of a single feature: Pretest_Score_Level. This was a static feature with a value that was determined prior to the student engaging with the online course; it was based entirely on the learner's pretest score, and it did not change during the training session. This state representation is denoted as `PretestOnly`.

The second state representation that we investigated consisted of a pair of features: Pretest_Score_Level and Current_Chapter. This state representation accounted for both the learner's prior knowledge and current progress within the online training course. Although the Pretest_Score_Level feature did not change during the session, the Current_Chapter feature incremented each time the learner progressed to a new lesson that coincided with a subsequent chapter. This state representation is denoted as `Pretest + Chapter`.

The third state representation consisted of three features: Pretest_Score_Level, Current_Chapter, and Remediation_Count_High_Low. This state contributed additional information about how often the learner had received remediation in the course thus far. The Remediation Count feature could fluctuate up and down depending on how many recall questions the learner missed in the course, and thus, how many instances of remediation he/she received in the course thus far. This state representation is denoted as `Pretest + Chapter + RemCount`.

The fourth state representation consisted of all four features: Pretest_Score_Level, Current_Chapter, Reme diation_Count_High_Low, and Previous_Remediation_ Type. This provided the richest state information, and thus the greatest amount of information, to inform pedagogical strategy decisions about remediation. The initial value of the Previous_Remediation_Type feature was None, and its value changed each time a learner missed a recall question and the system selected a form of ICAP-inspired remediation to deliver. This state representation is denoted as `AllFeatures`.

The selection of these specific state representations and features was informed by previous research,[35,36] as well as the results of Study 1. Results suggested that the effectiveness of ICAP-inspired remediation may change over time and under different conditions. Remediation that requires less cognitive engagement may be an effective instructional approach when a learner requires multiple instances of remediation or has reached a later stage of the online training course. We sought to devise empirically based state representations to investigate alternative designs of the MDP-based pedagogical model, enabling the creation of adaptive tutorial policies to control how and when different forms of remediation are delivered to learners.

Specifically, the PretestOnly state representation provided a personalized remediation policy that would not adapt its remediation behavior at run-time. Pretest scores are an important predictor of student learning outcomes. The PretestOnly state representation provided a means for inducing a baseline policy that would be adaptive to individual learners but not contextually adaptive during the online training course. In contrast, features such as Current_Chapter and Remediation_Count_High_Low could change in value during a student's interaction with the online training course. Previous research has shown that student engagement with ICAP-inspired remediation

decreases over time, and an increased number of remediation activities is associated with reduced learning.[35,36] Therefore, Current_Chapter and Remediation_Count_High_Low served as empirically based features to inform run-time decisions about remediation based upon a student's current learning conditions. The fourth feature, Previous_Remediation_Type, was selected to provide a mechanism for introducing variety to the types of remediation activities delivered to learners for the purpose of maintaining student interest and motivation. Previous_Remediation_Type was included as the last feature in the additive state representation. Although its inclusion was theoretically based, we did not have empirical evidence from Study 1 to directly link remediation variety with student interest.

Discretized features were used to minimize issues related to data sparsity, which are an important factor in tabular RL with policies induced from learning interaction data. Specifically, we used a binary split to encode the Pretest_Score_Level and Remediation_Count_High_Low features in order to increase the number of data points associated with each possible state of the MDP. Although the state space for MDPs in this work is highly constrained, it is in line with prior research on RL-based tutorial planning in advanced learning technologies.[8,9,23] An alternative approach to developing state representations in RL-based tutorial planning is to utilize automated feature selection, which involves developing a large pool of candidate features and then algorithmically selecting a small subset for use in the MDP.[9] We did not use this approach in the current study, because our aim was to examine how different combinations of empirically and theoretically based state features impact RL-based tutorial policies. Further investigation of richer state representations is a promising direction for future work, and it calls for utilizing function approximation techniques, which we do not consider here.

3.1.2. MDP rewards. In devising the RL tutorial policies, two different rewards models were utilized: `FullNLG` and `ChapterNLG`. Both reward models were based upon participants' NLGs. NLG was computed for each learner. The `FullNLG` reward model computed pretest and posttest scores based upon sum performance across the entire 12-item test. This metric is commonly used as a reward model in RL-based tutorial planning in advanced learning technologies.[8,23,25,26] A strength of this model was that it leveraged the full set of available items on the pretest and posttest to estimate learner knowledge. A weakness of this approach was that the connection between pretest/posttest score and a specific instance of remediation could be weak. For example, if a learner received remediation during Chapter 1 of the online training course, but demonstrated pre–post improvement only on items pertaining to Chapter 4, then a policy that optimized the `FullNLG` reward model would utilize the improvement in Chapter 4 items to reinforce pedagogical strategy decisions about remediation related to Chapter 1 concepts. This could introduce noise related to credit assignment within RL.

The `ChapterNLG` model was designed to address this issue by introducing a more granular representation of NLG that was connected to specific remediation decisions enacted during learner interactions with the training course. The `ChapterNLG` reward model computed proportional improvements in test scores, but the posttest and pretest scores were computed based upon performance across only those items that aligned with course chapters on which tutorial decisions about ICAP-inspired remediation were enacted. For example, if a student missed recall questions in Chapters 1 and 4 of the online course, then `ChapterNLG` would be computed based upon pretest and posttest scores on the subset of items that aligned with Chapters 1 and 4; gains or losses on items aligned with Chapters 2 and 3 would be ignored. This enhanced the connection between the effects of instructional remediation and reward, but it came at the cost of estimating learner knowledge from a smaller number of test items, raising potential reliability issues.

Both `FullNLG` and `ChapterNLG` reward were assigned at the conclusion of an episode corresponding to a single learner's training session with the online training course. They were formatted as a real-valued number with a range of -1 to $+1$. No incremental rewards were given within a training episode. Thus, `FullNLG` and `ChapterNLG` serve as the only optimization criteria considered for inducing tutorial policies during RL in this work.

3.2. Results

We induced eight different policies corresponding to different combinations of the four alternate state representations and two reward models described above. We examine a subset of the induced policies here in-depth, and we report descriptive statistics on remediation strategies recommended by each policy trained using the training interaction dataset.

