Behavior Research Methods (2022) 54:216–232
https://doi.org/10.3758/s13428-021-01602-9

Building an intelligent recommendation system for personalized test scheduling in computerized assessments: A reinforcement learning approach

Jinnie Shin[1] · **Okan Bulut**[2]

Accepted: 20 April 2021 / Published online: 15 June 2021

Abstract

The introduction of computerized formative assessments in the classroom has opened a new area of effective progress monitoring with more accessible test administrations. With computerized formative assessments, all students could be tested at the same time and with the same number of test administrations within a school year. Alternatively, the decision for the number and frequency of such tests could be made by teachers based on their observations and personal judgments about students. However, this often results in rigid test scheduling that fails to take into account the pace at which students acquire knowledge. To administer computerized formative assessments efficiently, teachers should be provided with systematic guidance regarding effective test scheduling based on each student's level of progress. In this study, we introduce an intelligent recommendation system that can gauge the optimal number and timing of testing for each student. We discuss how to build an intelligent recommendation system using a reinforcement learning approach. Then, we present a case study with a large sample of students' test results in a computerized formative assessment. We show that the intelligent recommendation system can significantly reduce the number of testing for the students by eliminating unnecessary test administrations where students do not show significant progress (i.e., growth). Also, the proposed recommendation system is capable of identifying the optimal test time for students to demonstrate adequate progress from one test administration to another. Implications for future research on personalized assessment scheduling are discussed.

Keywords Test administration optimization · Reinforcement learning · Computerized formative assessment · Personalized learning

Introduction

Computerized formative assessments establish a connection between computer technologies and formative assessments to effectively manage and deliver classroom assessments (Webb et al., 2013). Formative assessments refer to the general process that engages students and teachers to evaluate student learning, provide feedback, and improve learning outcomes using formative evaluation (McManus, 2008). In a critical review of research on formative assessments, Dunn and Mulvenon (2009) collectively defined formative evaluation as the evidence-based evaluation aiming to "inform teachers, students, and educational stakeholders about the teaching and learning process" (p.4, Dunn & Mulvenon, 2009). A diverse format of tools can be applied to formative assessment to effectively collect evidence of student learning (Bennett, 2011). Such tools include, but are not limited to, peer feedback and assessment (e.g., Volante & Beckett 2011), questioning (e.g., Black & Harrison, 2001), self-assessment (e.g., Andrade, 2019), and pedagogical documentation, such as portfolios (e.g., Buldu, 2010).

Appropriate use of formative assessment tools in the classroom has previously been identified to result in a profound student learning (McManus, 2008; Kingston & Nash, 2011). Meta-analyses on the effect of classroom formative assessment have revealed that formative assessments show predominant effects to improve learning outcomes

✉ Jinnie Shin
jinnie.shin@coe.ufl.edu

Okan Bulut
bulut@ualberta.ca

[1] Institute for Advanced learning Technologies, College of Education, University of Florida, Gainesville, FL, USA

[2] Centre for Research in Applied Measurement and Evaluation, University of Alberta, Alberta, Canada

with the effect size of 0.4 to 0.7 (Black & Wiliam, 1998; 2010), while the statistical evidence of the exact effect size has been criticized and remains controversial. There is a consensus on the importance and positive impact of formative assessments in education (Kingston & Nash, 2011; McMillan et al., 2013).

Computerized formative assessments were designed and introduced to continue the benefits of traditional formative assessment with the ease of assessment practices using technological aids. Computerized formative assessments or computer-assisted formative assessments refer to various types of formative assessments that involve computer technology in delivering, administering, and grading the assessment (Wilson et al., 2011). Traditional formative assessments often involve paper and pencil-based assessment tools (Joyce, 2018). On the other hand, computerized formative assessments allow online assessments. The drastic shift in the mode of assessment practices significantly helps reduce educators' burden in organizing and managing test administration processes. For example, using computerized formative assessments, teachers can effectively manage the continuous evaluation of students' academic growth without much effort in generating, evaluating, and interpreting assessment results (Tomasik et al., 2018). Moreover, with a meticulously designed scoring system yielding comparable scores, teachers could closely monitor the learning trajectories and performance improvement of individual students. Hence, they could make more accurate and effective decisions regarding students' learning progress. In turn, students could receive more individualized and customized feedback about their performance, so that they can learn about their strengths and weaknesses to develop a better insight into their current performance and future focus (Dopper & Sjoer, 2004).

With such substantial benefits, computerized formative assessments have been introduced to many classrooms as a replacement for traditional (i.e., paper-and-pencil) formative assessments to provide better learning and assessment experiences to students (Angus & Watson, 2009). Computerized formative assessments not only reduce teacher workload in administering assessments but also create a great opportunity for personalized learning through systematic performance monitoring and timely feedback. In fact, many studies focused on generating guidelines for developing effective computerized formative assessment frameworks to support personalized learning. For example, generating high-quality items automatically was introduced as an important precondition to bridge the gap between the traditional item writing process in paper-and-pencil testing to accelerate the transition to the digital assessment (Gierl & Lai, 2018). In addition, providing feedback in a timely and cost-effective manner has been pursued by many researchers (Bulut et al., 2019; Gierl et al., 2018).

Such frameworks aimed to satisfy the two principles underlying formative assessments: to deliver frequent and continuous assessments embedded within classroom instruction and to inform both teachers and students by effectively capturing students' learning progress (Dopper & Sjoer, 2004; Wongwatkit et al., 2017). Hence, the frameworks focused on meeting the assessment requirements prior to the test administration (e.g., item generation) and utilizing the acquired information regarding student performance after the assessments are conducted (e.g., feedback generation). However, despite continuous efforts, attempts to improve the assessment practices during assessment periods have remained relatively limited in the previous literature. Moreover, little to no studies have explored new ways to enhance the effectiveness of computerized formative assessments, to our knowledge.

In practice, scheduling of computerized formative assessments heavily relies on teachers' personal judgment and observation of students. In fact, teachers often act as a sole authority to decide which students should be tested, at which point of the school year, and how frequently such students should be tested (Redecker & Johannessen, 2013). This often leads to rigid test scheduling that fails to consider the pace at which students acquire knowledge. For example, teachers are expected to record and track the performance trajectories for all students who participated in each assessment so they can provide meaningful decisions for the necessity of a future administration and the optimal timing of the next assessment. This is quite a daunting task for teachers as they are expected to closely monitor 20 to 30 students' academic progress in an environment where students could participate in up to 10–20 formative assessments in a school year. Furthermore, the price of misinterpreting students' learning progress and administering an excessive number of assessments could result in losing the diagnostic value of formative assessments. Other negative effects of prolonged and overly frequent testing include decreasing students' learning motivation, failing to early detect at-risk students effectively, and preventing teachers from allocating enough time for teaching and instruction (Sharkey & Murnane, 2006).

To prevent such problems and maximize the benefits of computerized formative assessments, it is critical to provide teachers with systematic guidance to make effective test administration decisions based on each student's level of progress. Therefore, the purpose of this study is to build an intelligent recommendation system using an interactive optimization model to determine the optimal number and timing of test administrations for each student. In the current study, the optimality of test administration is defined as the practice of minimizing the number of assessments provided to students while maximizing the observed score improvement between the test administrations.

With a goal-oriented learning approach, the proposed system aims to accomplish the goal of selecting a meaningful test administration schedule for students, while reducing the overly excessive number of test administrations. Using a case study with a large sample of students' test results in a computerized reading assessment, we explored the capacity of using a reinforcement learning-based system to identify meaningful test performance history from students' learning trajectories. The algorithm capitalizes on students' performance history in computerized formative assessments and their text participation during the school year to make meaningful future test scheduling decisions for individual students.

Test optimization in computerized formative assessments

A computerized formative assessment is an instrument to measure students' real-time mastery of skills and knowledge. The results of a computerized formative assessment are expected to inform teachers regarding students' learning needs. Computerized formative assessment is praised for several aspects by teachers, educators, and researchers (Sharkey & Murnane, 2006). First, it is able to inform teachers on students' level of knowledge in a timely and continuous manner. Second, it can provide schools with information on developing additional instructional plans for students lagging in core subject areas such as reading, math, and science. Third, it can help the superintendent to identify schools that need extra help and to report general progress to school boards. Finally, it helps students understand their strengths and weaknesses to modify and self-regulate their learning strategies.

