The effects of a personalized recommendation system on students' high-stakes achievement scores: A field experiment

Nilanjana Chakraborty
Department of Statistics
University of Florida
nchakraborty@ufl.edu

Samrat Roy
Department of Statistics
University of Florida
samratroy@ufl.edu

Walter L. Leite
College of Education
University of Florida
walter.leite@coe.ufl.edu

Mohamad Kazem Shirani Faradonbeh
Department of Statistics
University of Georgia
mohamadksf@uga.edu

George Michailidis
Department of Statistics and the Informatics Institute
University of Florida
gmichail@.ufl.edu

ABSTRACT

This study examines data from a field experiment investigating the effects of a personalized recommendation algorithm that proposes to students which videosto watch next, after they complete mini-assessments for algebra that available on the Math Nation intelligent virtual learning environment (IVLE). The end users of Math Nation are students enrolled in an Algebra 1 course in middle and high schools of the state of Florida, and the IVLE is used both during and out of school time. The objective of the developed recommendation algorithm is to increase student preparation to take the state-mandated End-of-Course (EoC) Algebra 1 assessment at the end of the school year. The algorithm is based on a Markov Decision Process framework that uses as input the students' responses to a series of mini-assessment tests. The current study randomly assigned 16,406 students to either treatment or control conditions, which were blind to both students and teachers. The results indicate that the effects of the recommendation algorithm depend on the level of usage of students, showing significant improvements on EoC test scores of students who have a moderate level of usage. However, there was no effect for low usage students. The study also shows that students practicing with the mini-assessments available on Math Nation, helps them improve by a small margin their performance on the End-of-Course test, irrespective of the usage level. Finally, the study provides insights on challenges posed for implementing personalized recommendation algorithms at a large scale, related both to student self-regulation and teacher orchestration of technology use in the classroom.

Nilanjana Chakraborty, Samrat Roy, Walter Leite and George Michailidis "The effects of a personalized recommendation system on students' high-stakes achievement scores: A field experiment". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 588-594. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education

Keywords

Personalized Recommendation, Randomized Control Study, Hierarchical Clustering, Markov Decision Process, Algebra

1. INTRODUCTION

There is a growing trend in employing intelligent virtual learning environments (IVLE) to aid students in improving their math performance in K-12 education [8, 26, 23]. While there is a robust body of literature that shows that students' preparedness together with various demographic and school characteristics are key factors for predicting students' performance in various math tests [15], IVLE have been viewed as an especially promising way of improving students' achievements in mathematics. Given the investment of resources into technology products and the time and effort needed to integrate them into the curriculum, there has been considerable interest in determining their effectiveness. A number of studies have reported positive effects based both on small scale randomized control trials and longer term interventions [14, 13, 19, 16, 15], as well as based on observational data [12]. There have also been a series of meta-analysis studies showing that IVLE have substantial effects on student outcomes [10, 11, 24, 27].

IVLE have the potential of offering personalized learning experiences. The latter refer to instruction "in which the pace of learning the instructional approach are optimized for the needs of each learner", according to the United States National Education Technology Plan 2017. IVLE that offer some degree of personalization include Khan Academy at the K-12 level and Newton at the higher education level. As discussed in [3], at the core of personalized learning strategies is a recommendation algorithm aiming to propose appropriate learning materials and topics to the student at the right time, leveraging the student's prior history of interactions with the IVLE.

Many personalized learning strategies leverage ideas and tools from the field of Reinforcement Learning [4, 5, 9, 20]. The key components of a reinforcement learning based algorithm are the triplet of *state, reward, action*. The state reflects information on the student's knowledge and skills set on the topic(s) under consideration, the reward relates to the goals of the strategy (e.g. performance on tests, engagement with the IVLE, etc.) and the action refers to an activity, (e.g. watch a video on a topic of interest, take an assessment test, etc.) that, based on the current state information, aims to maximize the expected reward.