3.2.1. Research question 1: to what extent does MDP-based tutorial planning reproduce the pairwise ordering predicted by the ICAP framework? To investigate this question, we examined policies induced for the `Pretest + Chapter + RemCount` state representation. Tables 3 and 4 show the full induced policies for this state representation and the `ChapterNLG and FullNLG`

Table 3. Induced policy for `Pretest + Chapter + RemCount` state representation and `ChapterNLG` reward.

Action 1: no remediation	Action 2: passive	Action 3: active	Action 4: constructive
[0,3,1] [0,4,1]	[0,4,0] [0,1,1] [1,1,0]	[0,2,0] [0,3,0] [1,2,0] [1,3,0] [1,4,0] [1,1,1] [1,2,1] [1,3,1] [1,4,1]	[0,1,0] [0,2,1]

Table 4. Induced Policy for `Pretest + Chapter + RemCount` state representation and `FullNLG` reward.

Action 1: no remediation	Action 2: passive	Action 3: active	Action 4: constructive
[0,3,1] [0,4,1]	[0,4,0] [0,1,1] [1,1,0]	[0,3,0] [1,2,0] [1,3,0] [1,4,0] [1,1,1] [1,3,1] [1,4,1]	[0,1,0] [0,2,0] [0,2,1] [1,2,1]

reward models, respectively. Within the tables, the sets of tutorial planner states mapped to each type of remediation (i.e., planner action) are shown in the columns. States are formatted as feature vectors within square brackets. The first feature of each vector is Pretest_ Score_Level. The second feature is Current_Chapter, and the third is Remediation_Count_High_Low.

Overall, both policies selected active remediation as the preferred intervention in the majority of states. In the case of the `FullNLG` reward model, the second most frequently selected action was constructive remediation. However, passive remediation was the second most frequently selected action for the `ChapterNLG` reward model. In total, active and constructive remediation were selected in 69% of states.

Active remediation was most strongly represented among states associated with high Pretest_Score_Level. For the `ChapterNLG` reward model, active remediation was selected in all but one high Pretest_Score_Level state. For the `FullNLG` reward model, active remediation was selected in all but two high Pretest_Score_Level states. Constructive remediation was selected only during Chapters 1 and 2 of the course. For the `FullNLG` reward model, constructive remediation was selected in three-quarters of states in which a learner had a low

Table 5. Induced policy for `Pretest + Chapter` state representation and `ChapterNLG` reward.

Action 1: no remediation	Action 2: passive	Action 3: active	Action 4: constructive
[0,4]	[0,1]	[0,3] [1,2] [1,3] [1,4]	[0,2] [1,1]

Table 6. Induced policy for `Pretest + Chapter` state representation and `FullNLG` reward.

Action 1: no remediation	Action 2: passive	Action 3: active	Action 4: constructive
[0,4] [1,3]	[0,1]	[0,3] [1,2] [1,4]	[0,2] [1,1]

Pretest_Score_Level and was in the first half of the course. Conversely, the policy recommended the ''no remediation'' action only during the latter half of the course. Furthermore, it was reserved for learners who had already received a high amount of remediation thus far.

Similar trends were observed for policies induced with the `Pretest + Chapter` state representation (Tables 5 and 6). Active remediation was again the most commonly selected intervention in these policies, particularly for learners with high prior knowledge about COIN. Constructive remediation was selected for a mixture of high prior knowledge learners, as well as low prior knowledge learners who were early in the course. The remaining remediation selections were distributed across passive and no interventions. Delivering no remediation was reserved for learners in the latter half of the course.

The `PretestOnly` state representation yielded policies that exclusively selected active remediation regardless of Pretest_Score_Level. The `AllFeatures` state representation yielded policies that did not clearly favor particular remediation strategies over others. In summary, neither of these policies showed strong resemblance to the ICAP model.

3.2.2. Research question 2: how do alternative representations for encoding MDP states and rewards impact induced RL-based tutorial policies that control ICAP-inspired remediation? To investigate how the induced policies prioritized different ICAP-inspired remediation approaches across different state representations and reward models, we examined the action frequency distributions across states for the induced policies. Figure 5(a) shows the frequency distributions

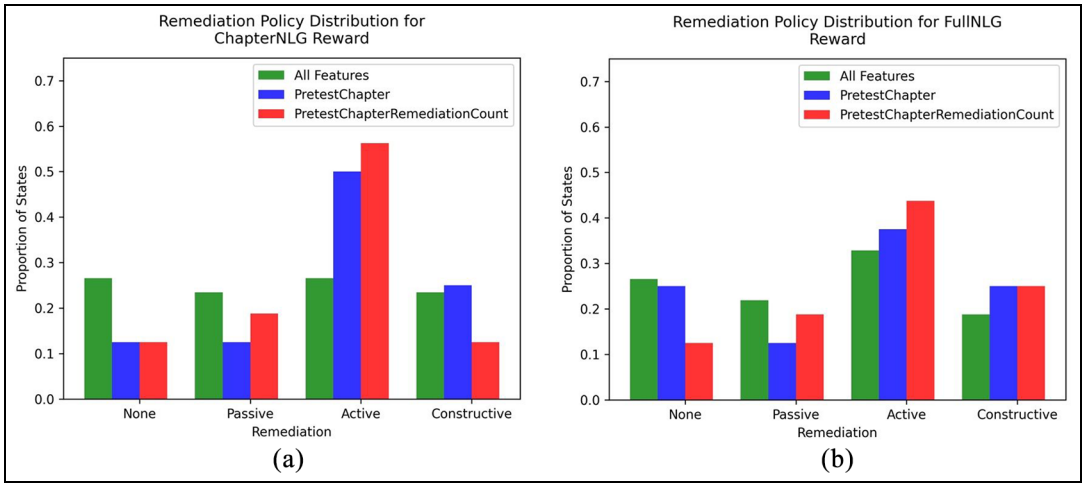

Figure 5. Bar charts showing overall frequencies of remediation strategies across alternate state representations and reward models: (a) policies induced with `ChapterNLG` reward; (b) policies induced with `FullNLG` reward. (Color online only.)

across different types of remediation associated with the `ChapterNLG` reward model. Figure 5(b) shows the frequency distributions across different types of remediation associated with the `FullNLG` reward model. In the figures, the four types of remediation are shown along the *x*-axis. The proportion of states in which a particular type of remediation was selected for a given policy is shown on the *y*-axis. The frequency distribution for the `AllFeatures` policy is shown in green. The `Pretest + Chapter + RemCount` policy is shown in red. The `Pretest + Chapter` policy is shown in blue. The `Pretest` remediation policy distribution is not shown. As mentioned previously, the `PretestOnly` policy always selected active remediation regardless of *Pretest_Score_Level*, which is a significant departure from the ICAP model. Therefore, we omit it from this analysis.