Despite such benefits, there are still controversies remaining on the use of computerized formative assessment effectively and efficiently in practice. Sharkey and Murnane (2006) pointed out several challenges in designing a successful computerized formative assessment. One of the dilemmas that stood out was about employing cost-effective formative assessment systems in schools, in particular, when using commercial computerized formative assessments. Commercial computerized formative assessments have numerous advantages of fast implementation, high item quality, and better alignment with state standards for assessments. However, employing a commercial formative assessment from testing agencies could cost from $5 to $75 per student per year, which might be a burden to school boards, especially if the number of students and the number of test administrations are high.

In addition to the cost-effectiveness of computerized formative assessments, frequency and timing of assessment have been considered important factors to maximize the diagnostic benefits of computerized formative assessments. In fact, a few studies have indicated that too frequent and prolonged testing could negatively affect students' test-taking behaviors and engagement (Sharkey & Murnane, 2006; van den Berg et al., 2018). For example, students can easily get bored and feel reluctant to complete a consecutive series of formative assessments, which may lead to behaviors like gaming the system or simply refusing to complete the assessments. In that case, the goal of using formative assessments to identify students' academic growth will not be achieved. Several studies argued that assessing students too frequently fails to prioritize learning and alienates students from learning to improve their performance (Sherrington, 2018). According to van den Berg et al. (2018), testing students more frequently using formative assessments did not significantly increase student performance in mathematics. Most importantly, prioritizing test administration and test experiences over instruction could significantly reduce the necessary classroom learning, resulting in "missing instruction time" and "wasted resources" (p.3, Bulut et al., 2020).

On the other hand, in the context of progress monitoring, more frequent testing has been praised for increasing the precision of students' academic growth estimation by reducing the measurement error (Mellard et al., 2009; Christ et al., 2012). For example, January et al. (2019) investigated the impact of testing frequency and the density of progress monitoring schedules on the accuracy of performance growth estimation in reading for second and fourth graders. The findings indicated that assessing students more frequently (e.g., twice a week rather than once a week) could significantly improve the confidence of accurately measuring students' academic growth. In other words, providing assessments more frequently could help diagnose students who struggle with reading.

The lack of consensus on the number of test administrations and the testing frequency for computerized formative assessments warrants the need for finding a systematic way to optimize test administration decisions. More specifically, minimizing the number of test administrations, attempting to avoid "over-testing" to keep students engaged and motivated, and maintaining a sufficient amount of information for accurate student monitoring are highly essential for the computerized assessment systems.

Test optimization using big data in education

Despite the potential impact of poorly optimized test administration decisions on students, research for optimizing the test administration has been a relatively unexplored area. More recently, a few studies have explored the possibility of

utilizing big data in education to provide individualized recommendations for test administration in computerized formative assessments (Bulut et al., 2020). The studies focused on providing systematic help to the teachers, to increase the capacity to individually monitor, evaluate, and make decisions based on students' learning trajectory (January et al., 2019; Dede, 2016; Fischer et al., 2020).

For instance, an intelligent recommender system has been used to provide customized and optimize solutions to for computerized formative assessments in grade 2 and grade 4 mathematics (Bulut et al., 2020). The study focused on identifying whether the use of an intelligence recommender system could minimize the number of test administrations for students. Also, the authors focused on investigating whether robust recommendations could be identified for students with abnormal growth (i.e., nonlinear increase/decrease, no score increase). The results indicated that the system could reduce the number of test administrations for both grade levels. Moreover, the system could maximize the score difference between the test administrations to increase the diagnostic information captured in each test administration for students.

Similarly, optimizing test length or test duration based on students' ability levels has been rigorously studied to advance computerized adaptive testing (CAT). The main objective of CAT differs from computerized formative assessments as the test optimization in CAT focuses on finding the most optimal (i.e., informative) items for each student, not the number of test administrations. However, the underlying optimization mechanism remains the same with the goal of minimizing the test burden. Further, CAT could solve problems in the traditional form of fixed-form assessment. For instance, CAT identifies and selects the most suitable item for each student based on their test history (Weiss & Kingsbury, 1984). In other words, CAT identifies the evidence to make future choices (e.g., the number of test items, upcoming test items) using the previous history (e.g., previous test items, correctly answered test items).

Recent CAT literature offers new solutions to map the unseen test administration of the students in the future by identifying the most optimal items for each student. For example, Nurakhmetov (2019) introduced how item selection in CAT could be investigated as a sequential decision process using reinforcement learning. To optimize the item selection process, the author focused on formalizing the task as a sequential classification problem using a partially observed Markov decision process (POMDP). Students' response patterns in the previously administered test items can be considered as partially observed sequence data of their test score based on the test items selected for them. Then, the future test item selection based on this partially observed test data can be framed as a sequential decision-making process. Using these properties, the system could identify the optimal selection of upcoming items based on students' partially observed item response patterns. This classification was conducted by the actor–critic algorithm. The proposed system could achieve 89 to 92% accuracy in making correct item selection decisions with 10 to 15 items.

The current study provides an interesting and novel framework to understand and optimize the sequence of test administration. First, the study framework introduces how the students' performance from the start to the end of the administration cycle could be understood as partially observed sequential data. We present more detailed information about the partially observed Markov decision process with regards to this framework in the next sections. Second, the study introduces how the various previous challenges in the optimization of test administrations can be mitigated using a flexible modeling capacity of the reinforcement learning framework. In the next section, we introduce reinforcement learning algorithms, which provide structural and theoretical guidelines integral to the current study.

Reinforcement learning algorithms

Reinforcement learning refers to interactive learning approaches where the learners attempt to discover the ideal actions while interacting with the environment (Sutton & Barto, 2018). Unlike the supervised learning algorithms, in which the system is guided to obtain desired examples using the labeled data, reinforcement learning does not confine the agent to identify a set of rules, but to explore all the possible actions while interacting and receiving feedback from the given scenario. Hence, the objective of reinforcement learning is to generate a robust system that could yield reasonable and appropriate decisions when given a novel scenario. Introducing the concept of reinforcement learning is outside the scope of this study. Instead, a list of comprehensive literature is provided for the readers to understand the important theoretical foundation of reinforcement learning (Sutton & Barto, 2018; Szepesvári, 2010; Mannor & Shimkin, 2004) as well as the various application of reinforcement learning in education (Iglesias et al., 2009; Dorça et al., 2013; Thomaz et al., 2006; Chi et al., 2011). Considering the journey of the reinforcement learning system to interact and try out various options, the reinforcement learning problem could be formalized as a Markov decision process with the decision-maker, or the environment that the agent interacts with (Fig. 1).

Within the reinforcement learning framework, the agent interacts with the environment while receiving the environments' state (S) based on its interaction or action

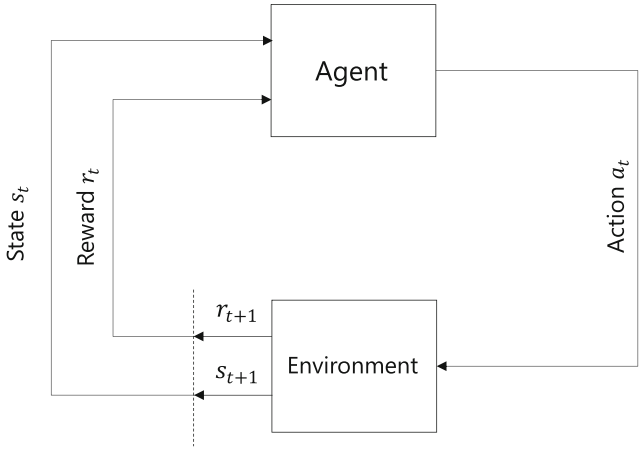

Fig. 1 A conceptual representation of reinforcement learning frameworks

(A) at each time step. In the meantime, the agent receives a certain reward (R) from such interactions while influencing or changing the current state of the environment (S′). Thereby, the ultimate objective of the agent in learning is to maximize the cumulative reward (R). Simply put, the agent is supposed to behave in a way that could choose the best action based on the given state yielding the highest reward. Therefore, almost all reinforcement learning problems are described formally as estimating how optimal for the agent to be in a given state (state-value function) or how optimal for the agent to perform a certain action in a given state (action-value function). Selecting and updating the optimal probability to choose the optimal action based on the state is critical to increasing the cumulative reward in the long run. For example, the policy indicates the probability of selecting the action a given the state s, $\pi(a|s)$. Throughout the process, the policy should be evaluated and improved reflecting the changes in the interaction. This is called the policy iteration because it attempts to evaluate and improve the policy π using a new value-function, v_π, until we reach the optimal policy π_*. As a result, we can gradually build a better policy that could yield higher rewards.

