mathematics

Article

Learning Analytics and Computerized Formative Assessments: An Application of Dijkstra's Shortest Path Algorithm for Personalized Test Scheduling

Okan Bulut [1,*], Jinnie Shin [2] and Damien C. Cormier [3]

[1] Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB T6G 2G5, Canada
[2] College of Education, University of Florida, Gainesville, FL 32611, USA; jinnie.shin@coe.ufl.edu
[3] Department of Educational Psychology, University of Alberta, Edmonton, AB T6G 2G5, Canada; damien.cormier@ualberta.ca
* Correspondence: bulut@ualberta.ca

Abstract: The use of computerized formative assessments in K–12 classrooms has yielded valuable data that can be utilized by learning analytics (LA) systems to produce actionable insights for teachers and other school-based professionals. For example, LA systems utilizing computerized formative assessments can be used for monitoring students' progress in reading and identifying struggling readers. Using such LA systems, teachers can also determine whether progress is adequate as the student works towards their instructional goal. However, due to the lack of guidelines on the timing, number, and frequency of computerized formative assessments, teachers often follow a one-size-fits-all approach by testing all students together on pre-determined dates. This approach leads to a rigid test scheduling that ignores the pace at which students improve their reading skills. In some cases, the consequence is testing that yields little to no useful data, while increasing the amount of instructional time that students miss. In this study, we propose an intelligent recommender system (IRS) based on Dijkstra's shortest path algorithm that can produce an optimal assessment schedule for each student based on their reading progress throughout the school year. We demonstrated the feasibility of the IRS using real data from a large sample of students in grade two (n = 668,324) and grade four (n = 727,147) who participated in a series of computerized reading assessments. Also, we conducted a Monte Carlo simulation study to evaluate the performance of the IRS in the presence of unusual growth trajectories in reading (e.g., negative growth, no growth, and plateau). Our results showed that the IRS could reduce the number of test administrations required at both grade levels by eliminating test administrations in which students' reading growth did not change substantially. In addition, the simulation results indicated that the IRS could yield robust results with meaningful recommendations under relatively extreme growth trajectories. Implications for the use of recommender systems in K–12 education and recommendations for future research are discussed.

Keywords: recommender system; shortest path; Dijkstra; reading; formative assessment; personalized learning

MSC: 05C12

Citation: Bulut, O.; Shin, J.; Cormier, D.C. Learning Analytics and Computerized Formative Assessments: An Application of Dijkstra's Shortest Path Algorithm for Personalized Test Scheduling. *Mathematics* **2022**, *10*, 2230. https://doi.org/10.3390/math10132230

Academic Editor: Joaquín Paredes

Received: 31 May 2022
Accepted: 22 June 2022
Published: 25 June 2022

1. Introduction

Over the past few decades, the paradigm shift from assessment *of* learning to assessment *for* learning has transformed teaching, learning, and assessment practices at all levels of education [1]. Today's modern educational systems emphasize the importance of improving student learning by using formative assessments that gauge students' academic progress rather than how much students know at a certain point in time [2]. In other words, the acquisition of academic skills is now viewed as a process, instead of an outcome. To this

end, results from formative assessments administered throughout the school year can help teachers document each student's academic progress, identify students who are in need of extra supports, and make appropriate instructional adjustments. Moreover, assessments administered earlier in the school year can be used for the purpose of screening students who are at risk of significant learning difficulties. Notably, the numerous benefits of formative assessments rely on a strong commitment to frequent, albeit relatively brief, assessments being administered to students. Therefore, to optimize the symbiotic relationship between instruction and assessment, teachers have increased their use of digital assessment tools, such as computerized formative assessments, in K–12 classrooms. This is, perhaps, unsurprising given their ability to efficiently generate and score meaningful assessments of student learning in a variety of academic domains and sub-domains.

In a broader educational context, data from computerized formative assessments can also be used for establishing a learning analytics (LA) framework where teachers can implement data-informed (or data-driven) decision making and optimize student learning in real time [3–5]. The primary goal of LA is to "exploit data generated in educational settings for purposes of optimizing learning and the environments in which it occurs" [6]. Researchers often build predictive LA models based on historical student data to help teachers make predictions about future educational outcomes (e.g., students' final course grades or course failure) and take appropriate actions [7]. However, predictive LA models can also be used more dynamically for making instructional decisions based on students' strengths and challenges. For example, the teacher can review how students' scores in computerized formative assessments change over time and generate timely feedback for individual students [8,9].

Although the increasing use of computerized formative assessments has improved the quality and quantity of data going into predictive LA models, it has also led to practical challenges, such as identifying which students need to be assessed, when they should be assessed, and how frequently computerized formative assessments should be administered. Generally, teachers are responsible for monitoring growth trajectories for all students and deciding the necessity of a future test administration and the optimal timing of the next assessment. However, in classroom settings, this task can be practically very challenging for teachers. Thus, teachers may have to follow a one-size-fits-all approach (e.g., monthly testing for all students) that disregards the pace at which each student acquires knowledge. In addition, following a frequent testing approach (e.g., weekly testing) would reduce the instructional time that students would receive in the classroom.

Results from previous studies suggested that teachers may need systematic guidance to determine an optimal assessment schedule based on each student's unique progress [10,11]. Although several researchers shared practical guidelines regarding the number, timing, and frequency of computerized formative assessments focusing on reading [8,12,13], there is no consensus on the number of test administrations and the testing frequency for computerized formative assessments. Furthermore, grade-level guidelines for administrating computerized formative assessments focus on students achieving expected growth and thereby fail to consider students with unique learning needs [10]. Therefore, either placing the onus on teachers to make all scheduling decisions or following general guidelines about assessment schedules may not be effective solutions for maximizing the benefits of computerized formative assessments.

Recent studies showed that intelligent recommender systems (IRSs) can be used to generate a personalized assessment schedule based on each student's progress [10,11]. The main goal of the IRS is to exploit existing information about items (or users) to provide suggestions of new items for users (e.g., what products to buy, what videos to watch, or what online news to follow) [14]. In the context of computerized formative assessments, "item" refers to test administrations to be recommended and "user" refers to students. The IRS could harness historical assessment data to learn about different growth trajectories among students and then use this information to recommend test administrations. As students participate in computerized formative assessments throughout the school

year, the IRS could predict the number, frequency, and time of future test administrations for each student. There are two key benefits of developing this type of system. First, it addresses concerns from parents and teachers by potentially reducing the amount of time that is dedicated to testing over the course of a school year (i.e., increasing instructional time while minimizing the possibility of over-testing). Second, it helps teachers to ensure that all assessment data collected from students are of high quality and are useful in the decision-making process.

In this study, we aim to demonstrate how to develop a recommender system that can produce optimal test administration schedules by minimizing the number of test administrations without sacrificing the quality of the data collected (i.e., minimizing the influence of error on the decision-making process). To develop the IRS, we propose a user-based collaborative filtering approach based on Dijkstra's shortest path first (SPF) algorithm [15]. The SPF is a greedy algorithm for solving single-source, shortest path problems (e.g., finding the shortest route to take from one city to another). We selected the SPF algorithm due to its lower computational cost and scalability within a large-scale LA ecosystem. In the following sections, we describe the mathematical model underlying our recommender system and then demonstrate the feasibility of our proposed IRS approach using real and simulated assessment data.

2. Intelligent Recommender System

2.1. Recommender Systems for Educational Assessments

A recommender system is a program that uses existing data to learn different characteristics of users (e.g., students) and items (e.g., course materials) and then recommend items or actions to new users [16]. To date, various recommender systems have been introduced by educational researchers and computing science communities to support instructional practices and promote student learning. Researchers have mostly attempted to use recommender systems for providing individualized guidance to students on educational content (e.g., courses, course modules, and learning resources). Manouselis et al. [17]'s edited book on recommender systems provides examples of innovative recommender system applications in the context of technology-enhanced learning (TEL), such as learning plan recommendations in gamed-based learning, learning resource recommendations, and learning object recommendations in adaptive learning systems. Some researchers also used recommender systems as a predictive modeling technique. For example, Thai-Nghe et al. [18] used recommender system techniques for predicting student performance (i.e., whether students successfully completed a particular step of the item on their first attempt) in intelligent tutoring systems. The authors found that recommender system techniques such as matrix factorization and collaborative filtering outperformed the traditional regression methods in predicting student performance. Recently, more advanced algorithms and approaches such as reinforcement learning [11,19] have been proposed to improve the capacity, architectural flexibility, and performance accuracy of the recommender systems adopted within educational contexts.