For both reward models, the `Pretest + Chapter + RemCount` policy and `Pretest + Chapter` policy shared similar frequency distributions across the four types of remediation. Both policies selected active remediation in 38–56% of states, depending on the reward model. In comparison, the policies selected constructive remediation in 12–25% of states, passive remediation in 12–19% of states, and no remediation in 12–25% of states. These frequency distributions are contrasted with the `All Features` policies, which yielded a nearly uniform distribution across the four different types of remediation. The `AllFeatures` policies were the configuration that *least* resembled the pairwise ordering predicted by ICAP. For the `ChapterNLG` reward model, the `AllFeatures` policy selected every type of remediation in 23–27% of states. For the `FullNLG` reward model, the `All Features` policy selected every type of remediation in 19–33% of states. This is notable because the `All Features` policies had the greatest access to state information relative to the policies that used competing state representations.

A possible explanation for the disconnect between the ICAP model and the `AllFeatures` state representation is data sparsity; the training dataset may have simply contained too little data to train an effective tabular policy using certainty equivalent RL techniques. Table 7 shows descriptive statistics summarizing the number of data points utilized to compute value-function estimates for each state–action pair within the induced policies. From the table, it is apparent that as the richness of the state representation increased, the number of data points available to estimate the values of state–action pairs decreased. For example, value estimates in the `AllFeatures` policies were computed based upon 19 data points, on average, from the training dataset. In contrast, value estimates in the `Pretest + Chapter` policies were computed based upon over 180 data points, on average. Furthermore, one can observe that some state–action pairs in the `AllFeatures` policy were never observed in the training data at all. This inverse relationship has the effect of increasing the uncertainty of value estimates for policies that leverage richer state representations. In other words, there was greater uncertainty about whether an induced policy was in fact optimal for that state representation and reward model whenever more information was made available to the tutorial planner through the state representation.

Table 7. Descriptive statistics of observation counts for each state–action pair in induced reinforcement learning-based tutorial policies.

Policy	Reward	Mean	St. dev	Max	Min	Mode
`AllFeatures`	`ChapterNLG`	19	28.5	132	0	2
`AllFeatures`	`FullNLG`	18.8	26.6	125	0	2
`Pretest+Chapter+RemCnt`	`ChapterNLG`	100.6	98.4	299	2	–
`Pretest+Chapter+RemCnt`	`FullNLG`	101.8	100.8	299	2	–
`Pretest+Chapter`	`ChapterNLG`	223.2	151.9	415	11	–
`Pretest+Chapter`	`FullNLG`	182.1	138.3	401	11	–
`Pretest`	`ChapterNLG`	1037	292	1329	745	–
`Pretest`	`FullNLG`	1037	292	1329	745	–

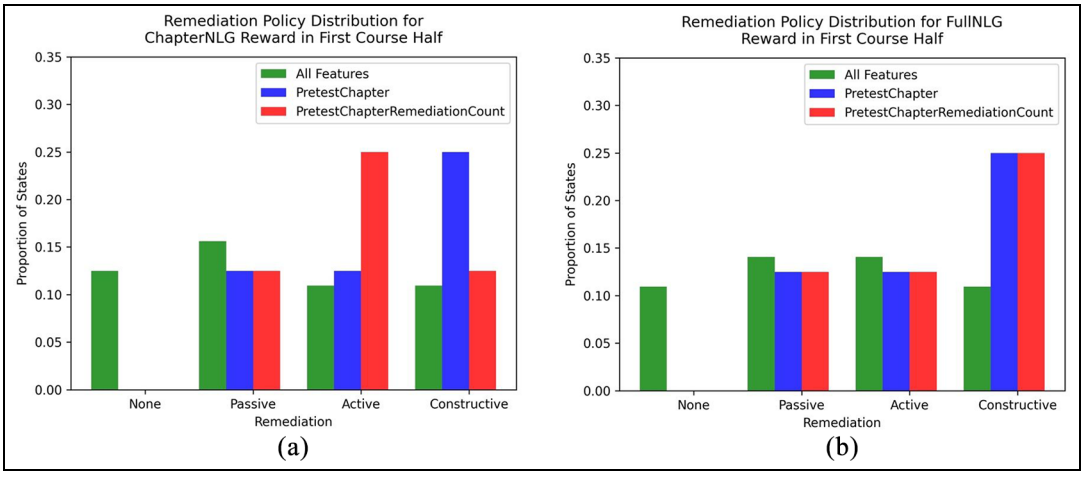

Figure 6. Bar charts showing overall frequencies of remediation strategies across alternate state representations and reward models during the initial half (Chapter = 1 or 2) of the online training course: (a) policies induced with `ChapterNLG` reward; (b) policies induced with `FullNLG` reward. (Color online only.)

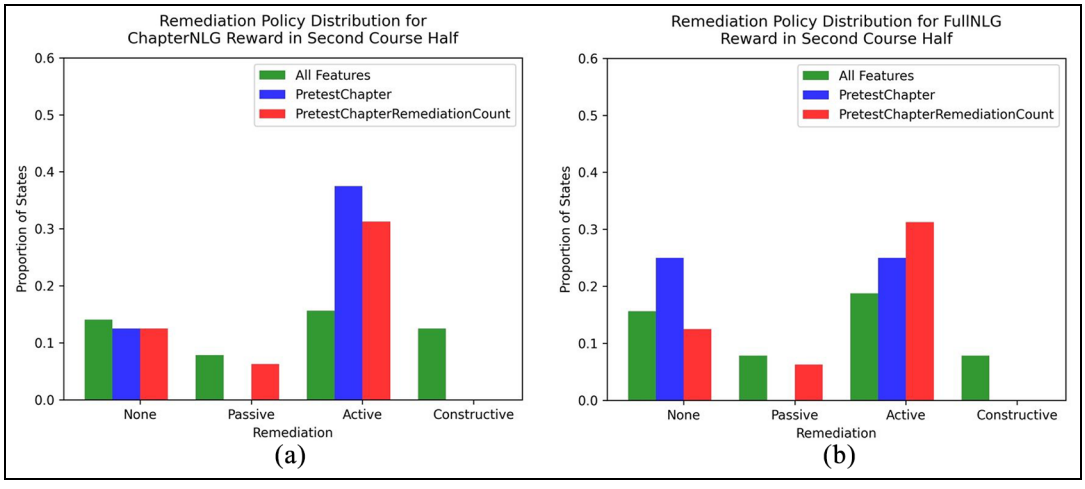

Figure 7. Bar charts showing overall frequencies of remediation strategies across alternate state representations and reward models during the latter half (Chapter = 3 or 4) of the online training course: (a) policies induced with `ChapterNLG` reward; (b) policies induced with `FullNLG` reward. (Color online only.)