Optimizing test schedules in computerized formative assessment ultimately boils down to understanding how to effectively select important test points from a sequence of test window options. The selected test points are required to provide essential information to diagnose students' learning progress, by identifying and classifying students based on their growth and future performance. Hence, test administration points that appear redundant or less informative could be filtered, or not recommended. Such complex decision-making processes of selecting and recommending test points require a framework that could classify students based on their future success with students' test performance in the selected test points. However, it is important to note that the classifier will not be able to capitalize on the full history of students' test performance results due to the interactive nature of the computerized formative assess-

the future assessment participation.

Partially observed Markov decision process

A partially observed Markov decision process, or POMDP, is a generalized extension of the Markov decision process, in which the states are not completely observed (Krishnamurthy, 2016; Papadimitriou & Tsitsiklis, 1987; Thorbergsson & Hooker, 2018). Similar to the underlying Markov decision process in which the problem was formalized with the state (S), action (A), state transition probability (T), and the reward function (R), the POMDP introduces the observations (Ω), and the conditional observation probability function, (O). These six elements, (S, A, O, T, R, Ω), could represent the problem in a POMDP setting. More specifically, at each time point t, the environment is in some state s. The agent chooses an action $a \in A$, which causes the environment to transition to state $s' \in S$ with probability $T(s'|s, a)$. At the same time, the agent receives an observation $o \in \Omega$, which depends on the new state of the environment with probability $O(o|s', a)$. Finally, the agent receives a reward, $R(s, a)$, based on the chosen action. In addition, unlike a typical Markov decision process, the POMPD introduces a new concept called a belief state (b). The belief state refers to the probability denoted to represent the most likely state that the agent is in currently. In other words, given the K number of states that the agent could be in, the belief state of the agent could be denoted as follows:

$$b = [b(s_1), b(s_2), b(s_3), b(s_4), ..., b(s_k)], \quad (1)$$

where $s_j \in S(1 \leq K)$. Because the agent cannot fully observe the current state, the belief system approximates the current state using a probability distribution. In this way, the agent could summarize its experience to a belief state. Then, the agent keeps updating its belief state using the last action (a_{t-1}), current observation (o_t), and the previous belief state (b_{t-1}). Figure 2 demonstrates a visual representation of the update rules.

The probability of observing a certain observation o and the state change to s' following a certain action a can be expressed as follows:

$$P(o|a) = \sum_{s \in S} b(s) \sum_{s' \in S} P(s'|s, a) P(o|a, s'). \quad (2)$$

Then, the belief state is updated to $b(s')$ as the state changes, based on the three elements introduced earlier, the last

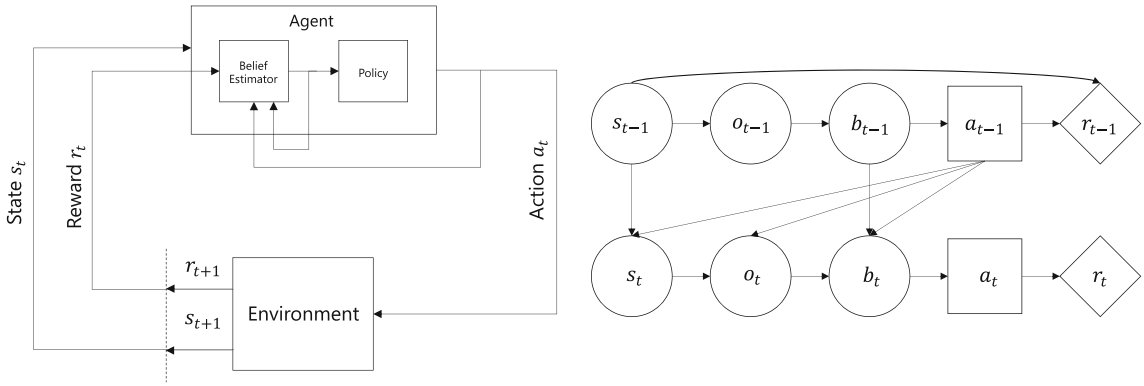

Fig. 2 A conceptual and a graph representation of the POMDP model

action (a_{t-1}), current observation (o_t), and the previous belief system (b_{t-1}).

$$\begin{aligned} b(s') &= P(s'|b,a,o) \\ &= P(o|a,s') \sum_{s \in S} b(s) P(s'|s,a) / P(o|a) \end{aligned} \tag{3}$$

The goal is identical to the Markov decision process (Feinberg & Shwartz, 2012), which attempts to maximize the accumulated reward in the long run. However, unlike how the policy was represented as a function of a state, the policy in a POMDP is a function of a belief state, b. In other words, solving the POMDP is formalized as a problem to estimate the value function of a belief state, rather than a state as in Markov decision processes. In summary, the value of the belief state given some policy π can be expressed as the cumulative reward as expressed in Eq. 4. Then, given an optimal policy π^* which is defined to achieve the highest long-term reward, we could apply the Bellman's equation (Bellman, 1954) to express the recursive value function of a belief state, which can be expressed as in Eq. 5,

$$v^{\pi}(b) = \sum_{t=0}^{\infty} r(b_t, a_t) = \sum_{t=0}^{\infty} \gamma^t E[R(s_t, a_t)|b_0, \pi] \tag{4}$$

$$\begin{aligned} v^*(b) &= max_{a \in A}[r(b,a) \\ &\quad + \gamma \sum_{o \in \Omega} P(o|b,a) v^*(\tau(b,a,o))] \end{aligned} \tag{5}$$

where $r(b,a)$ represents the immediate reward and γ represents a discounting factor to provide weights to the delayed or immediate rewards. Hence, when γ approaches 0, the agent will weigh the actions that could yield maximum immediate reward, while the value closer to 1 would prioritize maximizing the accumulated future rewards. Then, $\tau(b,a,o)$ refers to the belief state transition function, indicating how the current belief state b_t changes based on the acquired action a and observation b.

Present study

Previous studies have unanimously agreed on the potential of computerized formative assessments to provide personalized and adaptive learning in the classroom (Christ et al., 2012; January et al., 2019; Sharkey & Murnane, 2006; Wongwatkit et al., 2017). However, systematically deciding when (i.e., timing) and how often (i.e., density) to administer the assessments to students has remained an unresolved issue, in part, due to the competing perspectives in optimal test scheduling (Christ et al., 2012; January et al., 2019; Mellard et al., 2009; Sharkey & Murnane, 2006; Sherrington, 2018; van den Berg et al., 2018). In order to provide a more systematic solution to this issue, we adopted the reinforcement learning framework introduced by Nurakhmetov (2019) to the problem of selecting the optimal test administration for each student. We introduced an intelligent recommendation system to optimize the test administration decisions for computerized formative assessments.

We defined the optimality of the test administration using two criteria. First, the optimal test administration should be provided with the minimum number of assessments administered to each student. Second, the amount of information obtained from the subsequent assessments should be maximized to secure the diagnostic properties of the assessments. That is, a student's performance scores from one test administration to another should demonstrate adequate change. Overall, the explored recommendation framework focused on providing a balance between reducing the number of tests while securing sufficient information about the student's performance improvement. Hence, in our study, we formalized our problem as a sequential decision-making framework and attempted to formulate and solve the problem using the POMDP framework with reinforcement learning approaches. We used the actor–critic algorithm to recommend the optimal test administration points for each student.

Using reinforcement learning in test optimization could provide practical and methodological benefits to the test

optimization literature. First, unlike other neural approaches in machine learning, reinforcement learning could best synthesize the solutions to the problems where the interactions and the dynamics of the situation change the sequences of decisions drastically. For instance, in the test administration optimization, the teacher should interact with a student's test performance to make decisions about whether to continue assessing them and when to evaluate them again in a timely manner. This is a highly complex learning setting where the immediate reward (to the test a student or not) should determine the sequence of the process. This objective can be sufficiently delivered in the modeling and construction of the algorithm. In other words, reinforcement learning provides a flexible framework to best imitate the decision-making process of a teacher. For instance, the teacher will first attempt to observe the performance of a student and start making decisions about whether to test them or not based on the reward they are observing (e.g., is the score increasing or decreasing?). Also, multiple objectives can be easily provided as a constraint to help the system make decisions while providing viable solutions to the teacher, such as not over-testing the students. Second, previous studies have been conducted to understand and explore the capacity of reinforcement learning algorithms in various domains. Still, relatively few studies have been introduced to maximize the modeling and predicting the capacity of reinforcement learning to solve the problems and issues in educational settings. Hence, the current study could provide a novel example of how the use of reinforcement learning could benefit the advancement of artificial intelligence in education.

Three primary sections are introduced to clearly communicate the theoretical framework of the current recommender system design (Method), demonstration of the framework's capacity with the empirical dataset (Experiment), and how the system could effectively help administer computerized formative assessments (Results). Also, we discuss the practical and methodological implications and contribution of the study (Conclusion and Future Direction).