Despite their popularity in educational applications focusing on TEL, recommender systems have not been widely utilized in educational assessment contexts. To date, few studies have investigated how recommender systems could enhance formative and summative assessment practices in education. For example, de Oliveira et al. [20] proposed a recommender system for students participating in a computer programming course. The system created a user profile for each student based on their performance in formative assessments and recommended different classes of activities that students had to solve to improve their performance. Some researchers also integrated psychometric modeling approaches (e.g., item response theory (IRT)) into recommender systems to create personalized learning environments for students. For example, Baylari and Montazer [21] built a multi-agent e-learning system based on IRT and artificial neural networks. The system delivers a review assessment, analyzes the student's responses, diagnoses the student's potential learning problems, and then recommends the appropriate learning materials to

the student. The results of this study showed that the proposed system could recommend appropriate course materials with high degree of accuracy. Chen et al. [22] also proposed a personalized e-learning system combining IRT and recommender systems to provide adaptive learning opportunities to students. Their proposed system contained a course recommendation engine that dynamically selected appropriate course materials based on each student's ability level. Experimental results showed that the proposed system could improve students' learning efficiency and effectiveness by recommending course materials based on their ability levels.

To enhance the delivery and use of educational assessments in the classroom, de Schipper et al. [23] built and implemented a recommender system that provided automated and personalized feedback to secondary school students in the Netherlands. Using techniques such as singular value decomposition and collaborative filtering, the system recommended a set of personalized practice questions to students following a high-stakes summative assessment. The findings of this study showed that the recommender system could provide useful feedback to students in the form of personalized practice materials by making use of readily available data from summative assessments. Bulut et al. [10] proposed an IRS approach to optimize the scheduling of test administrations for a computerized formative assessment focusing on the mathematical abilities of students in grades K–12. The findings of their study indicated that the IRS could yield a significant reduction in the total number of test administrations required to make accurate decisions on students' academic growth. Similarly, Shin and Bulut [11] also proposed a recommender system using a reinforcement learning algorithm that aimed to optimize test administration schedules for students. The proposed system could identify the critical time points for students to demonstrate their academic growth. The results indicated that the algorithm could successfully reduce the number of test administrations without compromising the amount of information gathered about the students' learning progress. Furthermore, Kundu et al. [24] provided a comprehensive overview of how the big data in education, especially in educational assessments, enabled the adoption of recommender systems with various applicable examples.

In this study, we propose to use recommender systems to produce personalized test administration schedules for students participating in computerized formative assessments. Using data from computerized formative assessments (i.e., students' scores), our system aims to produce a test administration schedule that could enable students to demonstrate sufficient academic growth with the minimum number of test administrations. In the following sections, we describe the mathematical modeling approach (i.e., Dijkstra's SPF algorithm) underlying our recommender system.

2.2. Directed Graph to Represent the Test Administration Sequence

In mathematics, graph theory refers to the study of graphs to learn pairwise relations between the two or more entities. In our study, we used students' test performance history with the finite set of testing window as an individual entity in a directed graph. A directed graph, $G = (V, E)$, consists of a non-empty finite set V of elements called vertices or nodes. A finite set of E represents the edges with or without the weight to demonstrate the relationships between two nodes, V and V'. In our recommender system, we let $G = (V, E)$ to represent a directed graph for a personalized path to participate in a sequence of computerized formative assessments. Each node represents a single test administration within a particular testing window. For example, the following sequences (S) in Equation (1) represent students' test administration records (T) for a total of n students within k testing windows (e.g., semi-monthly testing windows throughout a school year):

$$A = \left\{ \begin{array}{l} S_1 = T_1 \rightarrow T_2 \rightarrow T_3 \rightarrow T_4 \rightarrow \cdots \rightarrow T_k \\ S_2 = T_1 \rightarrow T_3 \rightarrow T_4 \rightarrow T_5 \rightarrow \cdots \rightarrow T_{k-1} \\ S_n = T_3 \rightarrow T_4 \rightarrow T_5 \rightarrow T_{12} \rightarrow \cdots \rightarrow T_{k-2} \end{array} \right\}. \quad (1)$$

To identify the pairwise relationship between two nodes, we could define the edges and their association weights in multiple ways. The connecting edges represent the direct

sequence that students are involved in their test participation. For example, $T_1 \rightarrow T_3$ represents a student who participated in the computerized formative assessment in the first and third testing windows. To define a directed graph with respect to students' learning progress (i.e., academic growth based on their scores), the strength of the pairwise relationship between the two entities should represent how closely the assessment scores are related to each other. Hence, we used the average positive score change observed between the two entities (testing windows) to represent their relationship. Assuming that i and i' are two connected entities for student j, m is the highest assessment score, and u is the lowest assessment score that student j could achieve in a series of test administrations; the relationship between the two test administrations ($W_{ii'}$) can be expressed as follows:

$$W_{ii'} = log(L_{ii'}), \tag{2}$$

where $L_{ii'}$ can be computed as

$$L_{ii'} = \frac{1}{\frac{1}{(m-u)}\Sigma_{j=1}^{n}|T_{ji} - T_{ji'}|}. \tag{3}$$

In Equation (2), we transform the probability value ($L_{ii'}$) into a log value ($W_{ii'}$) for two reasons. First, since the probability values of $L_{ii'}$ are always between zero and one, the weights calculated in the log space will be negative values so that the shortest path can be assigned to the highest weight. Second, the inverse of the assessment scores was used to represent a "closer" connection between the two nodes when there is a significant academic growth (i.e., a large, positive or negative change between two assessment scores). Therefore, test administrations with sufficiently high (or low) assessment scores will be critical points to understand students' overall learning progress. Figure 1 illustrates how a test administration sequence could be transformed into a directed graph. The figure at the bottom provides a simple demonstration of how the association weights for the edges can be computed.

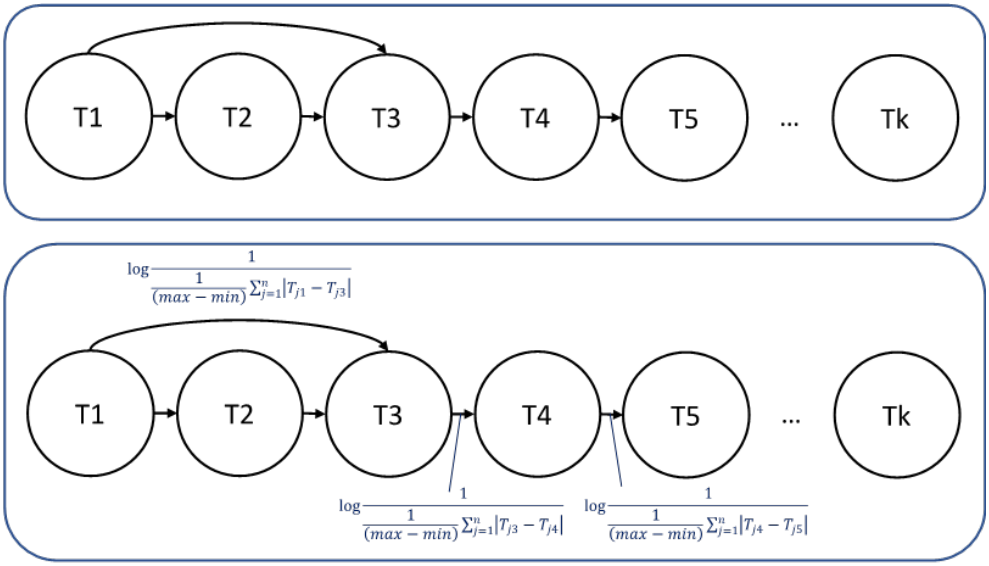

Figure 1. A directed graph representation based on the test administration sequence.

2.3. Shortest Path Similarity as Recommendations

In this study, we utilized shortest path similarity to identify the most efficient test sequences (or paths) that could be recommendable to other students. The SPF algorithm [15] identifies the shortest path from a starting entity to the target entity in a weighted graph. The algorithm creates a tree that examines the distance (i.e., path) from the starting entity to all other points (i.e., nodes) in the graph and then selects the shortest path as a solution.