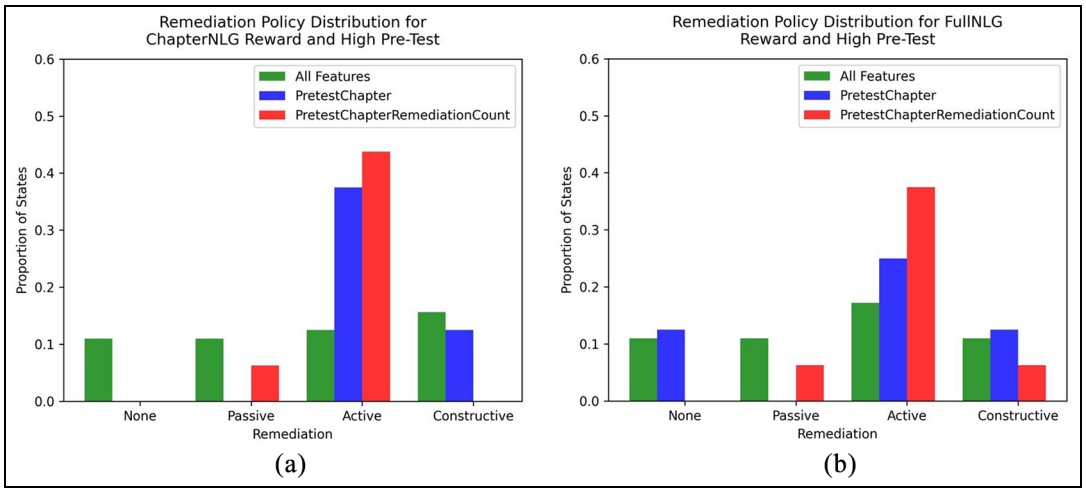

Figure 8. Bar charts showing overall frequencies of remediation strategies across alternate state representations and reward models for high Pretest_Score_Level learners: (a) policies induced with `ChapterNLG` reward; (b) policies induced with `FullNLG` reward.

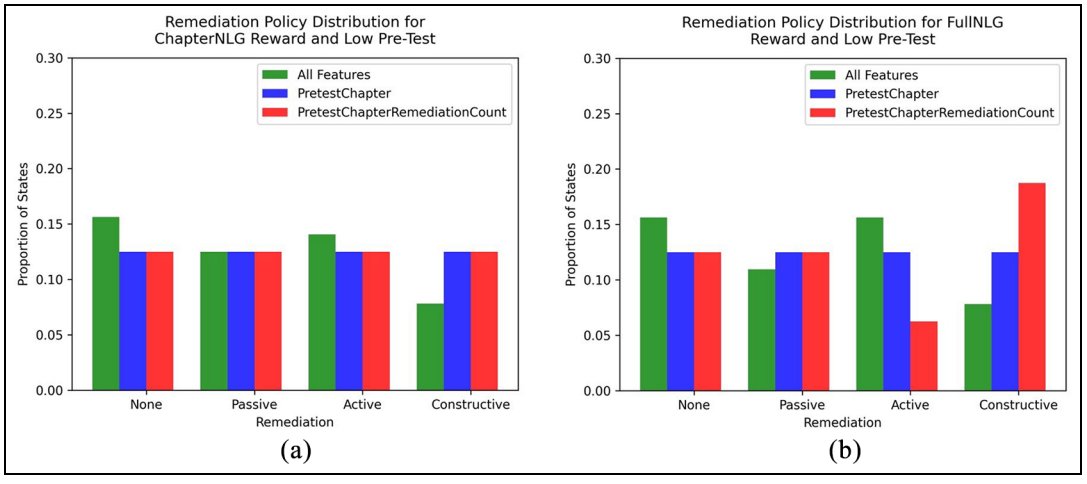

Figure 9. Bar charts showing overall frequencies of remediation strategies across alternate state representations and reward models for low Pretest_Score_Level learners: (a) policies induced with `ChapterNLG` reward; (b) policies induced with `FullNLG` reward.

3.2.3. Research question 3: under what circumstances do RL-based tutorial policies for ICAP-inspired remediation select constructive interventions? Active interventions? Passive interventions? No intervention?. To investigate how the induced tutorial policies selected ICAP-inspired remediation under different conditions, we examined the policies' action frequency distributions during different phases of the training course and for different learners. Figures 6 and 7 show the frequency distributions across different types of remediation during the initial half of the course and the latter half of the course, respectively. Axes and bar colors are the same as previous frequency distribution figures.

During the first half of the course, the `Pretest+Chapter` and `Pretest+Chapter+RemCount` state representations yielded policies that partially resembled the pairwise ordering predicted by the ICAP framework. Regardless of the reward model, the `Pretest+Chapter` policies selected constructive remediation in twice as many states as active and passive remediation, respectively, during this phase of the course (Figures 6(a) and (b)). They refrained from selecting "no remediation" in any states. The `Pretest+Chapter+RemCount` state representation combined with the `FullNLG` reward model (Figure 6(b)) produced the same

frequency distribution: constructive remediation was selected in twice as many states as active and passive remediation, and ''no remediation'' was never selected.

The latter half of the course yielded a significantly different remediation action frequency distribution (Figure 7). The `Pretest + Chapter` and `Pretest + Chapter + RemCount` policies did not select constructive remediation in any states during the second half of the course. Rather, active remediation and no remediation were selected most often for both sets of state representations and reward models. Across both halves of the course, the `AllFeatures` state representation yielded policies with action frequency distributions that were relatively close to uniformly distributed across different remediation types.