Methods

In this section, we introduce a set of theoretical and methodological guidelines, which were used to construct the current system's framework. More specifically, we describe the task of optimizing the administration of computerized formative assessment using a partially observed Markov decision process framework. Then, we provide a specific reinforcement learning algorithm used to optimize test administration, which is the actor–critic algorithms. Last, we present the performance evaluation framework of the system's capacity to effectively optimize test administration.

Task description

The main objective of our recommendation system was to select the minimal number of critical test administration points for each student during the school year. Hence, our system was supposed to utilize the information regarding students' previous and current test performance, to provide recommendations regarding a sequence of optimal future test administration points, or timing. To make sure that our system does not support the prolonged and excessive number of tests, it was required to pre-select a small number of test administration size (e.g., the maximum number of tests administered for each student). Then, within the allowed number of test administrations, our system was supposed to find the optimal sequence to administer assessments to each student. Moreover, a selected sequence of test administrations should accurately measure the progress of students' learning by predicting students' achievement at the end of the school year as well as their overall learning-progress slopes with high accuracy. This way, we could provide optimal test scheduling for individual students with a drastically reduced number of test administrations while preserving the diagnostic value of computerized formative assessment. In other words, students' test performance at the recommended "critical points" of test administration should accurately reflect student
resenting the
student
of the school
year ar
emic growth
(i.e., sl
used as the
criteria
ion, or label,
for eac

To
the current
study i
schedule is
optimiz
ons about the
student
naximization
of the
tem to learn
how to
interventions
or feed
ure success.
To do
or the local
positive
and thereby,
letting
ment. If the
student
ar growth in
their p
e to measure
them a
ch as at the
very b
year, and at
the end
's scores are
fluctuat
the student's
progres
es adequate

The main goal of the recommendation system is to minimize the number of critical test administration points for each student throughout the school year while still accurately measuring their learning progress. The system uses information from students' past and current test performances to recommend an optimal sequence of future test timings.

To avoid excessive testing, the system pre-sets a limit on the number of tests a student can take. Within this limit, the system finds the most effective times to administer tests that reflect the student's learning progress accurately. The selected test points should:

1. Represent the student's final performance score at the end of the year.

2. Accurately track the student's learning growth (or progress slope).

The system's objective is to provide the best diagnostic decisions by maximizing the positive changes in student scores. It continuously monitors performance changes, ensuring appropriate interventions or feedback are provided.

• For students with steady progress, the system may schedule tests only a few times (beginning, middle, and end of the year).

• For students with fluctuating scores, more frequent monitoring is required to ensure they receive adequate support.

Ultimately, the system aims to reduce testing to maximize instructional time while ensuring the tests capture students' growth and learning outcomes effectively.

support (or intervention) to improve their scores. Still, minimizing the number of test administrations would help the teacher secure more instructional time. These are the primary reasons underlying the goal of minimizing the number of the test while maximizing the positive score growth.

Task formalization with POMDP

Due to such unique problem conditions, our system must provide prediction decisions regarding students' progress dynamically, instead of making a decision based on a full vector of scores (i.e., full student performance history during the school year). Therefore, our task is formalized into a POMDP model problem referencing on Nurakhmetov (2019)'s reinforcement learning framework. Our task of identifying the best test administration schedule for each student while minimizing the number of test administrations required specifications of the POMDP elements to adequately describe the current scenario.

More specifically, we defined each episode to represent a test administration. During each episode (t), students' previous and current test scores are converted into a test history. The test history stores all possible combinations of the observed test performance scores into a set of sequences. For instance, a student's test history at the episode (T=3) could be represented as a set, such as $((S_1), (S_2), (S_1, S_2), ())$, where S_i is the student's performance score at time point i. Test scores were considered comparable across all time points in this computerized formative assessment. Then, the system has to map such performance history with the final student success classification label, whether the student could demonstrate adequate learning progress with the acceptable final performance. To do so, we need a classifier that allows an input of sequences to map it to an output label, such as the recurrent neural networks (RNN) or the long short-term memory networks (LSTM).

We compiled such specific elements to formalize our problems using POMDP. More specifically, the action (A) of an agent is provided regarding the choice of locating among the possible future test administrations. In other words, the agent decides which test window to select for the next test administration (a_t). Next, the state (S) represented the test administrators' decision-making process. The state at each time point (S_t) is defined by three components: (1) the current test score of the student, x, (2) the test score history of the student up until the current time point, h_{t-1}, and (3) the reward (R). As mentioned earlier, the test score history is represented as a set of possible sequences of the student's previous performance scores to preserve students' learning trajectory information. Using such information, the system may classify the internal belief of the state, b_t with a sequential classifier with memory networks. In the current study, we used the Gated Recurrent Units as an internal belief classifier. Thus, the new observation determining the selected test window can be expressed as a task to update the internal belief using the three elements, $O_t = b(x, h_{t-1}, K)$. The results of this will be represented as a probability of how such observations will be labeled at the end of the school year regarding the students' performance success. Thus, the last component, the reward (R), was defined based on the classification results and this is further discussed in detail in the following section.

System development using actor–critic algorithms

Among the three categories of popular reinforcement learning algorithms –*actor only, critic only, and actor–critic*–, actor–critic algorithms attempt to take advantage of the other two models. For example, the actor–critic algorithms include a critic who can evaluate the current policy and an actor who could produce continuous actions like actor-only algorithms (Grondman, 2015). In the actor–critic algorithm, the agent is separated into two roles, which are the actor and the critic. The major role of the actor is to iterate to locate the current best policy, in other words, explore the best and the vast options in the environment. The actors often are represented with some neural networks to predict the best action given the state. Then, the critic adjusts the behavior of the actor by providing adequate feedback about the chosen action. Therefore, the critic can be expressed as another type of function approximator that produces the action value, or the advantage value based on the environment and the action chosen by the actor.

Based on the interaction of the two agents, the ultimate objective of the actor–critic algorithm is to maximize the expected reward based on some parameterized policy. Again, policy (π) determines the probability of a certain action given the current state. Similar to how supervised algorithms are optimized, we can apply the gradient ascent (or descent) to optimize the policy, using a policy gradient. Hence, given the value of the current action in the state ($Q^{\pi}(s_t, a_t)$) and the average baseline return ($V^{\pi}(s_t)$), the policy gradient, $\nabla_{\theta} J(\Theta)$, can be expressed as:

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_{\theta} \log \pi_{\theta}(a_{i,t}|s_{i,t})(Q(s_{it}, a_{it}) - V(s_{it})) \quad (6)$$

$$Q^{\pi}(s_t, a_t) = \sum_{t'=t}^{T} E_{\pi_{\theta}}[r(s_{t'}, a_{t'})|s_t, a_t] \quad (7)$$

$$V^{\pi}(s_t) = E_{a_t \sim \pi_{\theta}(a_t|s_t)}[Q^{\pi}(s_t, a_t)] \quad (8)$$

More specifically, the current algorithm, in particular the critic, attempts to maximize the advantage function. The advantage function, $A^{\pi}(s_t, a_t)$, approximates how good the

reward produced from the current action, $Q^{\pi}(s_t, at_t)$, when it is compared to the baseline, $V^{\pi}(s_t)$.

$$A^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) - V^{\pi}(s_t) \tag{9}$$

In the current study, it is critical to modify and control the maximum number of test administrations to investigate the possibility of minimizing the number of test administrations. Hence, we attempted to mitigate this problem by changing the number of the terminal episode (T) for each round of experiments to understand the capacity of the algorithms to minimize the number of test administrations without losing much information about students' performance. In terms of the reward, the agent is provided $r = 1$ or $r = -1$ depending on whether the classification was correct about students' group classification, only at the terminal state, T. Also, in the current study, we used the discounting value closer to 1 ($\gamma = 0.99$) to ensure that the agent considers for the accumulated rewards over the episodes.