In our case, the SPF algorithm is adopted to identify the shortest path between test administrations to produce an optimal test administration schedule for each student. We define the optimal test administration based on two criteria: the number of test administrations and the score change between test administrations (i.e., academic growth). We aim to select the minimum number of test administrations to reduce the instructional time that students are likely to miss while participating in a computerized formative assessment outside the classroom. Additionally, we aim to avoid test administrations in which students are not likely to demonstrate a significant score change from the previous test administration. That is, students should not be required to participate in any test administration until they are able to demonstrate adequate improvement.

To identify the shortest path between test administrations, we set the distance from one entity to another connected entity (i.e., test administrations from two testing windows) to zero for the initial node and to ∞ for all other nodes. This will serve as a tentative distance value between the entities (see step 1 in Figure 2). The SPF algorithm attempts to iteratively update this distance value starting from the initial node with the randomly assigned distance between the other nodes (i.e., $\{2, 4, 3, 5, \ldots\}$ in Figure 2). The algorithm updates the value until it is represented by the smallest weight. For example, the randomly assigned distance value between the first and the second node in Figure 2 is set to two. Hence, the SPF algorithm updates the distance value of the second node to two (i.e., $0 + 2 = 2$). In the second step, the third node's distance value is updated to six (i.e., $2 + 4 = 6$), but then it is corrected to have the smallest distance of value two due to its shorter distance relationship with the first node. The order of updating this rule for every entity (i.e., testing window) in the graph is controlled by a concept called a priority queue. The priority queue ranks the entities in a specific order based on their initial connecting weights. This helps identify which path needs to be updated first in order for effective iterations (for more information about the priority queue, see Chen et al. [25]).

As explained above, the SPF algorithm iteratively finds the paths that have the minimum weight, in other words, the shortest distance (see Figure 2). In our proposed IRS, the SPF algorithm updates the solution based on the relationship between different test administrations. That is, the algorithm finds the test administrations where the change in students' scores in the computerized formative assessment was the most significant. Note that we use the inverse value of the score change to represent their relationships so that the shortest distance corresponds to the largest score change. The SPF algorithm identifies a list of test administration paths connecting from one node to another with minimized weights. As more information on students (i.e., new scores in the computerized formative assessment) becomes available, the SPF algorithm adapts accordingly and continues to look for the shortest path (i.e., the test administration schedule with the smallest number of test administrations and maximum score change).

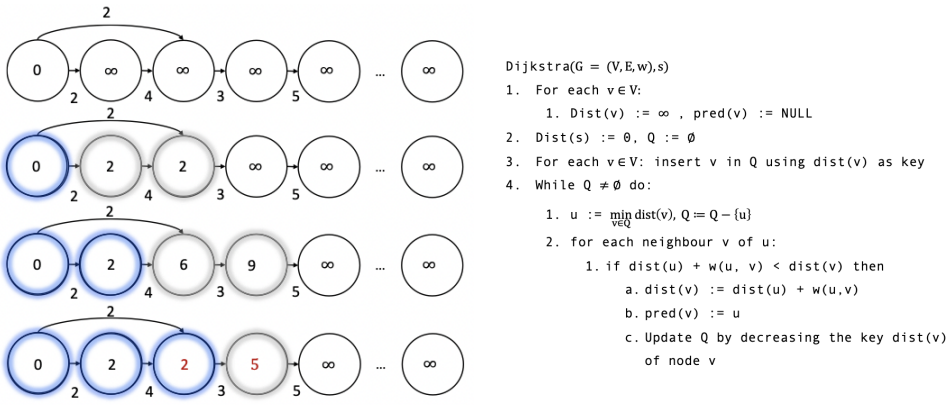

Figure 2. A conceptual representation of the shortest-path similarity update with the test sequence X (pseudo-code on the right).

In our study, we adopted the SPF algorithm to identify a set of test administrations that would be suitable for students who are at increased risk for reading difficulties based on their performance in a computerized formative assessment. The shortest path recommendations were derived from the assessment history of exemplary students who demonstrated sufficient academic growth throughout the school year (details of the exemplary and at-risk categorizations are available in the Methods section). First, we applied the SPF algorithm based on a directed graph generated using students' test administration sequences (i.e., students' assessment scores over multiple testing windows) and calculated the edge weights. Second, we mapped the shortest path results onto students' test performance information. In the final step (i.e., recommendation phase), we matched at-risk students and exemplary students who demonstrated a similar assessment performance within the same testing window using the Euclidean distance. Given two students, one from the exemplary group who followed the shortest paths, $S_u = (s_{1u}, s_{2u}, \ldots, s_{mu})$, and another from the at-risk group, $S_l = (s_{1l}, s_{2l}, \ldots, s_{nu})$, we iteratively identify the students by evaluating their Euclidean distance of their test score. That is, the nth student in the at-risk group, s_{nu}, is compared to the mth student $s_{m,l}$ from the exemplary group with the shortest path, and then a test administration recommendation is made by locating the testing sequence of the mth student that maximizes the Euclidean distance:

$$\underset{x}{\text{argmax}} = \sqrt{\Sigma_{i=1}^{k}}(T_{s_{nu},i} - T_{s_{ml},i})^2. \quad (4)$$

Using the SPF algorithm, we designed an IRS that could recommend individualized progress monitoring schedules, minimizing the number of test administrations without sacrificing the quality of the data collected (i.e., minimizing the influence of error on the decision-making process). To demonstrate how our IRS produces optimal test administration schedules for at-risk students, we conducted two studies: a real-data study and a Monte Carlo simulation study. In the real-data study, we used existing test scores from a large group of students who participated in a series of computerized reading assessments. Using the longitudinal assessment data obtained from the students, we considered a hypothetical scenario in which we explored which test administrations the IRS would recommend to at-risk students, compared with standard practice (i.e., actual test administration decisions made by the teachers). In the simulation study, we examined the performance of the IRS when students demonstrated unusual growth trajectories (i.e., negative growth, no growth, and positive growth followed by a plateau). The following research questions guided the real data and simulation studies:

1. Does the IRS yield optimal test administration schedules with the minimum number of test administrations?
2. Does the IRS produce robust recommendations for students with unusual growth trajectories (e.g., decreasing trajectory, flat growth trajectory)?

In the following sections, we explain the details of each study, summarize the results, and discuss the implications of our findings.

3. Methods

3.1. Real Data Study

3.1.1. Sample and Instrument

The sample of the real-data study consisted of students in grade 2 (n = 668,324) and grade 4 (n = 727,147) in the United States who participated in a number of Star Reading assessments during the 2017–2018 school year. Star Reading [26] is a computerized adaptive test that measures a variety of reading skills, such as vocabulary knowledge, comprehension strategies, and literary text analysis. The purpose of Star Reading is to provide meaningful information to teachers to inform their classroom instruction. It can also provide information about the likelihood that a student will progress well in response to classroom instruction throughout the year and perform well on the state test at the end

of the school year. Each administration of Star Reading consists of 34 multiple-choice items that can be completed within an average administration time of approximately 20 min. Star Reading can be administered to students with a sight-word vocabulary of at least 100 words from kindergarten to grade 12. Strong evidence of reliability and validity are described in detail in the Star Reading technical manual [26].

In the data, the total number and frequency of Star Reading administrations varied by students because teachers could determine the test schedule for their students. For example, a large number of students participated in Star Reading several times at the beginning of the school year and during the last couple of weeks of the school year. In addition, some students participated in the assessment very frequently (e.g., multiple times within a week). To make test scheduling recommendations based on this dataset, we decided to determine a reasonable testing window based on the amount of time required for students to demonstrate enough growth in reading. With a monthly testing window, there would be too much time for some students, especially those who may be struggling to demonstrate reasonable growth in reading. Similarly, a weekly testing window would not provide enough time for students to receive an adequate amount of instruction to be able to improve their reading skills. In addition, when recommendations are made from these testing windows, giving teachers a specific day or week for testing each student might pose a significant logistical challenge for them. Therefore, we decided that a two-week testing window would be a reasonable duration and organized students' test participation history as nineteen testing windows from the start (August 2017) until the end (June 2018) of the school year, with roughly two testing windows per month. The number of individual administrations of the Star Reading assessment are displayed in Figure 3.