Figures 8 and 9 show remediation action frequency distributions of the induced policies for learners with high *Pretest_Score_Level* and low *Pretest_Score_Level*, respectively. Figure 8 shows that both `Pretest + Chapter` and `Pretest + Chapter + RemCount` policies selected active remediation much more frequently than they selected other forms of remediation for learners with high prior COIN content knowledge. In contrast, Figure 9 shows relatively little differentiation in the frequency of remediation selections across induced policies for learners with low prior COIN content knowledge.

3.3. Discussion

Overall, the results of Study 2 provide partial support for the hypothesis that RL-based tutorial planning will automatically reconstruct the pairwise ordering predicted by the ICAP model for remediation in adaptive online training environments. For the `Pretest + Chapter + RemCount` policies, active remediation was consistently prioritized over passive remediation, which was selected more often than no remediation. Constructive remediation was selected less often than anticipated, but it was prioritized for those learners who had lower prior knowledge and missed recall questions during the first half of the course. This pattern is consistent with the instructional technique of *fading* scaffolding over the span of a training course.[40] Furthermore, the results were consistent with expectations about the presence of a trade-off between, on the one hand, higher levels of cognitive engagement, and on the other hand, time and cognitive load associated with ICAP-inspired remediation. Rather than universally prioritize constructive remediation in the majority of states, the induced policies differentially selected remediation requiring varying levels of cognitive engagement in different states.

In our examination of how alternative representations for encoding MDP states and rewards impact induced RL-based tutorial policies that control ICAP-inspired remediation, results showed that for the different state representations, the policies selected active remediation most frequently except in the case of the AllFeatures state representation. Results also showed that differences between policies that shared the same state representation but optimized different reward models were relatively minor. For the `Pretest + Chapter + RemCount` state representation, the `FullNLG` reward model yielded policies that selected constructive remediation in more states than the `ChapterNLG` reward model. However, there were no differences in constructive remediation selections between policies for the `Pretest + Chapter` state representation. In contrast, the `FullNLG` reward model yielded a policy for the `Pretest + Chapter` state representation that more frequently selected no intervention than did the `ChapterNLG` reward model. In general, policies induced with the same state representation but different reward models were more than 87% similar for the `Pretest + Chapter` and `Pretest + Chapter + RemCount` state representations.

Our third research question asked under which circumstances do the tutorial policies select the different kinds of ICAP-inspired remediation. Results showed the selected policy depended on where students were in the course, prior knowledge, and the number of remediation instances they had received.

4. General discussion

This paper presents results from a pair of studies that aimed to devise and investigate data-driven tutorial planning policies using RL techniques to provide learners with adaptive remediation. The dataset included training interaction data from over 500 learners who completed a 90-minute online training course and received different forms of remediation that were based on the ICAP framework for cognitive engagement. Preliminary analysis of the training data showed learners demonstrated modest learning gains by completing the course and that constructive and active forms of remediation, which are more cognitively engaging, helped learners correct errors that led to incorrect responses on recall quiz items. Preliminary analyses from the training dataset also suggest that the effectiveness of ICAP-inspired remediation may change over time and under different conditions, pointing toward the need for adaptive tutorial policies to control how and when different forms of remediation are delivered to learners.

The RL policy analyses, or Study 2, demonstrated that certainty equivalent RL can yield instructional policies that operationalize key aspects of the ICAP model of cognitive engagement in learning. Results showed that induced policies reconstructed several components of the pairwise ordering predicted by the ICAP framework: a policy that

utilized the `Pretest + Chapter + RemCount` state representation and in which the `FullNLG` reward model selected active remediation more frequently than passive or no remediation. The policy also selected constructive remediation more frequently than passive or no remediation. In addition, it selected passive remediation more frequently than no remediation. The pairwise ordering was especially prominent during the first half of the training course. The induced policy prioritized active remediation for learners with higher prior content knowledge – this shows promise for yielding more efficient training times – and it prioritized constructive remediation for learners with lower prior knowledge during the first half of the course. Finally, the policy only selected no remediation during the latter half of the course for learners who had already received an above-threshold amount of remediation. This component of the policy was consistent with the instructional approach of fading, which is a surprising but interesting result. Overall, the findings provide evidence that lends support to the ICAP model, and they show that RL-based tutorial planning yields adaptive remediation policies that show significant promise in online training environments.

Notably, Study 1 demonstrated that constructive remediation was most helpful for correctly addressing a missed recall question, whereas Study 2 indicated that active remediation was most helpful for promoting learning gains. This difference in prioritized remediation strategies between the studies could be explained by a number of factors, including differences in the dependent variables and reward states used in each study. Study 1 used an operationalization of remediation effectiveness that examined the probability that a remediation activity led to a successful response on a previously missed quiz question. Study 2 utilized NLGs from pretest to posttest, at the chapter and course levels, as reward states while also accounting for differences in pretest scores, course progress, and remediation interaction history. The differences in the dependent variables and states may have driven the differences between the behavioral results and the RL results. Reward engineering is a critical issue in the design of RL systems, and this includes RL-based tutorial planners. NLGs have been previously used as a reward in adaptive learning environments, but there are other rewards that merit investigation. This is a promising direction for future work.

Results of the RL policy investigation also showed that state representation has an important effect on the content and action frequency distribution of induced tutorial policies. State representations that contain too few features (e.g., `PretestOnly`) or too many features (e.g., `AllFeatures`) produced policies that were ineffective at differentiating how to select remediation in different circumstances. For policies with too many features, data sparsity issues may arise that increase the uncertainty of computed value-function estimates, which are utilized to identify the optimal tutorial policy. The results of Study 2 did not find evidence showing that alternative reward models – particularly rewards computed from granular versus aggregate measures of learning – had a major impact on RL-based tutorial policies.

Results also suggested that RL-based tutorial planning may produce different policies for learners with higher prior content knowledge than learners with lower prior content knowledge. For learners with higher prior knowledge, induced policies selected active remediation most frequently, whereas for learners with lower prior knowledge, induced policies selected remediation strategies with near-uniform probability. These findings are reminiscent of related work on RL-based tutorial planning that has found evidence of an aptitude–treatment interaction effect in studies involving student classroom interactions with an intelligent tutoring system.[9] The results point toward RL-based tutorial policies that implement different remediation strategies, depending on learners' prior knowledge, and furthermore, have different levels of effectiveness in promoting learning outcomes. Given that a key objective of adaptive learning technologies is to aid learners who might struggle without individualized support, investigating these trends further is an important direction of future work.