Performance evaluation

The performance of our recommendation system based on reinforcement learning will be evaluated based on three criteria. First, we will focus on whether the system could accurately classify the students based on their performance at the end of the school as well as their performance growth throughout the school year. More specifically, we identified the students who were at risk based on their pace of improvement in each assessment as well as their final score point at the end of the school year. For instance, if either a student's final assessment score was below the 25th percentile or his/her median score growth (i.e., slope) was below the median score growth of the entire sample, then the student was identified as at-risk (1), while the rest was identified as a successful or an exemplary student (0). Thus, the capacity of our system to correctly identify the students who are at-risk with a relatively limited amount of information from the restricted number of test points is evaluated. This evaluation metric could help us understand the diagnostic properties that could be preserved with limited test points by locating "critical test points" using our system. Second, we used the total amount of observed score change as a single evaluation criterion. This was under the expectation that the amount of observed score change should not significantly decrease after reducing the number of tests administered to each student to preserve the diagnostic value of the assessments. Similarly, we used the average amount of positive score change (APSC) observed in each test administration to evaluate the importance of a given test point, and this was included as the final evaluation criterion of the system performance. Formally, the three evaluation criteria can be expressed as follows:

$$Acc(\%) : \frac{Correct_{success} + Correct_{atrisk}}{N_{success} + N_{atrisk}} \tag{10}$$

$$APSC : \frac{1}{N}\sum_{i}^{N}\sum_{j}^{K}(S_{ij} - S_{ij-1}), \tag{11}$$

$$APSC/test : \frac{1}{K}\frac{1}{N}\sum_{i}^{N}\sum_{j}^{K}(S_{ij} - S_{ij-1}), \tag{12}$$

where i is the index for the students ($i \in \{1, ..., N\}$) and j is the index for test administrations ($j \in \{1, ..., K\}$).

To delineate these evaluation criteria, assume a student who participated in the formative assessment five times within a school year and scored 95, 97, 103, 100, 96 in the assessments. Based on the change between the student's scores in consecutive assessments, the total positive score change can be calculated as $(97 - 95) + (103 - 97) + (100 - 103) = 8 + (-3) = 5$. Hence, the total positive score change for this particular student would be 5. As the student participated in the assessment five times, then the total positive score change per test would be $5/5 = 1$. While evaluating the positive score changes, the system will select the most appropriate number of test administrations and the timing of test administrations. This decision would be evaluated based on whether the student's performance on the selected test evaluation windows could successfully identify at-risk students. This step will be repeated for the entire sample of students and the average of the total positive score change and the score change per test administration will be evaluated.

Experiment

In this section, we demonstrate the capacity of the current system in optimizing the administration in computerized formative assessment using an empirical dataset. We provide information about the experimental settings. The characteristics, descriptive statistics, as well as evaluation criteria related to this experiment, are provided in detail.[1]

Participants and the instrument

The sample of this experiment consisted of 727,147 fourth-grade students in the U.S. who participated in computerized reading assessments throughout the school year. The

[1] https://github.com/jinnieshinufl/Personalized-Test-Scheduling-RL

reading assessment was designed to measure a variety of reading skills, such as vocabulary knowledge, comprehension strategies, and analyzing literary text. The acquired data included students' scores from multiple test administrations during the 2017–2018 school year. The data sets analyzed during the current study are not publicly available due to proprietary reasons.

The current dataset is a fully adaptive computerized adaptive assessment. The formative assessments aim to closely monitor students learning and competencies in reading comprehension. A broad range of reading comprehension skills is measured based on the grade level. For instance, the five key domains of students' word knowledge and skills, comprehension strategies and constructing meaning, analyzing literary text, understanding the author's craft, analyzing an argument, and evaluating text. The test is constructed with multiple-choice questions and the students commonly take around 15 minutes to complete the text, on average.

Data processing and test window selection

Students could participate in the assessment following the teachers' guidance without any specific limitation in terms of the frequency or the density of their participation. Given that the empirical dataset consists of students of grades 1 and 2, the teachers play important roles to set assessment time, send students to participate in the test, and administer the test during the school year.

In fact, we could identify quite extreme cases of participation patterns, where a noticeable size of students would participate in the assessment multiple times within a few days. Also, there were relatively high proportions of students who actively participated in the assessments earlier in the school year and not taking part in any further assessments until the last couple of weeks of the school year. This caused a quite complex problem of having vastly large test scheduling options to recommend, as students could individually participate in the assessment on a daily basis during the school year. More specifically, given that the aim of the scheduling system is to provide a sensible recommendation regarding the test time point within the school year, providing an exact date for individual students as critical test points seems less feasible. Also, we identified that providing a general testing period, or a test window, rather than an exact day-based recommendation, could provide more flexibility to test administrators. To provide a reasonable solution to this problem, we organized students' test participation history and their performance from the start (August 2017) until the end (May 2018) of the school year, with roughly two windows in each month. This way, we could limit the number of possible recommendations that our systems could select from (19 test windows). The selection of the intervals between the test windows (i.e. 2 weeks to a month) was chosen to find relatively large sample sizes for each test window period, while considering the general administrative schedules of a typical school year with a consideration of school breaks, such as Christmas and the spring break.

To summarize, the input of the test schedule optimization system is a vector of students' test performance in the available 19 test windows. For instance, if a student participated in the exam during the school year six times and achieved the score of 978, 979, 1000, 992, 995, and 1002, the performance vector would be as follows:

[nan, nan, ***978****, nan,* ***979****, nan, nan, nan,* ***1000****,* ***992****, nan, nan, nan,* ***995****, nan, nan, nan, nan,* ***1002*** *]*

The performance vector places the student's performance information considering the schedule of his/her test administration. Thus, the missingness in the dataset, for instance, if the student did not participate in some of the tests (i.e., no score information), will be coded as not applicable, or nan.

Figure 3 provides an overview of the number of test participants and their average score in the resulting 19 test windows throughout their school year. The total number of students who participated in each test administration window fluctuated quite dramatically across the school year. For example, the highest number of participation occurred at the first test window of the school year (i.e., early August), since most teachers wanted to know their students' levels at the beginning of the school year. Then the participation rate fluctuated until the students showed the lowest rate of participation in test window 9 (i.e., late December). In terms of the average performance score student achieved, the overall trend indicated a steady improvement in students' performance over the school year. The scale score for each assessment ranged from 600 to 1400. The average score of the students in their earliest participation window was 978. Then, the average score of the last administration window (i.e., late June) showed a noticeable improvement to 1007. While there were a few points of fluctuation, most students indicated an increasing trend (i.e., growth) as they acquired more reading knowledge throughout the school year.

At-risk student identification

Prior to the development of our recommendation system, we had to identify the students who showed adequate growth in the assessments that they participated in over the school year. These students (called the exemplary group) were selected based on their median score changes across the school year and the percentile rank of their score in their last participation. The students who showed faster score growth, or improvement, than the other students (above the median growth) and the students who were above the

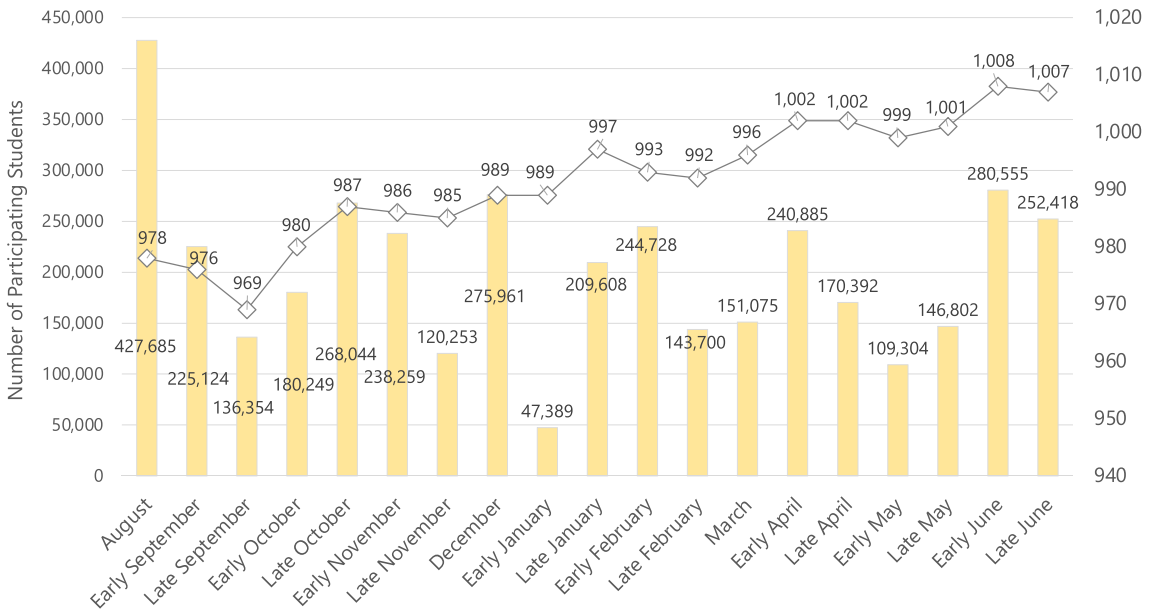

Fig. 3 Total number of participants and the average reading scores in each test window

25th percentile in their last assessment participation were selected as the exemplary group. Then, the rest of the students were considered the students who require some intervention, for instance, the change of administration scheduling, to closely monitor their performance. This step was to ensure that our recommendation system could learn from the exemplary cases to make recommendations for the other students for whom the number of test administrations might not be optimal or the timing of the test administrations might not be ideal. Based on the selection criteria, nearly 52.9% of the students were classified into the exemplary group, while the rest of the students were labeled as "at-risk". More information about the final class categorization is presented in Table 1.