This study reports the results of a large-scale randomized field experiment that focuses on the impact of a simple personalized strategy implemented on the Math Nation IVLE, on a high-stakes, state-mandated End-of-Course (EoC) algebra test. Although many evaluations of IVLE have been published, most of them rely on locally developed standardized tests, rather than high-stakes statewide tests [10]. Math Nation, is an online video-based tutoring program aiming to prepare students in the state of Florida for the EoC, which is required for high school graduation. The platform offers videos on various algebra topics recorded by different tutors, explaining the main concepts and walking the student through related examples, 3-question assessments for each topic and 10-question assessments for sets of related topics, with video explanations for each question. Therefore, students can assess their progress by taking both the short (3-question) and the long (10-question) tests. Further, the platform offers a monitored discussion area, wherein students can pose questions to peers and volunteer tutors. Hence, at launch time, it shared a number of characteristics with Khan Academy, both being self-guided and easy to use on an ad hoc basis, without the need for extensive professional development training for teachers. The content of the videos and assessments are aligned with the curriculum adopted by the state and also the content and format of the EoC test.

A new feature of Math Nation is the introduction of an algorithm to recommend videos to students, leveraging information on their performance on the mini-assessments associated with each video. Specifically, Math Nation divides the whole Algebra 1 course materials into 10 sections. Each section is further divided into several topics, thus resulting in a total of 93 topics for the entire course. For each topic, there is a tutorial video associated with it, recorded by different tutors. At the end of the video the student is presented with a 3-question assessment (henceforth called a mini-assessment) and based on the score obtained, a video recommendation (the action) is offered aiming to maximize the student's expected score (the reward) on these mini-assessments. The student can follow the recommendation or decide to ignore it and select another video of her/his own choice by the same or another tutor. To compare the effectiveness of the recommendation algorithm, a "business-as-usual" competitor is implemented, which recommends the next video in a predetermined sequence related to the structure of the algebra state curriculum, irrespective of the score achieved in the mini-assessment.

The objectives of the study are twofold: (i) estimate the average treatment effect of the recommendation algorithm vis-a-vis its competitor together with its interactions with previous achievement and level of usage of the algorithm, and (ii) understand the relationship between performance in the mini-assessments and the EoC test, after accounting for math preparedness and school characteristics of the students that participated in this randomized control study.

The remainder of the paper is structured as follows. Section 2 presents the developed personalized recommendation strategy. Section 3 describes in detail the data recorded from the algorithm, as well as other covariates used in the analysis. Section 4 presents the statistical methods used in the analysis and the main results of the study. Finally, Section 5 discusses the implications of our findings and suggestions to modify the recommendation algorithm.

2. PERSONALIZED RECOMMENDATION STRATEGY

Next, we describe the data-driven algorithm for recommending a suitable tutoring video to each individual student. As previously mentioned, the content of the course is divided into 93 topics, with each topic accompanied by a video recorded by 5 tutors in English and 1 tutor in Spanish. Students can freely select the tutor for each video.

To rigorously set the stage for the video recommendation algorithm, fix a single student, and let $s_k(t)$ be the corresponding "mini-score" for topic $k \in \{1, 2, \cdots, 93\}$, at time $t = 0, 1, \cdots$. These mini-scores, representing the knowledge level of the student, are obtained by assessing responses to the mini-assessments comprising of 4-choice questions, with a single correct choice. Thus, the set of possible outcomes consists of i correct answer(s), together with $3-i$ wrong answer(s), for $i = 0, 1, 2, 3$. Then, we center and normalize the corresponding scores (henceforth referred to as mini-scores), so that on average, simply guessing the answers lead to a zero score. Thus, we have $s_k(t) \in \{-3, 1, 5, 9\}$, and if the answers are selected completely at random, $\mathbb{E}[s_k(t)] = 0$.