4.1. Practical implications

What do the results of our study mean in the context of using theoretically grounded frameworks, such as ICAP, to guide pedagogical decisions in AISs? The results of the current study lend support for the predictions made by the ICAP framework and provide evidence that an adaptive approach to operationalizing ICAP-inspired remediation in AISs, as opposed to strictly following the prescribed ordering of ICAP, is appropriate, at least in the specific context examined in this work. AISs that have the flexibility to tailor the type of ICAP-inspired activity provided to learners offer a promising approach for maintaining student engagement and maximizing learning outcomes. Rather than receiving one form of remediation over the course of training or receiving different types of remediation that are scaffolded according to a fixed schedule (e.g., constructive, active, passive), the current results show that using RL to learn tutorial policies from data offers an approach for delivering tailored remediation in AISs. Using a data-driven tutorial planning approach to learn which remediation activities work better for different

learners under different circumstances, and further refining these policies over time as more data becomes available, offers a pathway to deeply adaptive training experiences.

4.2. Directions for future research

There are several limitations of this work that merit acknowledgment. Firstly, the study would have benefited from a larger sample size. RL analyses are data-intensive and require large training datasets to avoid data sparsity issues. Our sample included responses from over 500 students. Training datasets of this size are not uncommon when applying RL to induce policies with human-subjects' datasets, particularly in education-based settings where training datasets can range from 200 to more than 1000 subjects.[42–44] Developing RL methods that account for limited samples sizes is an active area of research.[45] Secondly, the analyses presented in Study 1 could have been strengthened if a control condition were included in the study. This would have allowed us to examine if the learning gains observed were a result of being exposed to remediation or an artifact of completing the training. Because the purpose of this data collection was to generate a training dataset and not to conduct an experiment, a control group was not included in the study design. Thirdly, our analysis of the RL policies did not include statistical hypothesis testing, which raises questions about reliability. An approach to address this issue would be to run the RL analyses multiple times to examine whether the observed differences between remediation types, state representations, and reward models are observed consistently. We have elected not to do this in the current work because it would involve running value iteration multiple times on randomly selected subsets of the data, which would exacerbate data sparsity concerns, which we have highlighted as being an important issue. The real test of the reliability of the RL-based tutorial planning results would be to observe how frequently constructive, active, passive, and no remediation activities are delivered to learners in a run-time setting, and then evaluate whether there are significant differences in their frequency. This is a promising direction for future work.

There are several promising directions for future research on the design, development, and evaluation of RL-based tutorial policies in online training environments. Recent years have seen dramatic advances in RL techniques and applications, especially in the area of deep RL, which leverages deep neural networks to capture patterns in high-dimensional input data and approximate nonlinear value-functions, and optimize parameterized policy representations.[41] There is also growing interest in human-in-the-loop RL, which seeks to augment RL with human input to improve the efficiency and effectiveness of induced control policies.[46] These methods show significant promise for devising data-driven tutorial planners in AISs. They provide principled mechanisms for implementing richer state representations with large numbers of input features, as well as methods for utilizing human demonstrations and feedback to guide the learning process. Deep RL techniques, such as deep Q-networks and A3C, have been investigated for tutorial planning in game-based learning environments and intelligent tutoring systems, respectively.[9,22,25,26] Research on Deep RL frameworks for tutorial planning has focused primarily on adaptive learning technologies for K-12 and undergraduate students, as well as traditional academic subjects, such as microbiology and logic.[22,23,25] Deep RL has not been widely investigated for data-driven tutorial planning in military training domains, which is an important direction for investigating generalizability to alternative subjects and educational contexts.

A second promising direction is validating the framework with data from a military population. This project included participants from Amazon MTurk to investigate the impact of ICAP-inspired remediation and the creation of RL-based tutorial policies for an online training course on COIN. A major benefit of working with the MTurk population was the relative ease of collecting data to train RL-based tutorial planning models. However, study participants may have differed in their prior knowledge, experience, motivation, and demographics relative to participants from a military population. Investigating how well the adaptive training materials and tutorial planning methods utilized in this study translate to a military audience is an important future step for the work. Similarly, investigating how the RL-based tutorial planning approach and ICAP-inspired remediation policies transfer to other training domains and educational settings are key future directions.

An attractive attribute of the policies induced in this project is their relative simplicity, making their implementation within an adaptive training platform, such as GIFT, relatively straightforward. Although results suggest that RL-based tutorial planning can yield data-driven remediation policies that reproduce major components of the ICAP model, investigating their impact on learning outcomes through randomized experimentation is the ''gold standard'' for empirical evaluation. There are several study designs that merit consideration. Induced policies could be compared to a control condition that provides no remediation. Alternatively, induced policies could be compared to a random policy that emulates the remediation strategy utilized in the MTurk study. Induced policies could also be

compared to a heuristic model that utilizes a set of consistent rules for driving remediation (e.g., always deliver passive remediation, always deliver constructive remediation, etc.). By isolating the impact of induced remediation policies on learning outcomes, the efficacy of RL-based tutorial planning can be examined and our understanding of data-driven approaches for design and development of adaptive training environments can be extended.

Finally, future research should examine the impact of RL-based instructional policies within a run-time virtual training environment. Across the military, virtual and game-based training is being applied more than ever to assist in skill acquisition through rapid exposure to sets and reps that provide novel opportunities to practice. Extending the RL-based policies to support remedial coaching interventions that target human-performance dimensions is an important next step.

5. Conclusion

Data-driven approaches to tutorial planning, such as RL, show significant promise for devising effective models of instructional techniques and strategies for complex domains and learning environments. This paper summarized research on a RL-based approach for data-driven tutorial planning. We investigated tutorial planning in the domain of COIN training, with a focus on online training environments. We devised a set of adaptive remediation policies that are applicable to multiple learning environments and that are inspired by the ICAP framework.[10] Results from a study involving more than 500 participants recruited through Amazon MTurk found that participants achieved significant learning gains by completing the course. The effectiveness of different remediation interventions largely followed the ICAP model.[10,29] An examination of RL-induced policies found that they reproduced key components of the ICAP model. Induced policies selected active remediation more frequently than passive and no remediation. Similarly, constructive remediation was selected more often than passive and no remediation, and it was prioritized for learners with low prior knowledge during the first half of the course, when remediation was observed to be most effective. Passive remediation was selected more frequently than no remediation, and no remediation was reserved for the latter half of the course, which was consistent with a pedagogical strategy of faded scaffolding. These findings demonstrate that RL-based tutorial planning shows significant promise as a framework for devising generalizable, data-driven tutorial planning models that automatically improve instructional techniques, strategies, and tactics based upon data from learner interactions with online training environments.