In this study, we aimed to answer the main research question of whether our recommendation system could help minimize the number of test administrations without losing too much information about student performance. Hence, the experiment exploited the advantage actor–critic algorithm with our main research question formalized as a POMDP problem. More specifically, we separated the original data set into 90% for training, and the rest 10% for testing. Due to the large size of the original data set, with only 10% of the data set aside for testing, we could secure a large and representative sample of 72,714 students for testing the classification accuracy. Most importantly, the last assessment score of the students was removed, since it was used to derive the outcome label, which represents whether the student was exemplary or at risk.

Table 1 The final data categorization based on the performance grouping

	Participants	(%)
Exemplary	384,752	52.9%
At Risk	342,395	47.1%
Total	727,147	100%

Selection of test administration size

We selected a range for the maximum number of test administrations that could be administered to each student. This is equivalent to providing the terminal episode, T, so that we can investigate the effects of using a reduced number of test administrations. To locate a reasonable range, we first identified the average number of tests, or the test length, for the students. During the 2017–2018 school year, the students participated in the computerized formative assessment roughly five times (5.31) on average (Fig. 4). More specifically, students who were classified as 'at-risk' group participated in the assessment 5.48 times during the school year, while the 'exemplary' group students participated in 5.14 times. Using the average number of test administrations, we experimented with a varying batch size of 2 to 5. This is because we did not want students to participate in the test more than what is suggested from the average number of test administrations in the original data set. Also, participating only one time in the assessment could not provide enough information about students' improvement or growth. Hence, the test size of 1 was also omitted from the range. In summary, we experimented with limiting the size of the terminal state (T) using the selected range of batch size (i.e., 2 to 5 tests per student) to find

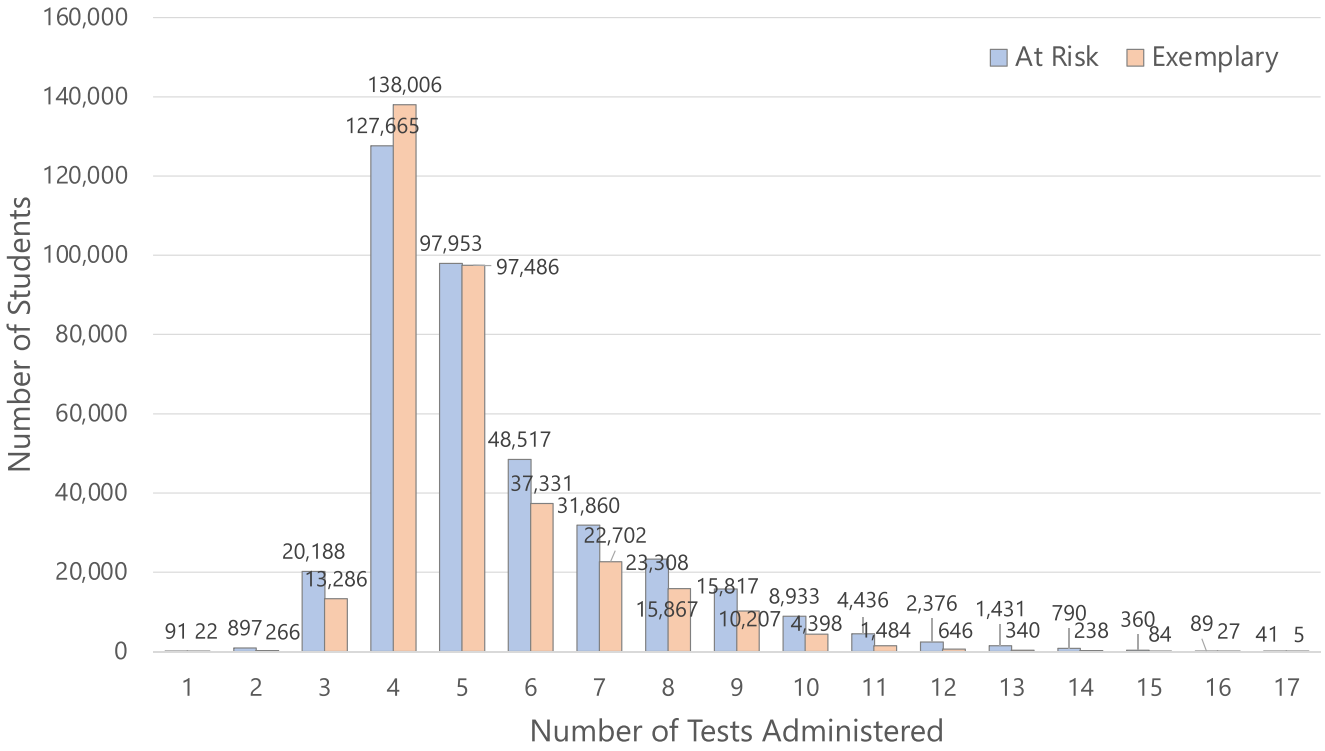

Fig. 4 Number of test participation during the 2017-18 school year

the optimal number of test administration and their specific administration points across the school year.

Evaluation criteria

To evaluate and compare the outcome of the varying size of the test window, we introduced two types of evaluation metrics. The first set of evaluation indices focused on the classification performance of the recommendation system. Therefore, we used two common classification indices, classification accuracy, and F1 score. Table 2 shows a confusion matrix for the binary classification of at-risk and exemplary students. Using the values in this matrix, the accuracy and F1 score were calculated as follows:

$$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)}, \text{ and} \tag{13}$$

$$F1 = 2\frac{\frac{TP}{(TP+FP)} * \frac{TP}{(TP+FN)}}{\frac{TP}{(TP+FP)} + \frac{TP}{(TP+FN)}}. \tag{14}$$

Second, to investigate whether the reduced number of test administrations or test window size affects the information acquired regarding students' performance improvement, we used the average positive score change (APSC) and the average positive score change per test (APSC per test). The main purpose of the computerized formative assessment is to ensure that students' reading performance gradually improves from early fall until the end of the school year. This knowledge or skill improvement should be observed in their score improvement or growth over the school year. Hence, given the current system is forced to minimize the number of test administrations to a certain degree, it is critical to evaluate the captured score improvement for each student. This could provide us with a good understanding of whether the recommended test administrations (i.e., recommended test schedule for each student) could serve the same purpose. More importantly, the average positive score change per test could represent how meaningful each test administration was in terms of capturing positive score growth or improvement. In short, the test performance of correctly classifying whether the student is at risk or not served as one type of evaluation criteria. Then, we evaluated how meaningful the selected test windows are based using the captured positive score improvement, or score change. We constructed the reinforcement learning algorithms using Python. We used TensorFlow to construct deep learning algorithms (e.g., LSTM, RNN) in the modeling process.

Table 2 The classification of at-risk and exemplary students

	Actual class (Real Data)	
Predicted (System)	At Risk (1)	Exemplary (0)
At Risk (1)	True Positive (TP)	False Positive (FP)
Exemplary (0)	False Negative (FN)	True Negative (TN)

Results

Table 3 presents the results for the recommendation system based on the aforementioned evaluation criteria. The results indicated that the proposed recommendation system could classify students based on whether they are at-risk or successful with fairly high accuracy. In terms of the classification results, we explored the accuracy and F1 score to investigate the classification accuracy of the current system. We found that the accuracy of the classification and the F1 score across the different test window sizes showed comparable results.

Regardless of the test window size, the recommendation system could achieve relatively high accuracy, close to 90%, and the F1 scores above 0.90. More specifically, we found that the system could achieve the highest accuracy and the F1 score when setting the maximum test size as 3 with an accuracy of 89% and the F1 score of 0.93. However, the differences with the other test window sizes are very minimal. In short, the results indicated with the drastically reduced number of test administrations, we could identify whether students were successful or at-risk with fairly high accuracy. In the most dramatic case, the recommendation system only required two test points to understand and categorize students correctly.