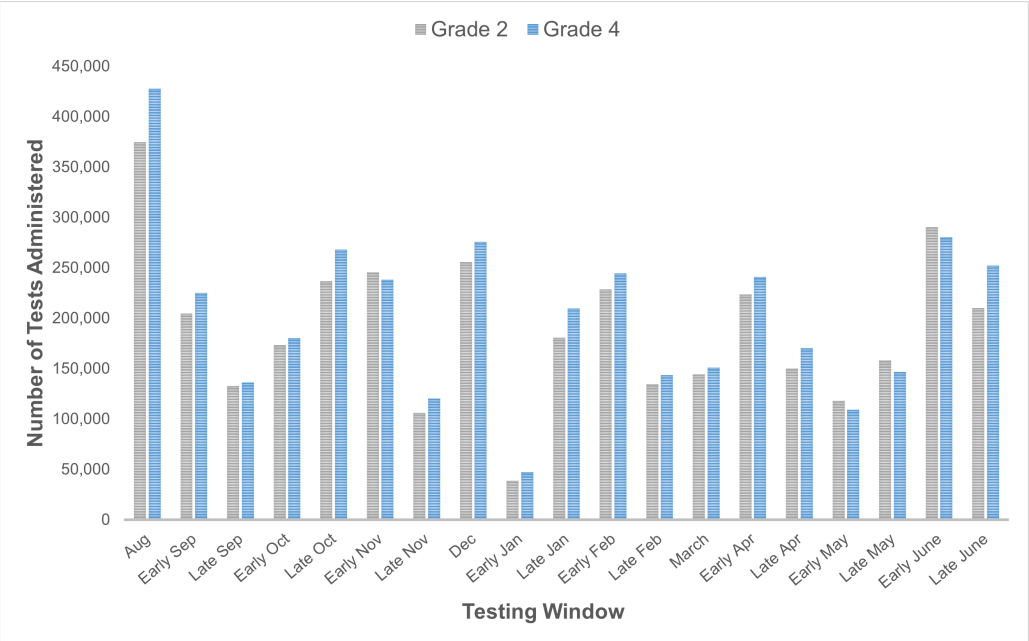

Figure 3. The total number of test administrations by testing window.

Star Reading scores are reported on a unified scale score metric ranging from 600 to 1400. Higher scores indicate better performance in reading. Figure 4 shows the average scale scores in Star Reading by testing window. For both grades 2 and 4, the average Star Reading score gradually increased as students acquired more knowledge and skills in reading. It should be noted that although students are expected to show a linear growth pattern throughout the school year, this does not necessarily occur for all students. Some students' scores may show a nonlinear pattern, a flat pattern (i.e., no improvement), or a decreasing pattern. Students with such growth patterns need to be closely monitored and provided interventions or remedial reading programs to improve their reading skills.

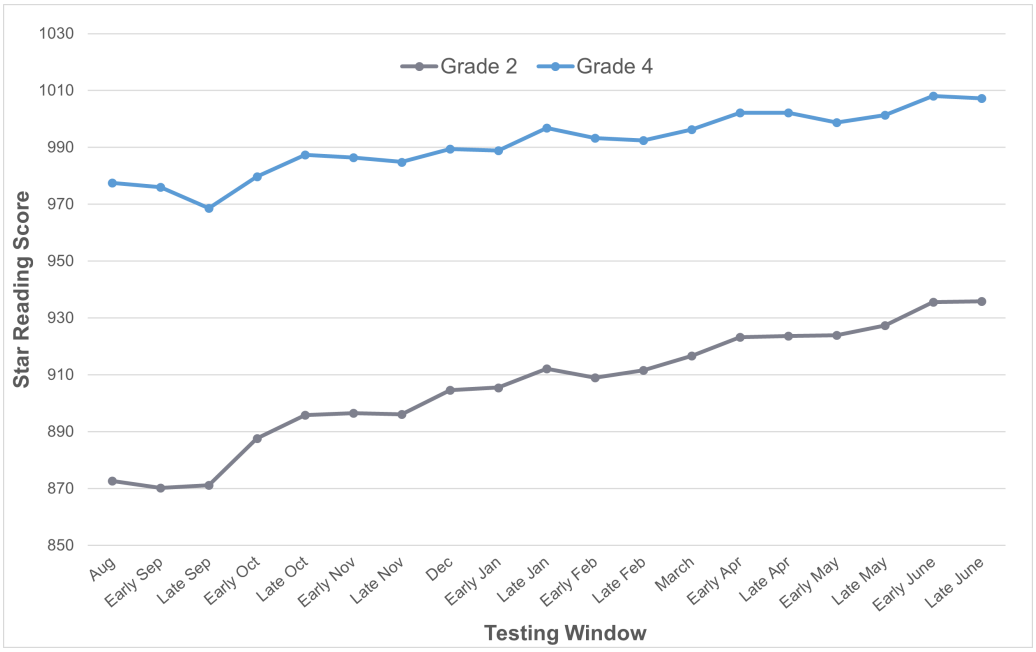

Figure 4. The average Star Reading scores in each test window.

3.1.2. Data Preprocessing

To build the IRS using the Dijkstra's SPF algorithm, we applied a data preprocessing procedure. First, we calculated students' academic growth in reading during the 2017–2018 school year. To estimate students' growth (i.e., slopes), we used the Theil–Sen estimation method [27]. The Theil–Sen estimator is similar to traditional regression approaches such as ordinary least-squares, but it is robust to outliers [8]. The following formula shows the slope calculation using the Theil–Sen method:

$$\text{Slope} = \frac{\text{Star Reading Score}_{\text{Time2}} - \text{Star Reading Score}_{\text{Time1}}}{\text{Date}_{\text{Time2}} - \text{Date}_{\text{Time1}}}, \quad (5)$$

where Star Reading $\text{Score}_{\text{Time 1}}$ and Star Reading $\text{Score}_{\text{Time 2}}$ are the student's Star Reading scores from two administrations, $\text{Date}_{\text{Time 1}}$ and $\text{Date}_{\text{Time 2}}$ are are the dates that the test administrations occurred, and Slope is the growth estimated based on the average daily change between the two scores. This calculation is repeated for all possible slopes for a given student, and then the median value of the calculated slopes is used as the best estimate of student growth.

Next, we split the sample for each grade level into two samples: training and test. The training set consisted of students who showed adequate growth in the assessments (hereinafter referred to as the *exemplary* group). To identify adequate growth, we used 2 rules: the student's Theil–Sen slope must be larger than the median value of all slope estimates in the sample and the student's final Star Reading score must be above the 25th percentile, which is considered a cut-off for identifying at-risk students in reading [11]. After applying these selection criteria, the grade 2 training sample included 276,087 students and the remaining 392,237 students were included in the test sample. For grade 4, the training sample included 278,442 students while the test sample included the remaining 448,705 students. The training set was used to train the IRS based on the exemplary students and to make recommendations for the students in the test set for whom either the number of test administrations or the timing of the test administrations might not be ideal (hereinafter referred to as the *at-risk* students).

3.1.3. Data Analysis

As explained earlier, we trained the IRS using Dijkstra's SPF algorithm with a priority queue [15]. The primary goal of the IRS was to find an optimal test schedule with the least number of test administrations and the maximum positive score change in Star Reading across the test windows. After a target student (i.e., students in the test dataset) participated in two administrations of Star Reading, the student's academic growth (i.e., Theil–Sen slope) in reading was calculated. Then, the Euclidean distance was used to identify a list of exemplary students who indicated similar growth trajectories as the target student within the same testing window. For the identified exemplary students, the positive score difference between subsequent testing windows was calculated. The larger the score difference, the shorter the distance between the testing windows. In the recommendation phase, the student whose test schedule yielded the largest positive score change was identified as the shortest path and their test schedule was recommended to the target student. After each test administration, the IRS recalculated the target student's growth trajectory and determined whether the student should stay in the current test schedule or switch to an alternative test schedule that is more aligned with their growth trajectory. In the real-data study, we established a hypothetical scenario in which we examined which test administrations the IRS would recommend to students in the test dataset. Since the number of test administrations varied by students, students did not have a valid score for each of the nineteen testing windows. Therefore, we used linear interpolation to estimate missing scores for the testing windows that the students did not participate in Star Reading. Our goal was to compare the number of test administrations in standard practice (i.e., testing decisions being made by the teachers) against the number of test administrations recommended by the IRS. Using the complete dataset, we evaluated the performance of the IRS based on three criteria: (1) the average number of test administrations, (2) the average positive score change between test windows, and (3) the minimum and maximum number of recommended test administrations. All of the analyses were conducted using Python version 3.8.0 [28].