With the above setting, the full state of the student at time t is given by $S(t) = [s_1(t), \cdots, s_{93}(t)]' \in \{-3, 1, 5, 9\}^{93}$, while $\|S(t)\| = \sum_{k=1}^{93} s_k(t)$ reflects the (total) score of the student under consideration at time t. The dynamical model for topic k consists of a Markov chain for which the state is $s_k(t)$. For the time being, suppose that the parameters of the Markov chain consisting of 4×4 tables of transition probabilities among the states $\{-3, 1, 5, 9\}$ are available. We will shortly discuss a statistical method leveraging transfer learning techniques, for estimating the Markov transition kernels according to the observed data.

The recommendation strategy is to propose to the student the tutoring video corresponding to the topic with the *largest predicted growth* in the mini-score. Formally, at time t, the IVLE recommends the student to watch the tutoring video of topic $k^\star$, wherein

$$k^\star = \arg\max_k \mathbb{E}\left[s_k(t+1) - s_k(t) \,\middle|\, S(t)\right],$$

where the notation " | " is used to indicate a conditional probability distribution. The student can either accept the recommendation, watch the video and take the mini-assessment, or can ignore the recommendation and select another video

to watch (by possibly another tutor).

Note that in order to compute the above expected values, for every topic $i \in \{1, \cdots, 93\}$ it suffices to have only the 4 probabilities corresponding to the transition of the Markov chain from the current state $s_i(t)$ to the next one $s_i(t+1)$. Intuitively, the difference quantity $s_k(t+1) - s_k(t)$ reflects the predicted growth of the student in topic k. Therefore, the high level idea of the recommendation strategy is to propose to the student to work on the topic (s)he is capable of improving her/his knowledge level the most. Therefore, the recommendation aligns with Vygotsky's theory [28] of zone of proximal development by providing a video that is neither too easy, nor too challenging. Further, the recommended topic is totally personalized to the student, since the state $S(t)$ at time t is unique to each student.

Finally, we describe the statistical learning procedure for estimating the Markov transition probabilities. For this purpose, the students are clustered in 12 different groups, based on their demographic and other background data, so that students of similar learning abilities will be assigned to the same group (cluster). The details of the clustering procedure are provided in Section 3. We assume that students in each group share the Markov transition probabilities reflecting their cognitive responses to watching the tutoring video of a specific topic. Thus, in order to estimate the transition probabilities for students in a fixed group, we divide the total number of transitions between every pair of the possible states $\{-3, 1, 5, 9\}$ in the group, with the total number of transitions in the group. We emphasize the following points. First, while the Markov transition probabilities are the same for all students in one demographic/background group, the states are uniquely personalized to each student. Second, the estimates of the transition probabilities change over time as the platform collects more data from the responses of the students to the mini-assessments. Further, when Math Nation starts being used by the students, the initial estimates of the transition probabilities are selected randomly, and are updated throughout the academic year as the students continue to use it. Finally, if there is more than one $k^\star$ maximizing the predicted growth, one will be selected at random.

Before the algorithm was deployed within Math Nation platform, it was extensively tested on synthetic data generated based on data collected in previous years from the platform. Specifically, students that have used the platform in previous years were clustered in 12 groups (see also Section 3) based on their demographic and background information. Note that the distributions of such data are very similar to those in the academic year that the recommendation algorithm was launched and evaluated in the current study. Subsequently, the response data to the mini-assessment tests of the students within each cluster were used to estimate the corresponding Markov transition probabilities. The latter were then used to initialize the recommendation algorithm and to generate synthetic data for students in different clusters. The upshot of this analysis was that the algorithm required adequate engagement ($t \geq 45$) to show significant improvement in performance in the mini-assessments. We revisit this point in the Discussion section.