In addition, to investigate whether the selected test points were meaningful to preserve the information about students' performance sufficiently, we evaluated the recommended test points using two evaluation metrics, APSC and APSC per test. As explained earlier, APSC quantifies the total amount of positive score changes between the student's consecutive test administrations. Therefore, by comparing the APSCs of the original test points and the recommended test points, we could understand how optimal the recommended test points are in terms of preserving the student's growth in reading. More specifically, the original test points indicated that students could achieve about 33.35 positive score improvement across the test participation on average. This indicates a student with an average or normal, score growth produced about 33 score improvement regardless of how many times she or he has participated in the school year. On the other hand, the importance of each test administration is evaluated by dividing the observed positive score change by the number of test administrations that the student participated in. In this case, the students showed a 6.7 score change per test administration on average. For the current study, this would serve as a more meaningful indicator to understand and to provide rationales regarding the decreased number of test administration.

Surprisingly, amongst the varying sizes of test windows, we identified that test windows 2 and 3 provided higher positive score change per test, 11.33 and 8.03, respectively, compared to the original score of 6.7. On the other hand, for test window size 4 and 5, while the number of test administration has increased, the increase of accumulative positive test score was relatively small compared to the test size 2 or 3. In fact, the accumulated test score per test was smaller than the original test points, 6.7, indicating that each test point could only preserve a positive score change of 6.03 and 5.17 for window sizes 4 and 5, respectively.

This finding indicates that selecting window sizes 4 and 5 could not add much information to identify the student's overall growth. That is, after window size 3, adding more test windows (i.e., administrations) would not bring valuable information to increase the classification accuracy based on the student's performance and future success at the end of the school year. Moreover, the positive score change per test became noticeably smaller than the original test points when the test size was bigger than 3. Adding more test administrations after the third test could not change the observed positive test score large enough to indicate a meaningful contribution from the additional test administrations.

Table 3 Experiment analysis results based on the proposed evaluation criteria

	Test administration size			
Evaluation criteria	2	3	4	5
Accuracy	0.89	0.89	0.89	0.89
F1-score	0.90	**0.93**	0.92	0.90
Original Test Length (ave /stdev)	5.32 (1.80)	5.31 (1.80)	5.31 (1.79)	5.33 (1.81)
Original APSC	33.35	33.43	33.24	33.16
Original APSC per test	6.72	6.77	6.71	6.67
New Test Length (ave)	2.00	3.00	4.00	5.00
New APSC	22.66	24.10	24.10	**25.83**
New APSC per test	**11.33**	8.03	6.03	5.17

Bold text indicates the best result in each evaluation metric for each test administration size (e.g., test size 3 resulted in the highest F1-score amongst the other test size values)

From the findings, we could conclude that by selecting the test window size as 3, our recommendation system could identify students' performance categories (i.e., exemplary or at risk) with high accuracy, while preserving a sufficient amount of information to observe positive score change (i.e., growth) between consecutive test administrations.

Conclusions and future directions

The technology-enhanced advancement in classroom assessment has brought a new perspective on how data-driven decision-making could emerge within the framework of effective individualized learning. For example, using computerized formative assessments, teachers can administer assessments and provide students with feedback in a timely manner with more frequent test administrations (Sharkey & Murnane, 2006). Traditionally, increasing the number and density of test administrations in classroom assessments has been widely recommended to achieve a more accurate estimation of student performance and their progress (January et al., 2019; Christ et al., 2012). Administering classroom assessments more frequently provides a chance for teachers and other school-based professionals to collect more evidence to support student learning. Hence, with the significant ease of burden in test preparation and administration, a computerized formative assessment could effectively enhance the practices of frequent testing and close monitoring.

However, such advancements have also brought an interesting and novel challenge to classroom assessment practices. Several studies found that increasing the frequency and density of classroom assessment could not improve the diagnostic value of the assessments (Sharkey & Murnane, 2006; Sherrington, 2018; van den Berg et al., 2018). These studies indicated that an excessive number of assessments could fail to prioritize learning and negatively affect students' test-taking behaviors as well as their engagement with the assessments. In other words, while the teachers might be able to get a better idea of how students' progress changes based on the increased number of test administrations, students might feel alienated from learning to improve their performance as they are not provided a sufficient amount of time for actual learning experiences.

To mitigate such complex problems, more interactive and data-driven decision-making processes are required to provide both teachers and students with optimal assessment experiences. This requires a systematic solution to locate the most "critical points" of assessing individual students during the school year by utilizing the rich information regarding student performance accumulated by computerized formative assessment systems. Locating critical points entails that a smaller amount of data points could be used to make informed decisions about students' progress while preserving the original diagnostic values. Hence, we proposed a reinforcement learning-based recommendation system that could locate such critical assessment points for individual students. We investigated whether our system could make meaningful classifications regarding students' improvement in reading skills and whether students' academic success at the end of the school year could be captured while reducing the number of test participation significantly with a pre-selected test window size (2 to 5 times during a school year).

After evaluating the performance of the recommendation system based on various metrics, we could conclude interesting implications from our findings. First, the overall evaluation results indicated that with a significantly reduced number of test administrations, such as participating in only two assessments, our system could learn to identify important testing points for individual students to understand their learning progress, with the accuracy of 0.89 and the F1 score of 0.90. Increasing the test size up to five times did not significantly improve the accuracy, but we could achieve the highest F1 score when setting the test size as three times (F1 score of 0.93). Considering that the students, on average, originally participated in the assessments more than five times during the school year, we could conclude that our system could effectively identify 'critical' and 'meaningful' test points under a very restrictive condition. More specifically, for test sizes 2 and 3, the average amount of positive score chance, or score improvement identified for each test point based on our recommendation (11.33 and 8.03) was higher than that of the original test points (6.72 and 6.77). This reveals that with fewer test administrations during the school year, we could significantly increase the amount of score changes observed between the test administrations. In other words, the recommendation system could eliminate the test administrations in which students did not indicate significant score changes.

Second, we found that our system could provide accurate recommendations for both groups of students (i.e., at-risk and exemplary). Moreover, the effects of recommended test points were most significant among the students with excessive and prolonged test scheduling in their original participation. In fact, 35.8 and 27.2% of the students from at-risk and exemplary groups, respectively, participated in the assessment more than five times during the school year. Also, more than 5% of at-risk students (N = 18,456) participated in the assessment twice as many times as the other students, with some extreme cases participating more than 17 times during the school year. The number of such cases with excessive testing was relatively lower among the high-performing students, but still resulting in a significant number of students (2%, N = 7222). These results could lead to two important conclusions. First, we

could hypothesize that progress monitoring was conducted more closely with more frequent testing for students in the low-performing, or at-risk group, which resulted in relatively lengthier, and prolonged test participation. In this case, providing systematic test recommendations using our system could reduce the number of test participation, in most dramatic cases, from 17 times to 2 times. In fact, on average, the number of test administrations could be reduced to 3.32 for each student using following our recommendation. In addition, it should be acknowledged that a large number of cases still resulted in an excessive number of testing for high-performing students as well (N = 7222). This reflects how challenging and daunting the task is to optimize the test scheduling (e.g., the frequency and density of test administrations) for teachers, school psychologists, and other school-based professionals. Consistent with our previous findings, the recommendation system could recommend fewer test administrations (i.e., three fewer administrations on average) for each student in the exemplary group as well.

To conclude, we evaluated our system based on its capacity to find the most critical and informative test points to diagnose students' progress with the minimum number of test administrations. The results indicated that our system could perform fairly well in optimizing the test administration schedule for both at-risk and exemplary groups in terms of reducing the number of excessive and prolonged test administrations. This could provide important implications for practice with regard to computerized formative assessments. First, by applying our recommendation system, teachers would be able to identify students who are at-risk earlier in the school year with a much smaller number of test administrations. Therefore, students could be provided with more timely intervention programs to help improve their reading performance. Second, our recommendation system could also detect the high-performing students with a significantly reduced number of test points. This will provide significant benefits to the high-performing students by providing them more opportunities to explore and expand their knowledge in the classroom, rather than participating in repetitive test administrations. Also, the system can transfer its learning from a large amount of data (e.g., current empirical datasets) to provide recommendations for smaller cases of samples. As long as the assessment scores are comparable, the learning acquired from the large sample of data should be able to help optimize the performance and evaluation patterns of students in classrooms. This will help mitigate issues of introducing and adopting technology-advanced solutions in smaller classrooms (Taylor & Stone, 2009).

Furthermore, the novel framework of the current study provides an important methodological contribution to the existing literature. While reinforcement learning has been widely adopted and explored in many domains, it is still quite new in the domain of education research. The flexibility of the model algorithms and the capacity to model decision-processes from noisy datasets provide promising applications of reinforcement learning in educational issues (Intayoad et al., 2020). The flexibility and the prediction capacity of the current system could provide a novel research framework for future advancement in the field of application of reinforcement learning in education research.