3.2. Simulation Study

In the Monte Carlo simulation study, we aimed to evaluate the performance of the IRS in the presence of unusual growth trajectories in the data. We considered three unusual patterns of growth trajectories: negative slopes (i.e., learning loss), zero slopes (i.e., no growth), and plateau slopes (i.e., academic growth or loss tapering off gradually). For each pattern, we considered a variety of scenarios. The negative slope condition assumed that there was either a linear or quadratic decay (i.e., learning loss) in the data. The zero slope condition assumed that the students' slopes were either zero or very close to zero. The plateau slope condition assumed that the students' trajectory (either negative or positive) became flat after the 14th, 15th, or 17th test administration. Figure 5 illustrates the three growth trajectories considered in the simulation study.

A linear regression model was used to simulate assessment scores based on the aforementioned growth trajectories:

$$y = a_0 + a_1 * d + \epsilon, \tag{6}$$

where y is the assessment score, a_0 represents the intercept (fixed to 600 for all simulation conditions), a_1 represents the slope, d represents the number of test administrations, and ϵ is the residual. We set $d = 20$ to obtain a full dataset (i.e., scores available for all testing windows) and $d < 20$ to obtain a sparse dataset (i.e., assessment scores available only for some testing windows). The negative slope condition involved a single negative slope (linear) or a mix of gradually decreasing, negative slope values (quadratic). The minimal slope condition was based on a mix of slope values close to zero. The plateau condition involved scores following either a positive or negative slope until the 14th, 15th, or 17th test administration, and then a slope close to zero. Table 1 presents the specification of the

data simulated using the linear and quadratic regression models. For each condition, we simulated 1000 students using the mblm package [29] in R [30]. Then, the IRS from the real data study was applied to the simulated dataset to generate test administration schedules for simulated students.

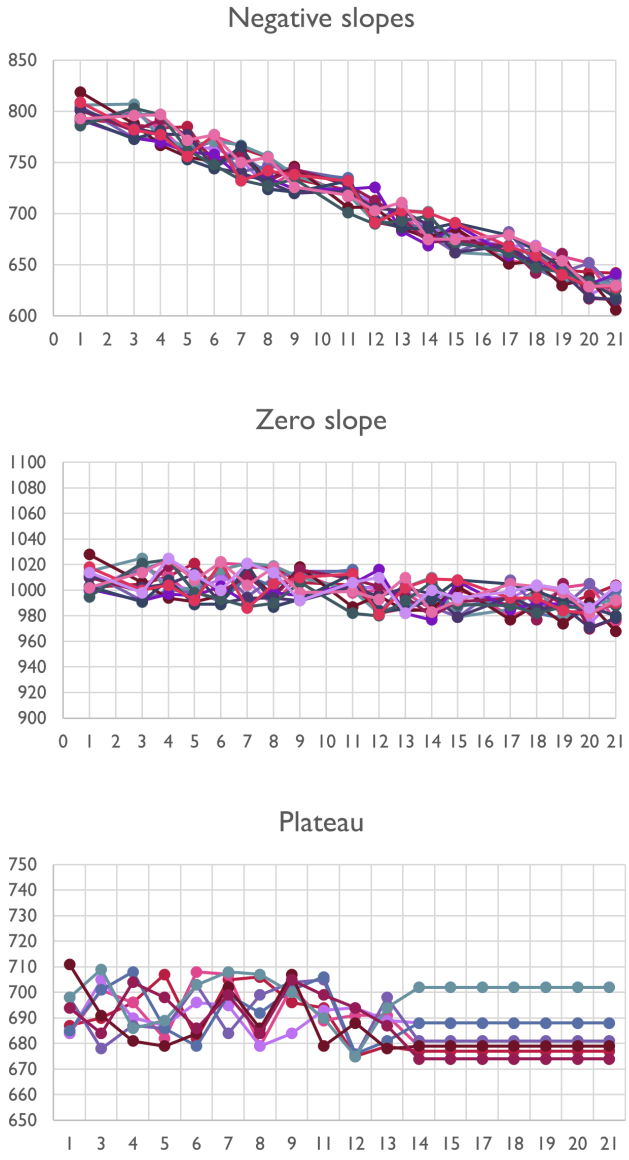

Figure 5. Academic growth trajectories in the simulated datasets.

The simulation study was conducted in four stages. First, for the sparse data (see Table 1), we attempted to generate the data points (students' scores) in randomly selected windows of $d \in \{3, 5, 10, 15\}$. For the full data, d was set to 20. The initial slope value was selected based on the three slope categories (i.e., negative, minimal changes, and plateau). For instance, for the negative slope category, the initial slope value was drawn from a uniform distribution following a selected range of $(-1.1, 0$; see Table 1). The residual was drawn from a Gaussian distribution, $N(0, d^2)$. Second, the linear regression model in Equation (6) was used to simulate the data. We first gathered the residual ϵ_d as the uniform deviated within the interval of $\epsilon_d \in \min(y - \hat{y}), \max(y - \hat{y})$. Then, the residual ϵ_d was added to the predicted value, $\hat{y_d}$. Third, the Theil–Sen estimation method was performed to obtain the intercept, slope, $\hat{y}$, and ϵ in Equation (6). Lastly, we repeated the same simulation procedure for a total number of 1000 samples (i.e., 1000 simulated students).

We evaluated the performance of the IRS under different simulation conditions based on the same three criteria: (1) the average number of test administrations, (2) the average positive score change between test windows, and (3) the number of recommended test administrations.

Table 1. A summary of the simulation parameters.

Slope Type		Sparse Data		Full Data		Sample Size
		Slope Range	Number of Tests *	Slope Range	Number of Tests	
Negative Slope	Linear	(−1.1, 0)	5.86	(−0.6, −0.3)	20	1000
	Quadratic	(−1.1, 0)	5.80	(−0.6, −0.4)	20	1000
Minimal Change	Zero slope	(−0.01, 0.0)	5.09	(−0.01, 0.0)	20	1000
Plateau	After 14	(−0.1, 0.1)	9.32	(−0.1, 0.1)	20	1000
	After 15	(−0.1, 0.1)	7.58	(−0.1, 0.1)	20	1000
	After 17	(−0.1, 0.1)	7.07	(−0.1, 0.1)	20	1000

* This refers to the average number of test administrations.

4. Results

4.1. Results of the Real-Data Study

Table 2 shows a comparison between the performance of the IRS and standard practice. As explained earlier, standard practice refers to the assessment practices followed by the teachers who were able to select the number, timing, and frequency of test administrations for their students. The results indicate that the IRS recommended significantly fewer test administrations while maximizing the positive score change between test administrations for both grade levels. Furthermore, the IRS was able to reduce the number of tests administered to as few as 5 tests per student in grade 2, and 6 tests per student in grade 4, compared with a maximum of 17 tests per student in standard practice. Previous research utilizing Star Reading suggests that at least five tests should be administered to students for accurate decision making [8]. However, the results of our study show that the IRS could produce personalized test administration schedules with fewer tests by maximizing the score change between test administrations.

Table 2. Results of the intelligent recommender system (IRS) compared to standard practice (SP).

Evaluation Criteria	Grade 2		Grade 4	
	SP	IRS	SP	IRS
Average number of tests	5.42	3.51	5.37	3.84
Average positive score change	8.32	12.25	3.49	4.63
Minimum number of test administrations *	1	1	1	1
Maximum number of test administrations *	17	5	17	6

* It excludes the first two test administrations necessary for the slope estimation.