Table 1: Distribution of the students across different Math Achievement Levels, School Grades and Student Grades

Achievement Level	No. of Students	School Grades	No. of Students	Student Grades	No. of Students
1	473	A	4,377	5	3
2	1,453	B	2,001	6	1463
3	3,487	C	4580	7	3599
4	2,711			8	5893
5	2,834				
Total	10,958	Total	10,958	Total	10,958

Note: Data based on previous school year performance

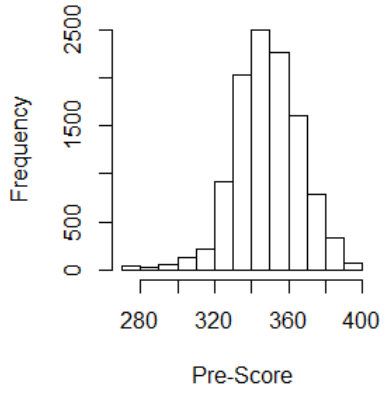

Figure 1: Distribution of Pre-Score

3. DATA DESCRIPTION

In this study, we randomly assign a sample of 16,406 middle and high school students enrolled in Algebra 1 in a large school district in the state of Florida, to a treatment (proposed recommendation strategy) or a control (business as usual recommendation strategy) group. The assignment was blind to students and teachers. The treatment group received video recommendations as described in the previous section, while the control group received a recommendation to watch the next video in the curriculum sequence. To initialize the recommendation, a randomized cluster design was employed. Specifically, students were first matched according to their grade, school characteristics and math preparedness test scores from the previous school year and then randomly assigned to the two groups. The variables used for matching purposes were the scores on the state standardized mathematics test, called the Mathematics Florida Standards Assessment[1](henceforth, referred to as Pre-Score and the corresponding test referred to as Pre-Test), as well as an achievement level assigned to them by their schools, while the quality of each school is reflected by a grade assigned to it by the state Department of Education[2]. The latter grades are based on several components and have five different levels ('A' being the highest level and 'F' being the lowest one). Due to lack of data for many of these variables, 5,448 students were removed from any further analysis and hence Table 1 that shows the distributions of the students across different Achievement Levels, School Grades and Student Grades and Figure 1 that depicts the distribution of the Pre-Score are based on the remaining 10,958 students.

[1]http://www.fldoe.org/accountability/assessments/k-12-student-assessment/fsa.stml

[2]http://www.fldoe.org/accountability/accountability-reporting/school-grades/

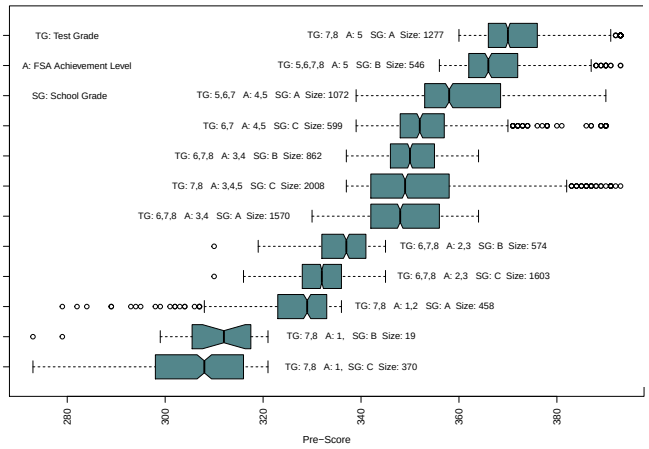

Figure 2: Boxplots of clusters hierarchically ordered based on Pre-score: For each cluster, School Grade, Achievement Level and Student Grade and Cluster Size are reported.

Using the above information, students were assigned to clusters/groups. This cluster assignment is used as a categorical variable in the analysis presented in Section 4. The clusters are designed in such a way that each of them corresponds to a group with a unique combination of math preparedness and school grade. In summary, the following four variables were considered by the clustering algorithm: Pre-Score, Math Achievement Level, School Grade and Student Grade. An agglomerative hierarchical clustering algorithm was employed for this task and using the dendrogram with Gower's distance metric, along with silhouette values [7], the number of clusters was chosen to be 12. Figure 2 provides a pictorial representation of the key features of the clusters. Specifically, for each cluster the Figure depicts the boxplot of the Pre-Score and also the corresponding Student Grade, Math Achievement Level, School Grade and Size of the cluster. For ease of comparison, the clusters are ordered according to the distribution of the Pre-Score. Hence, cluster 1 corresponds to the group of students having the lowest Pre-Score, while cluster 12 is the group with the highest Pre-Score. As Figure 2 shows, the size of cluster 2 was very small and hence it was merged with cluster 1 for the subsequent analyses.