Limitations and future research

While we carefully designed the experimental frameworks to thoroughly demonstrate the capacity of the current recommendation system, we identified a few concerns that might require further investigations. First, we evaluated the capacity of the recommendation system using reading assessment data collected from fourth-grade students. While our experimental data included a fairly large sample size to draw generalizable conclusions, we acknowledge that further research would be beneficial to investigate whether our system could yield comprehensive results in other subject domains with different progress trajectories (e.g., mathematics and science). Therefore, future studies would be encouraged to understand the capacity of the proposed system with other subject areas and grade levels.

Second, we encourage future studies on understanding whether the current system could be applied without a significant bias toward a specific sub-population of students with systematic differences in their learning progression. For example, previous studies have indicated that student-level variables, such as the students' language backgrounds (Gutiérrez & Vanderwood, 2013), grade level (Silberglitt et al., 2006), and gender (Kremer et al., 2016), could have significant impacts on students' reading performance trajectory over time. Hence, understanding whether our proposed system could provide an accurate and comprehensive recommendation for students with such diverse backgrounds would be highly beneficial. This, in fact, would help teachers and practitioners to better understand the capacity and the applicability of the current system in their diverse classroom settings.

Third, this study implemented a reinforcement learning approach that requires a large amount of interaction data to generate recommendations. Conventional algorithms such as collaborative filtering can be used for designing a recommendation system with smaller sample sizes. Therefore, future research should focus on item-based and user-based collaborative filtering for generating personalized testing schedules for computerized formative assessments.

Lastly, the current system's capacity was evaluated using a large dataset. While we could secure a relatively large portion of the data to provide a generalizable

understanding of the algorithm, future research on different schemes of evaluation would be beneficial. For instance, comparing whether the system could provide reasonable recommendations for high-achieving students (e.g., top 10% students) and low-achieving students (e.g., the bottom 10%). This would help further understand the capacity of the proposed recommendation system in various settings.

References

Andrade, H. L. (2019). A critical review of research on student self-assessment. In *Frontiers in Education*, (Vol. 4, p. 87): Frontiers.

Angus, S. D., & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? robust evidence from a large data set. *British Journal of Educational Technology*, *40*(2), 255–272.

Bellman, R. (1954). *The theory of dynamic programming*. Technical Report. Rand Corp Santa Monica CA.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

Black, P., & Harrison, C. (2001). Feedback in questioning and marking: The science teacher's role in formative assessment. *School Science Review*, *82*(301), 55–61.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90.

Buldu, M. (2010). Making learning visible in kindergarten classrooms: Pedagogical documentation as a formative assessment technique. *Teaching and Teacher Education*, *26*(7), 1439–1449.

Bulut, O., Cutumisu, M., Aquilina, A. M., & Singh, D. (2019). Effects of digital score reporting and feedback on students' learning in higher education. *Frontiers in Education*, *4*, 65. https://doi.org/10.3389/feduc.2019.00065

Bulut, O., Cormier, D. C., & Shin, J. (2020). An intelligent recommender system for personalized test administration scheduling with computerized formative assessments. *Frontiers in Education*, *5*, 182.

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education*, *21*(1-2), 83–113.

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children*, *78*(3), 356–373.

Dede, C. (2016). Next steps for "big data" in education: Utilizing data-intensive research. *Educational Technology*, 37–42.

Dopper, S. M., & Sjoer, E. (2004). Implementing formative assessment in engineering education: the use of the online assessment system etude. *European Journal of Engineering Education*, *29*(2), 259–266. https://doi.org/10.1080/0304379032000157187

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

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research, and Evaluation*, *14*(1), 7.

Feinberg, E. A., & Shwartz, A. (2012). *Handbook of Markov decision processes: methods and applications* Vol. 40. Berlin: Springer Science & Business Media.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ..., Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160.

Gierl, M., Bulut, O., & Zhang, X. (2018). Using computerized formative testing to support personalized learning in higher education: An application of two assessment technologies. In Zheng, R. (Ed.) *Digital technologies and instructional design for personalized learning*, (pp. 99–119). Hershey: IGI Global.

Gierl, M. J., & Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement*, *42*(1), 42–57.

Grondman, I. (2015). *Online model learning algorithms for actor–critic control*. Ph.D. Thesis, Technische Universiteit Delft.

Gutiérrez, G., & Vanderwood, M. L. (2013). A growth curve analysis of literacy performance among second-grade, Spanish-speaking, English-language learners. *School Psychology Review*, *42*(1), 3–21.

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

Intayoad, W., Kamyod, C., & Temdee, P. (2020). Reinforcement learning based on contextual bandits for personalized online learning recommendation systems. *Wireless Personal Communications*, 1–16.

January, S.-A. A., Van Norman, E. R., Christ, T. J., Ardoin, S. P., Eckert, T. L., & White, M. J. (2019). Evaluation of schedule frequency and density when monitoring progress with curriculum-based measurement. *School Psychology*, *34*(1), 119–127.

Joyce, P. (2018). The effectiveness of online and paper-based formative assessment in the learning of English as a second language. *PASAA: Journal of Language Teaching and Learning in Thailand*, *55*, 126–146.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37.

Kremer, K. P., Flower, A., Huang, J., & Vaughn, M. G. (2016). Behavior problems and children's academic achievement: A test of growth-curve models with gender and racial differences. *Children and Youth Services Review*, *67*, 95–104.

Krishnamurthy, V. (2016). *Partially observed Markov decision processes*. Cambridge University Press.

Mannor, S., & Shimkin, N. (2004). A geometric approach to multi-criterion reinforcement learning. *Journal of Machine Learning Research*, *5*, 325–360.

McManus, S. (2008). *Attributes of effective formative assessment*. Washington: Council of Chief State School Officers.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research, and Evaluation*, *18*(1), 2.

Mellard, D. F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research & Practice*, *24*(4), 186–195.

Nurakhmetov, D. (2019). Reinforcement learning applied to adaptive classification testing. In *Theoretical and Practical Advances in Computer-based Educational Measurement*, (pp. 325–336): Springer.

Papadimitriou, C. H., & Tsitsiklis, J. N. (1987). The complexity of Markov decision processes. *Mathematics of Operations Research*, *12*(3), 441–450.

Redecker, C., & Johannessen, O. (2013). Changing assessment–towards a new assessment paradigm using ICT. *European Journal of Education*, *48*(1), 79–96.

Sharkey, N. S., & Murnane, R. J. (2006). Tough choices in designing a formative assessment system. *American Journal of Education*, *112*(4), 572–588.

Sherrington, T. (2018). *Assessment too often fails to prioritise learning - let's change that*. Guardian News and Media.

Silberglitt, B., Appleton, J. J., Burns, M. K., & Jimerson, S. R. (2006). Examining the effects of grade retention on student reading performance: A longitudinal study. *Journal of School Psychology*, *44*(4), 255–270.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *4*(1), 1–103.

Taylor, M. E., & Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, *10*(7).

Thomaz, A. L., Hoffman, G., & Breazeal, C. (2006). Reinforcement learning with human teachers: Understanding how people want to teach robots. In IEEE (Ed.) *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 352–357).

Thorbergsson, L., & Hooker, G. (2018). Experimental design for partially observed Markov decision processes. *SIAM/ASA Journal on Uncertainty Quantification*, *6*(2), 549–567.

Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, *9*, 2245.

van den Berg, M., Bosker, R. J., & Suhre, C.or.J.M. (2018). Testing the effectiveness of classroom formative assessment in Dutch primary mathematics education. *School Effectiveness and School Improvement*, *29*(3), 339–361.

Volante, L., & Beckett, D. (2011). Formative assessment and the contemporary classroom: Synergies and tensions between research and practice. *Canadian Journal of Education*, *34*(2), 239–255.

Webb, M., Gibson, D., & Forkosh-Baruch, A. (2013). Challenges for information technology supporting educational assessment. *Journal of Computer Assisted Learning*, *29*(5), 451–462.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375.

Wilson, K., Boyd, C., Chen, L., & Jamal, S. (2011). Improving student performance in a first-year geography course: Examining the importance of computer-assisted formative assessment. *Computers & Education*, *57*(2), 1493–1500.

Wongwatkit, C., Srisawasdi, N., Hwang, G.-J., & Panjaburee, P. (2017). Influence of an integrated learning diagnosis and formative assessment-based personalized web learning approach on students learning performances and perceptions. *Interactive Learning Environments*, *25*(7), 889–903. https://doi.org/10.1080/10494820.2016.1224255