Figure 6 shows a grade two student with a positive growth trajectory in reading (top panel) and a grade four student who seems to be struggling to make adequate growth in reading (bottom panel). The grade 2 student participated in Star Reading 18 times and the grade 4 student (bottom panel) participated in Star Reading 17 times over the course of the academic year. The dashed line in each panel shows the students' growth trajectories based on their scores in Star Reading. The points with green shadowing represent the test administrations recommended by the IRS. Based on the recommended tests from the IRS, the total number of test administrations would reduce to three for both students. Unlike standard practice (i.e., test administration decisions made by teachers), the IRS monitored the students' progress and selected testing windows where they were most likely to demonstrate a significant score change. The first testing window was the starting point for both students, followed by a mid-year assessment in the 12th testing window, and an end-of-year assessment around the 18th or 19th testing window. Moreover, with the

tests recommended by the IRS, the positive score change per test would increase from 7.47 to 43.33 for the grade 2 student with a positive growth trajectory and from −0.29 to −1.67 for the grade 4 student with a negligible growth trajectory.

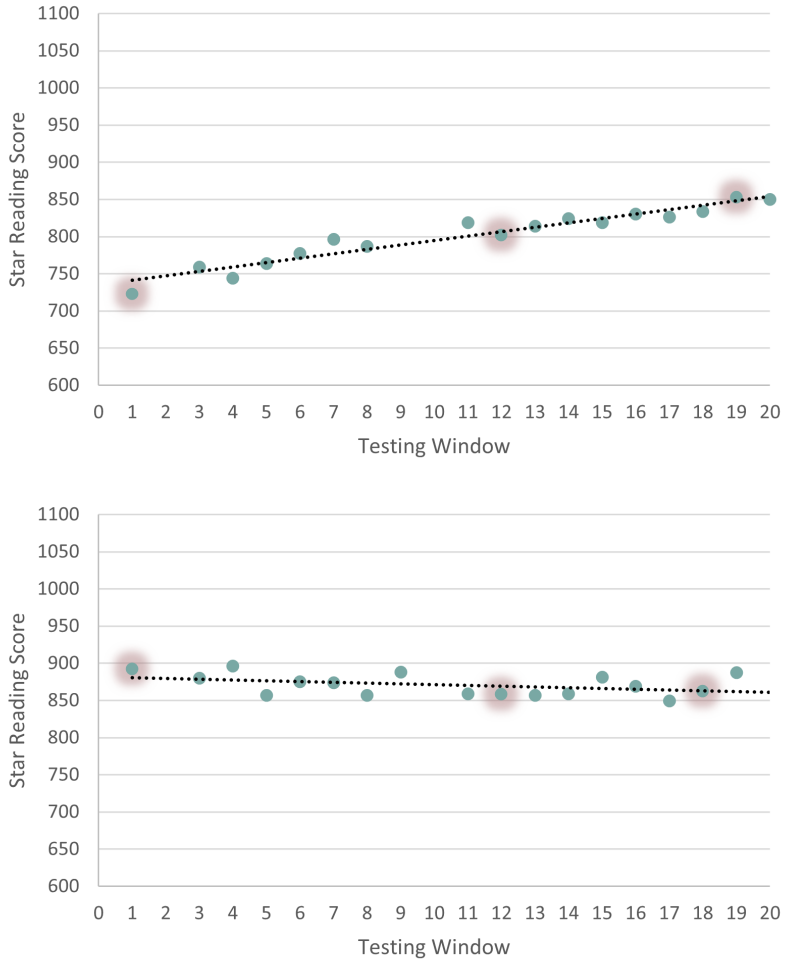

Figure 6. Original and recommended test administrations for two students.

Although the IRS produced testing schedule recommendations for the vast majority of the students, it was not able to generate a test administration schedule for less than 0.1% of the students at each grade level. To better understand the reasons for the inability of the IRS to generate a recommendation, we looked at these cases more carefully. At both grade levels, there appeared to be three reasons that no recommendation was made. First, students who only had a single test administration did not receive a recommendation. This is due to the fact that the system requires two data points before it will begin to generate a recommendation. This accounted for 0.02% and 0.01% of the cases in grade two and grade four, respectively. Second, when the number of test administrations increased and the observed score change did not increase significantly, no recommendation was produced by the IRS. In this case, the student should likely be tested more often because their growth is atypical. This group accounted for 0.02% of both the grade two and grade four cases. Finally, the third reason is that for some students, the number of test administrations decreased, but the observed score change did not change significantly. This group of cases represents an optimization of the testing schedule without greatly affecting the observed score change. For example, a student who originally participated in 4 tests with a relatively strong observed score change of 70 points was recommended to only take 3 tests, but this modification led the observed score change between the remaining data points to be 65. Although this difference could be considered a decrease in score, it is not necessarily large

enough to justify a change in the testing schedule. Thus, standard practice might be more suitable for such non-recommendable cases.

4.2. Results of the Simulation Study

Tables 3 and 4 present the results of the Monte Carlo simulation study for the sparse- and full-data conditions, respectively. Three patterns deserve to be scrutinized to better understand the results of the simulation study. First, we found that the IRS could reduce the number of test administrations significantly. This pattern remained consistent across all types of datasets simulated based on our three growth patterns (i.e., negative slope, minimal change, and plateau). In the sparse-data condition, the largest adjustment in the number of test administrations was observed for the plateau condition (i.e., $9.32 - 3.26 = 6.06$ difference). This finding suggests that, on average, the IRS identified six out of nine test administrations as redundant and, thus, did not recommend these test administrations. By contrast, the smallest adjustment in the number of test administrations occurred for the minimal change condition (i.e., $5.09 - 3.90 = 1.19$ difference). This finding indicates that, on average, the IRS identified only one test administration as redundant and, thus, did not recommend it. This was not a surprising finding because the simulated dataset for the sparse-data condition included the least number of test administrations under the minimal change condition.

Second, we found that the test administration schedules recommended by the IRS yielded a significant increase in the observed score change (see "Test Score Change" in Tables 3 and 4). The most noticeable changes were observed for the negative, linear and negative, quadratic slope conditions (i.e, roughly 19.12 difference in the sparse data and 60.67 difference in the full data). This finding suggests that the IRS could maximize the score change between different test administrations by selecting the testing windows where students were most likely to demonstrate a significant score change (either positive or negative). This is particularly important for the negative slope condition because following the test administration schedules recommended by the IRS would enable teachers to spend more instructional time for at-risk students who experience learning loss, instead of testing these students repeatedly.

Third, the superiority of the IRS in terms of the number of test administrations and observed score changes became more apparent in the full-data condition. In other words, generating a personalized test administration schedule would be most beneficial for classrooms where teachers are likely to follow a frequent testing approach to gauge academic growth for at-risk students. Our results showed that if significant score changes were not expected to be observed between subsequent test administrations, the IRS did not recommend new test administrations for at-risk students. This would help teachers focus on implementing differentiated instruction (or an academic intervention) that at-risk students often require to be able to show adequate growth in reading and other core subject areas.

Table 3. Results of the simulation study for the sparse-data condition.

Slope Type		Simulated Data		IRS	
		Number of Tests *	**Test Score Change**	**Number of Tests**	**Test Score Change**
Negative Slope	Linear	5.86	−44.66	3.75	−55.41
	Quadratic	5.80	−52.87	3.50	−71.99
Minimal Change	Zero slope	5.09	−3.22	3.90	−4.14
Plateau	After 14	9.32	−0.65	3.26	−1.92
	After 15	7.58	−0.54	3.23	−1.29
	After 17	7.07	−0.68	3.20	−1.38

* This refers to the average number of test administrations.

Table 4. Results of the simulation study for the full-data condition.

Slope Type		Simulated Data		IRS	
		Number of Tests	Test Score Change	Number of Tests	Test Score Change
Negative Slope	Linear	18	−14.59	4.65	−55.41
	Quadratic	18	−17.57	3.55	−78.24
Minimal Change	Zero slope	18	−1.41	4.30	−4.68
Plateau	After 14	18	−0.38	5.01	−1.36
	After 15	18	−0.28	4.98	−1.01
	After 17	18	−0.27	4.87	−0.97

5. Discussion

To date, many studies have discussed the importance of data-driven decision making to improve student learning and the quality of instruction in schools [31–34]. School-based professionals, such as classroom teachers and school psychologists, often collect qualitative and quantitative data on student learning through observations and classroom assessments. Then, they use the information to make decisions (e.g., applying academic interventions and tailoring the instruction to address students' learning needs). However, early applications of data-driven decision making were mostly limited to making one-size-fits-all decisions (e.g., making instructional adjustments based on grade-level guidelines produced by researchers) rather than personalized, student-level decisions. The evolution of big data and advanced data analytics approaches over the last two decades has resulted in the introduction of new educational frameworks such as LA [35] and educational data mining (EDM; [36]), which emphasize the use of data to provide personalized learning opportunities to students. Ongoing research in EDM and LA indicate that harnessing the power of data analytics helps educators improve the effectiveness of educational decision making in various areas such as personalizing teaching and learning, generating feedback for students, and individualized goal setting for students [37–39].