The number of times a particular student takes the mini-assessment after watching a video, is defined as the usage by that student. Figure 3 depicts the average usage per student for each of the clusters for both the control and treatment groups. It can readily be seen that the overall average across the study population is 2.88, with many clusters exhibiting significantly lower usage. There are also a few clusters exhibiting high usage; e.g. cluster 5 for the control group and cluster 9 for the treatment group.

4. METHODS AND RESULTS

The analyses described below, aim to provide answers to the two objectives outlined in Section 1. In our first analysis, we estimate the average treatment effect of the recommendation algorithm on EoC scores, using a simple linear regression model, with the following two categorical variables and the interaction between them; (i) the first categorical variable TC, comprises of two levels: the first represents the Treatment group that watched the personalized recommended videos and took the corresponding mini-assessments, and the second level corresponds to the Control group; (ii) the second categorical variable Previous Achievement Level, comprises of five categories, each corresponding to a different level of achievement in the Pre-test. Level 1 stands for the lowest achievement, whereas the highest level is coded by level 5. Then, the linear regression model with the above two predictors and their interaction is given by:

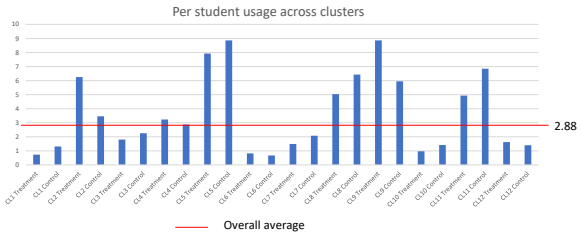

Figure 3: Average usage per student for different clusters, for both treatment and control groups

$$y = \mu + \beta_1\ TC + \beta_2\ Achievement + \beta_3\ (TC \times Achievement) + \epsilon \quad (1)$$

where y represents the EoC score and we further assume that $\epsilon \sim N(0, \sigma^2)$. Based on this model, the estimate of the average treatment effect of the personalized recommendation on EoC score, is the coefficient β_1 corresponding to the variable TC. Further, estimates of standard errors of the regression coefficients are based on cluster-robust estimators [2]. To answer the first research question discussed in Section 1, we test $H_0^{TC} : \beta_1 = 0$ vs. $H_1^{TC} : \beta_1 \neq 0$. The coefficient β_1 is the difference between the mean EoC score of the Treatment and the Control group, after accounting for the effect of all the other covariates. The estimated coefficients (scaled) and corresponding p-values are reported in Table 2. Table 2 shows that the achievement levels are statistically significant, while the treatment effect (i.e., the impact of the developed recommendation algorithm) is not. Further, there is a small positive significant effect for the interaction of the treatment with Achievement level 2. However, as shown in Figure 3, usage patterns vary widely across different groups (clusters) of students.

To that end, and in order to gain a deeper understanding of how the average treatment effect behaves across different IVLE usage levels, we fit model (1) separately on groups of students exhibiting different usage levels. After some initial exploratory analysis, we divided the students in approximately evenly distributed usage groups as shown in Table 3. The results are summarized in Table 3, whose first column specifies the usage levels of the group. As an example, students who have taken at least 10 mini-assessments tests, are categorized as a group with usage level 10 or higher.