Funding

This work was supported by the U.S. Army Research Laboratory (cooperative agreement W911NF-15-2-0030).

ORCID iDs

Randall Spain https://orcid.org/0000-0003-4067-7262
Andy Smith https://orcid.org/0000-0002-6577-7764

References

1. Chowdhary KR. *Fundamentals of artificial intelligence*. New Delhi: Springer, 2020, pp.603-649.
2. Noroozi O, Alikhani I, Järvelä S, et al. Multimodal data to design visual learning analytics for understanding regulation of learning. *Comput Hum Behav* 2019; 100: 298–304.
3. López CE, Cunningham J, Ashour O, et al. Deep reinforcement learning for procedural content generation of 3d virtual environments. *J Comput Inform Sci Eng* 2020; 20: 051005.
4. Doroudi S, Aleven V and Brunskill E. Where's the reward? *Int J Artif Intell Educ* 2019; 29: 568–620.
5. VanLehn K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ Psychol* 2011; 46: 197–221.
6. Durlach PJ and Ray JM. *Designing adaptive instructional environments: insights from empirical evidence*. Report for the US Army Research Inst For the Behavioral and Social Sciences, Orlando. Report no. 1297, 2011.
7. Murray T. An overview of intelligent tutoring system authoring tools: updated analysis of the state of the art. In: Murray T, Blessings S and Ainswirth S. (eds) *Authoring tools for advanced technology learning environments*. Dordrecht: Springer, 2003, pp.491–544.
8. Rowe JP and Lester JC. Improving student problem solving in narrative-centered learning environments: a modular reinforcement learning framework. In: Conati C, Heffernan N, Mitrovic A, et al. (eds) *international conference on artificial intelligence in education*, Madrid, Spain, 2015, pp.419–428. Cham: Springer.
9. Shen S, Mostafavi B, Barnes T, et al. Exploring induced pedagogical strategies through a Markov decision process framework: lessons learned. *J Educ Data Mining* 2018; 10: 27-68.
10. Chi MT. Active-constructive-interactive: a conceptual framework for differentiating learning activities. *Topic Cognit Sci* 2009; 1: 73–105.
11. Sottilare R, Barr A, Robson R, et al. Exploring the opportunities and benefits of standards for adaptive instructional systems (AISs). In: Sottilare R (eds) *proceedings of the adaptive instructional systems workshop in the industry track of the 14th international intelligent tutoring systems*, Montreal, Canada, 2018, pp.49–53.
12. Durlach P and Spain R. *Framework for instructional technology: methods of implementing adaptive training and education*. Report for the US Army Research Institute for the Behavioral and Social Sciences. Report No. 1335, Fort Belvoir, 2014.
13. Swart EK, Nielen TM and Sikkema-de Jong MT. Supporting learning from text: a meta-analysis on the timing and content of effective feedback. *Educ Res Rev* 2019; 28: 100296.

14. Walker E, Rummel N and Koedinger KR. Adaptive intelligent support to improve peer tutoring in algebra. *Int J Artif Intell Educ* 2014; 24: 33-61.
15. Chen O, Kalyuga S and Sweller J. The worked example effect, the generation effect, and element interactivity. *J Educ Psychol* 2015; 107: 689-704.
16. VanLehn K. The behavior of tutoring systems. *Int J Artif Intell Educ* 2006; 16: 227-265.
17. Aleven V, Mclaren BM, Sewall J, et al. A new paradigm for intelligent tutoring systems: example-tracing tutors. *Int J Artif Intell Educ* 2009; 19: 105–154.
18. Sottilare RA. Challenges to enhancing authoring tools and methods for intelligent tutoring systems. In: Sottilare R, Graesser A, Hu X, et al. (eds) *Design recommendations for intelligent tutoring systems: volume 3 - authoring tools and expert modeling techniques*. Orlando, FL: U.S. Army Research Laboratory, 2015, pp.3-7.
19. Williams JJ, Kim J, Rafferty A, et al. Axis: generating explanations at scale with learner sourcing and machine learning. In: *proceedings of the third (2016) ACM conference on learning@ scale*, Edinburgh, Scotland, 25–26 April 2016, pp.379–388.
20. Zhou G, Wang J, Lynch CF, et al. Towards closing the loop: bridging machine-induced pedagogical policies to learning theories. In: Hu X, Barnes T, Hershkovitz A, et al. (eds) *proceedings of the tenth international conference on educational data mining*, Wuhan, China, 2017, pp.112–119.
21. Sutton RS and Barto AG. *Reinforcement learning: an introduction*. Cambridge, Massachusetts: MIT Press, 2018.
22. Ausin MS, Azizsoltani H, Barnes T, et al Leveraging deep reinforcement learning for pedagogical policy induction in an intelligent tutoring system. In: Lynch C, Merceron A, Desmarais M, et al. (eds) *proceedings of the 12th international conference on educational data mining*, Montreal, Canada, 2–5 July 2019, pp.168–177.
23. Chi M, VanLehn K, Litman D, et al. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Model User Adapt Interact* 2011; 21: 137-180.
24. Rafferty A, Ying H and Williams J. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *J Educ Data Mining* 2019; 11: 47-79.
25. Wang P, Rowe JP, Min W, et al. Interactive narrative personalization with deep reinforcement learning. In: Sierra C (ed) *proceedings of the twenty-sixth international joint conference on artificial intelligence*, Melbourne, Australia, 19–25 August 2017, pp.3852–3858.
26. Wang P, Rowe JP, Min W, et al. High-fidelity simulated players for interactive narrative planning. In: Lang J (ed) *proceedings of the twenty-seventh international joint conference on artificial intelligence*, Stockholm, Sweden, 13–19 July 2018, pp.3884–3890.
27. Brunskill E and Russell S. Partially observable sequential decision making for problem selection in an intelligent tutoring system. In: Pechenizkiy M, Calders T, Conati C, et al. (eds) *proceedings of the 4th international conference on educational data mining*, 6-8 July 2011, pp.327–328. Eindhoven.
28. Folsom-Kovarik JT, Sukthankar G and Schatz S. Tractable POMDP representations for intelligent tutoring systems. *ACM Trans Intell Syst Technol* 2013; 4: 1–22.
29. Chi MT and Wylie R. The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ Psychol* 2014; 49: 219–243.
30. Sottilare RA, Brawner KW, Goldberg BS, et al. *The generalized intelligent framework for tutoring (GIFT)*. Orlando: US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED), 2012.
31. Sottilare RA, Brawner KW, Sinatra AM, et al. An updated concept for a Generalized Intelligent Framework for Tutoring (GIFT), GIFTtutoring.org (2017, accessed 29 June 2021).
32. McAlinden R, Gordon A, Lane HC, et al. UrbanSim: a game-based simulation for counterinsurgency and stability-focused operations. In: Craig S and Dicheva D (eds) *workshop on intelligent educational games at the 14th international conference on artificial intelligence in education*, Brighton, UK, 6–7 July 2009, pp.41–50.
33. Rowe J, Spain R, Pokorny B, et al. Design and development of an adaptive hypermedia course for counterinsurgency training in GIFT: Opportunities and lessons learned. In: Sottilare R (ed) *proceedings of the sixth annual gift user symposium (giftsym6)*, Orlando, Florida, 30 May–2 June 2018, pp.229–239. Orlando, FL: U.S. Army Research Laboratory.
34. Elliot AJ and Murayama K. On the measurement of achievement goals: critique, illustration, and application. *J Educ Psychol* 2008; 100: 613-628.
35. Spain R, Rowe J, Goldberg B, et al. Enhancing learning outcomes through adaptive remediation with GIFT. In: *proceedings of the 2019 interservice/industry training simulation and education conference (I/ITSEC)*, Orlando, Florida, 2–6 December 2019, paper no. 19275, pp.1–11.
36. Spain R, Rowe J, Goldberg B, et al. Towards data-driven tutorial planning for counterinsurgency training in GIFT: preliminary findings and lessons learned. In: Sottilare R (ed) *proceedings of the 7th annual gift users symposium*, Orlando, Florida, 17 May–19 May, 2019, pp.111–120. Orlando, FL: U.S. Army Research Laboratory.
37. Taub M, Sawyer R, Smith A, et al. The agency effect: the impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Comput Educ* 2020; 147: 103781.
38. Barnes T and Stamper J. Toward automatic hint generation for logic proof tutoring using historical student data. In: Woolf B, Aïmeur E, Nkambou R, et al. (eds) *international conference on intelligent tutoring systems*, Berlin, Heidelberg, 23–27 June 2008, pp.373–382. Montreal, Canada: Springer.
39. Rowe JP. *Narrative-centered tutorial planning with concurrent Markov decision processes*. PhD Thesis, North Carolina State University, Raleigh, NC, 2013.
40. Pea RD. The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *J Learn Sci* 2004; 13: 423–451.

41. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature* 2015; 518: 529–533.
42. Zhou G, Yang X, Azizsoltani H, et al. Improving student-system interaction through data-driven explanations of hierarchical reinforcement learning induced pedagogical policies. In: *proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, Genoa, Italy, 14–17 July 2020, pp.284–292.
43. Ausin MS, Maniktala M, Barnes T, et al. Exploring the impact of simple explanations and agency on batch deep reinforcement learning induced pedagogical policies. In: Bittencourt I, Cukurova M, Muldner K, et al. (eds) *international conference on artificial intelligence in education*, Ifrane, Morocco, 6 July 2020, pp.472–485. Cham: Springer.
44. Zhou G, Azizsoltani H, Ausin MS, et al. Hierarchical reinforcement learning for pedagogical policy induction. In: *international conference on artificial intelligence in education*, Chicago, IL, 25 June 2019, pp.544–556. Cham: Springer.
45. Yu Y. Towards sample efficient reinforcement learning. In: Lang J (ed) *proceedings of the twenty seventh international joint conference on artificial intelligence*, Stockholm, Sweden, 13 July 2018, pp.5739–5743.
46. Taylor ME. Improving reinforcement learning with human input. In: *proceedings of the twenty seventh international joint conference on artificial intelligence*, Stockholm, Sweden, 13 July 2018, pp.5724–5728.

Author biographies

Randall Spain is a research scientist in the Center for Educational Informatics at North Carolina State University where he uses principles, theories, and methods of applied psychology to design and evaluate the impact of advanced training technologies on learning and performance. He holds a PhD in Human Factors Psychology and serves on the editorial board for *Military Psychology*.

Jonathan Rowe is a research scientist in the Center for Educational Informatics at North Carolina State University, as well as an Adjunct assistant professor in the Department of Computer Science. His research focuses on AI in advanced learning technologies, with an emphasis on game-based learning environments, intelligent tutoring systems, multimodal learning analytics, learner modeling, and computational models of interactive narrative generation. He also serves as an Associate Editor for the *International Journal of Artificial Intelligence in Education* and *IEEE Transactions on Learning Technologies*.

Andy Smith is a research scientist in the Center for Educational Informatics at North Carolina State University. His research is focused on the intersection of AI and education, with emphasis on user modeling, game-based learning, and educational data mining.

Benjamin Goldberg is a senior scientist at the U.S. Army DEVCOM - Soldier Center, Simulation and Training Technology Center (STTC) in Orlando, FL. His research is focused on the application of intelligent tutoring and AI techniques to build adaptive training programs that improve performance and accelerate competency development. He is co-creator of GIFT, and holds a PhD in Modeling & Simulation from the University of Central Florida.

Robert Pokorny is a senior scientist at Affinity Associates. He earned his PhD in Experimental Psychology at the University of Oregon in 1985, and completed a postdoctoral appointment at the University of Texas at Austin in AI. He has participated in many cognitive science research projects, largely focused on designing simulation-based training, interface design of complex systems, and performance assessment.

Bradford Mott is a senior research scientist in the Center for Educational Informatics at North Carolina State University. His research interests include computer games, computational models of interactive narrative, and intelligent game-based learning environments.

James Lester is Distinguished University professor of Computer Science at North Carolina State University, where he is director of the Center for Educational Informatics. He is a fellow of the Association for the Advancement of Artificial Intelligence (AAAI).