With a gradual shift towards digital learning in K–12 education, the increasing volume and variety of educational data have also posed practical challenges for school-based professionals. One of these challenges has been developing guidelines for teachers to help them determine an optimal administration schedule for formative assessments used for monitoring students' academic growth in core subject areas such as reading and mathematics. An optimal test administration schedule requires a balance between having enough data to make good decisions and minimizing the effects of sacrificing instructional time to have students complete a formative assessment. Previous research suggests that teachers often follow a one-size-fits-all testing procedure (e.g., monthly testing for all students) that overlooks individual differences in student progress [10]. This is not necessarily surprising because finding an optimal assessment schedule for each student requires considerable time and effort, as well as knowledge of many technical concepts such as measurement error in assessments. Moreover, teachers should arguably prioritize developing and delivering high-quality instruction, instead of dealing with scheduling formative assessments. Therefore, it is important to provide teachers with the most simplified approach to identifying an optimized testing schedule based on each student's progress. This would minimize the time and expertise required to make these decisions.

In this study, we proposed an IRS approach to generate personalized assessment schedules for students who participate in computerized formative assessments focusing on reading skills. The IRS utilizes Dijkstra's SPF algorithm [15] to find an optimal testing path for each student by minimizing the number of test administrations and maximizing the observed score change between subsequent test administrations. The goal of the algorithm is to calculate the distance (i.e., observed score change between test administrations) and recommend the shortest path (i.e., the schedule that yields the maximum score change) as a solution. We trained the IRS using a large sample of exemplary students in grades two

and four who participated in Star Reading and demonstrated adequate growth in reading. Then, we tested the performance of the IRS using a sample of at-risk students who either experienced learning loss or failed to demonstrate adequate academic growth in reading during the 2017–2018 academic year. In addition to the real data study, we conducted a Monte Carlo simulation study to evaluate the performance of the IRS in recommending optimal test administrations for students with unusual growth trajectories (e.g., negative growth, no growth, and plateau).

The results from the real-data study showed that there was a drastic difference between the number of assessments completed during standard practices (i.e., test administration schedules followed by teachers) and the optimized number of tests recommended by the IRS. Based on the data used for this study, it appears that there are a number of additional tests being administered to students (i.e., over-testing), which takes away from instructional time and increases the likelihood of students experiencing testing fatigue. Despite the potentially negative consequences of this practice, researchers have generally continued to advocate for a relatively large number of tests be administered over a long period [13,40]. The fact that the IRS only requires two data points to make a strong assessment scheduling recommendation radically changes the way computerized formative assessments are conducted in schools. With the use of the IRS, the assessment scheduling process has become entirely automated, which frees up precious teacher resources so they can focus on other tasks—such as delivering high-quality instruction. Consequently, the use of an IRS represents a significant paradigm shift in educational assessment. The results from the simulation study also provided additional evidence regarding the robustness of the IRS in the presence of unusual growth trajectories. Regarding the number of test administrations and observed score changes, the advantage of the IRS was most apparent for the negative slope condition (i.e., students experiencing learning loss) and the full-data condition (i.e., students being over-tested). These findings suggested that the IRS served its purpose by optimizing the assessment schedules, especially for students who need increased instructional time instead of being tested needlessly.

5.1. Practical Implications

The results of our study suggest that it is possible to generate personalized test schedules for computerized formative assessments using an IRS. This innovative approach represents a significant step forward in data-based decision-making based on computerized formative assessments within an LA ecosystem. Namely, it comes with many benefits and it avoids many of the challenges associated with standard practice followed by teachers in the classroom. One of the primary benefits is that the IRS has the ability to produce an individualized pathway to successful educational growth for students. This positive framing of educational goals is likely to increase student motivation for growth in core curricular areas such as reading, regardless of their current level of performance. Furthermore, the IRS reduces the number of assessments required for strong decision making, which leads to less instructional time being sacrificed for administering computerized formative assessments. The added benefit is that it should also reduce the likelihood of student burn-out with respect to assessments. Finally, the IRS produces an efficient and relatively care-free test administration process for the teachers by giving them a two-week testing window for every single student in their classroom. This is not only a flexible approach to monitoring students' progress regularly, but it also reduces the need for the teachers to try to optimize their students' assessment schedules themselves. Lastly, the IRS approach presented in this study emphasizes the value of LA for K–12 education. K–12 schools and school authorities can use our IRS approach as a starting point for developing further LA applications that generate actionable insights for teachers and other stakeholders in education [41].

5.2. Limitations and Future Directions

This study has several limitations. First, we built and tested the IRS using existing data for two grades (grades two and four) in reading. Although the data were split and

the model testing represents what would likely happen if the IRS were to be implemented, the process has yet to be validated using real, on-the-fly recommendations in schools. In other words, we developed a system that is intended to be dynamic using static data. Therefore, it will be important to monitor, and potentially tweak, the IRS if it is implemented for generating personalized assessment schedules in schools. Future studies can evaluate the feasibility of the proposed IRS approach using real data from different grade levels, as well as different subject areas such as mathematics and science.

Second, the IRS proposed in this study is limited to producing test administration recommendations but does not necessarily make decisions about student progress. This means that the decision to provide differentiated instruction, implement an academic intervention, or refer a student for a psycho-educational assessment based on the results of computerized formative assessments are still left to teachers and other school-based professionals (e.g., school psychologists). Nonetheless, the data produced by the IRS should allow these professionals to make these decisions with more confidence, given that the trend lines produced from the relatively fewer test administrations would likely be very close to the trend line produced with the previously large number of test administrations.

Third, although the IRS essentially optimizes the assessment schedule based on its best prediction of what a student's growth trajectory is likely to be, it is not necessarily designed to be predictive of end-of-year academic performance. The individual assessment schedules would continuously be updated as students take more tests (i.e., as more data are collected). This means that predictions about students' long-term educational outcomes from one or more data points should be considered in addition to using the IRS. Future studies can expand our IRS approach by incorporating additional predictive algorithms that would prioritize the prediction of end-of-year academic performance after a certain point within a school year (e.g., after the start of the second semester).

Author Contributions: Conceptualization, O.B., D.C.C. and J.S.; methodology, O.B. and J.S.; software, J.S.; validation, O.B., D.C.C. and J.S.; formal analysis, J.S.; data curation, O.B. and D.C.C.; writing—original draft preparation, O.B. and J.S.; writing—review and editing, O.B., D.C.C. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed in this study was obtained from Renaissance, Inc., and the following restrictions apply: all data are solely owned and licensed by Renaissance, Inc. and thus cannot be discussed or shared by the researchers in any form or format. Requests to access these datasets should be directed to Eric Stickney, eric.stickney@renaissance.com.

Conflicts of Interest: O.B. and D.C.C. were paid consultants for Renaissance, Inc. during the design and implementation of the project that resulted in this manuscript. J.S. declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EDM	Educational data mining
IRS	Intelligent recommender system
IRT	Item response theory
LA	Learning analytics
SPF	Shortest path first
TEL	Technology-enhanced learning

References

1. Black, P.; Harrison, C.; Lee, C.; Marshall, B.; Wiliam, D. Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan* **2004**, *86*, 8–21. [CrossRef]
2. McMillan, J.H.; Andrade, H.L.; Heritage, M. *Using Formative Assessment to Enhance Learning, Achievement, and Academic Self-Regulation*; Routledge: London, UK, 2017.
3. Barana, A.; Conte, A.; Fissore, C.; Marchisio, M.; Rabellino, S. Learning analytics to improve formative assessment strategies. *J. E-Learn. Knowl. Soc.* **2019**, *15*, 75–88. [CrossRef]
4. Siemens, G.; Gasevic, D. Guest editorial-learning and knowledge analytics. *J. Educ. Technol. Soc.* **2012**, *15*, 1–2.
5. Admiraal, W.; Vermeulen, J.; Bulterman-Bos, J. Teaching with learning analytics: How to connect computer-based assessment data with classroom instruction? *Technol. Pedagog. Educ.* **2020**, *29*, 577–591. [CrossRef]
6. Nouri, J.; Ebner, M.; Ifenthaler, D.; Sqr, M.; Malmberg, J.; Khalil, M.; Bruun, J.; Viberg, O.; González, M.Á.C.; Papamitsiou, Z.; et al. Efforts in Europe for data-driven improvement of education—A review of learning analytics research in six countries. *Int. J. Learn. Anal. Artif. Intell. Educ.* **2019**, *1* 8–27. [CrossRef]
7. Dawson, S.; Gašević, D.; Siemens, G.; Joksimovic, S. Current state and future trends: A citation network analysis of the learning analytics field. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, Indianapolis, IN, USA, 24–28 March 2014; pp. 231–240. [CrossRef]
8. Bulut, O.; Cormier, D.C. Validity evidence for progress monitoring with Star Reading: Slope estimates, administration frequency, and number of data points. *Front. Educ.* **2018**, *3*, 68. [CrossRef]
9. Maier, U.; Wolf, N.; Randler, C. Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Comput. Educ.* **2016**, *95*, 85–98. [CrossRef]
10. Bulut, O.; Cormier, D.C.; Shin, J. An intelligent recommender system for personalized test administration scheduling with computerized formative assessments. *Front. Educ.* **2020**, *5*, 182. [CrossRef]
11. Shin, J.; Bulut, O. Building an intelligent recommendation system for personalized test scheduling in computerized assessments: A reinforcement learning approach. *Behav. Res. Methods* **2022**, *54*, 216–232. [CrossRef]
12. January, S.A.A.; Van Norman, E.R.; Christ, T.J.; Ardoin, S.P.; Eckert, T.L.; White, M.J. Progress monitoring in reading: Comparison of weekly, bimonthly, and monthly assessments for students at risk for reading difficulties in grades 2–4. *Sch. Psychol. Rev.* **2018**, *47*, 83–94. [CrossRef]
13. January, S.A.A.; Van Norman, E.R.; Christ, T.J.; Ardoin, S.P.; Eckert, T.L.; White, M.J. Evaluation of schedule frequency and density when monitoring progress with curriculum-based measurement. *Sch. Psychol.* **2019**, *34*, 119. [CrossRef]
14. Ricci, F.; Rokach, L.; Shapira, B. (Eds.) Recommender Systems: Introduction and Challenges. In *Recommender Systems Handbook*; Springer: Boston, MA, USA, 2015; pp. 1–34. [CrossRef]
15. Dijkstra, E.W. A note on two problems in connexion with graphs. *Numer. Math.* **1959**, *1*, 269–271. [CrossRef]
16. Zaiane, O. Building a recommender agent for e-learning systems. In Proceedings of the International Conference on Computers in Education, Auckland, New Zealand, 3–6 December 2002; pp. 55–59. [CrossRef]
17. Manouselis, N.; Drachsler, H.; Verbert, K.; Santos, O.C. (Eds.) *Recommender Systems for Technology Enhanced Learning: Research Trends and Applications*; Springer Science & Business Media: New York, NY, USA, 2014.
18. Thai-Nghe, N.; Drumond, L.; Krohn-Grimberghe, A.; Schmidt-Thieme, L. Recommender system for predicting student performance. *Procedia Comput. Sci.* **2010**, *1*, 2811–2819. [CrossRef]
19. Leite, W.; Roy, S.; Chakraborty, N.; Michailidis, G.; Huggins-Manley, A.C.; D'Mello, S.; Shirani Faradonbeh, M.K.; Jensen, E.; Kuang, H.; Jing, Z. A novel video recommendation system for algebra: An effectiveness evaluation study. In Proceedings of the LAK22: 12th International Learning Analytics and Knowledge Conference, Newport Beach, CA, USA, 21–25 March 2022; pp. 294–303.
20. De Oliveira, M.G.; Marques Ciarelli, P.; Oliveira, E. Recommendation of programming activities by multi-label classification for a formative assessment of students. *Expert Syst. Appl.* **2013**, *40*, 6641–6651. [CrossRef]
21. Baylari, A.; Montazer, G.A. Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Syst. Appl.* **2009**, *36*, 8013–8021. [CrossRef]
22. Chen, C.M.; Lee, H.M.; Chen, Y.H. Personalized e-learning system using item response theory. *Comput. Educ.* **2005**, *44*, 237–255. [CrossRef]
23. De Schipper, E.; Feskens, R.; Keuning, J. Personalized and automated feedback in summative assessment using recommender systems. *Front. Educ.* **2021**, *6*, 652070. [CrossRef]
24. Kundu, S.S.; Sarkar, D.; Jana, P.; Kole, D.K. Personalization in Education Using Recommendation System: An Overview. In *Computational Intelligence in Digital Pedagogy*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 85–111.
25. Chen, M.; Chowdhury, R.A.; Ramachandran, V.; Roche, D.L.; Tong, L. *Priority Queues and Dijkstra's Algorithm*; Computer Science Department, University of Texas at Austin: Austin, TX, USA, 2007.
26. Renaissance. *Star Assessments™ for Reading Technical Manual*; Technical Report; Renaissance : Wisconsin Rapids, WI, USA, 2018.
27. Vannest, K.J.; Parker, R.I.; Davis, J.L.; Soares, D.A.; Smith, S.L. The Theil–Sen slope for high-stakes decisions from progress monitoring. *Behav. Disord.* **2012**, *37*, 271–280. [CrossRef]
28. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.

29. Komsta, L. Mblm: Median-Based Linear Models. R Package Version 0.12.1. 2019. Available online: https://cran.r-project.org/web/packages/mblm/ (accessed on 30 May 2022).
30. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
31. Hoogland, I.; Schildkamp, K.; Van der Kleij, F.; Heitink, M.; Kippers, W.; Veldkamp, B.; Dijkstra, A.M. Prerequisites for data-based decision making in the classroom: Research evidence and practical illustrations. *Teach. Teach. Educ.* **2016**, *60*, 377–386. [CrossRef]
32. Kaufman, T.E.; Graham, C.R.; Picciano, A.G.; Popham, J.A.; Wiley, D. Data-driven decision making in the K-12 classroom. In *Handbook of Research on Educational Communications and Technology*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 337–346. [CrossRef]
33. Lai, M.K.; Schildkamp, K. Data-based Decision Making: An Overview. In *Data-Based Decision Making in Education: Challenges and Opportunities*; Springer: Dordrecht, The Netherlands, 2013; pp. 9–21. [CrossRef]
34. Marsh, J.A.; Pane, J.F.; Hamilton, L.S. *Making Sense of Data-Driven Decision Making in Education: Evidence from Recent RAND Research*; RAND Corporation: Santa Monica, CA, USA, 2006. [CrossRef]
35. Ferguson, R. Learning analytics: Drivers, developments and challenges. *Int. J. Technol. Enhanc. Learn.* **2012**, *4*, 304–317. [CrossRef]
36. Baker, R.S.; Yacef, K. The state of educational data mining in 2009: A review and future visions. *J. Educ. Data Min.* **2009**, *1*, 3–17. [CrossRef]
37. Baepler, P.; Murdoch, C.J. Academic analytics and data mining in higher education. *Int. J. Scholarsh. Teach. Learn.* **2010**, *4*, 1–9. [CrossRef]
38. Vatrapu, R.; Teplovs, C.; Fujita, N.; Bull, S. Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge, LAK '11, Banff, AB, Canada, 27 February–1 March 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 93–98.
39. Tempelaar, D.T.; Rienties, B.; Giesbers, B. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Comput. Hum. Behav.* **2015**, *47*, 157–167. [CrossRef]
40. Christ, T.J.; Zopluoglu, C.; Long, J.D.; Monaghen, B.D. Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Except. Child.* **2012**, *78*, 356–373. [CrossRef]
41. Baker, R.S.; Koedinger, K.R. Towards demonstrating the value of learning analytics for K-12 education. In *Learning Analytics in Education*; Niemi, D., Pea, R.D., Saxberg, B., Clark, R.E., Eds.; Information Age Publishing: Charlotte, NC, USA, 2018; Chapter 2, pp. 49–62.