Table 2: Estimated coefficients and corresponding p-values for Model (1)

Variable	Scaled Coefficient	p-value
Intercept	348.45	<0.001
Treatment	-0.99	0.32
Achievement level 2	11.28	<0.001
Achievement level 3	23.14	<0.001
Achievement level 4	35.31	<0.001
Achievement level 5	50.02	<0.001
TC*Achievement level 2	1.94	0.05
TC*Achievement level 3	0.94	0.34
TC*Achievement level 4	1.25	0.21
TC*Achievement level 5	1.09	0.27

Note: The scaled coefficients are obtained by dividing the estimated coefficients by their standard error.

The second and third columns contains= the p-values corresponding to the test $\beta_1 = 0$ and the scaled version of the estimated coefficients, respectively. As it is evident from Table 3, the first few rows that correspond to lower usage groups, have high p-values and thus the average treatment effect is not statistically significant. The treatment effect becomes significant for students who used the platform more extensively (≥ 48).

The model also controls for the level of achievement of students. Table 4 presents the results for the β_2 regression coefficient for different usage levels. The corresponding p-values are given in parentheses. It can be seen that the effect is statistically significant (marked in bold font) across almost all Previous Achievement levels and usage levels, as expected based on the overall results presented in Table 2. Further, this result is in accordance with a large body of literature that has found a positive association between level of math preparation and test scores (see, e.g.,[16, 15] and references therein). Further, the magnitude of the coefficient is larger for higher achievement levels.

Model (1) also estimates the interaction effect between the treatment and the Previous Achievement level. Table 5 summarizes the scaled estimates of the interaction effects and the p-values (given in parentheses). Since the Previous Achievement level has 5 categories, we obtain the estimates for all the levels except the baseline category, i.e., Previous Achievement level 1, which is absorbed in the intercept of the model. As usage increases, Table 5 displays more significant interaction effects (in bold font) between treatment and achievement level as compared to low usage groups. Note that due to lack of data in selected categories, some of the interaction effects could not be estimated and hence left blank.

Note that most of the interaction effects are not statistically significant. There are selected ones with a positive coefficient, corresponding to higher achievement levels (3 and above) for high usage groups (e.g., 33 and 65). Analogously, there are selected interaction effects with a negative coefficient corresponding to the lower achievement level 2, and relative high usage level.

To answer the second research question on the relationship between the performance of the students in mini-assessments and in the EoC test, we obtain the *Average Mini-Assessments*

Table 3: Usage-wise effect of the recommendation: p-values and scaled coefficients for different usage levels

Usage Level	p-value	Scaled Coefficient	Sample size
9	0.64	0.46	1097
13	0.40	0.85	932
27	0.29	1.07	515
33	0.17	1.39	411
48	0.02	2.41	254
52	0.01	2.56	230
55	0.05	1.93	207
59	0.06	1.91	183
65	0.02	2.31	140
74	0.08	1.78	92

Table 4: Usage-wise effect of the Previous Achievement level: p-values and scaled coefficients for different usage levels

Usage Level	Level 2	Level 3	Level 4	Level 5
9	4.65 (**<0.001**)	7.41 (**<0.001**)	10.67 (**<0.001**)	15.76 (**<0.001**)
13	4.94 (**<0.001**)	7.51 (**<0.001**)	- 10.74 (**<0.001**)	15.37 (**<0.001**)
27	1.87 (0.06)	2.64 (**0.008**)	4.61 (**<0.001**)	6.76 (**<0.001**)
33	2.45 (**0.01**)	3.02 (**0.002**)	4.72 (**<0.001**)	6.19 (**<0.001**)
48	3.23 (**0.001**)	3.19 (**0.001**)	4.18 (**<0.001**)	5.41 (**<0.001**)
52	3.16 (**0.002**)	3.29 ((**0.001**))	4.16 (**<0.001**)	5.42 (**<0.001**)
55	2.97 (**0.003**)	3.07 (**0.002**)	3.97 ((**<0.001**))	5.41 (**<0.001**)
59	2.05 (**0.04**)	1.92 (**0.05**)	2.65 (**0.008**)	3.58 (**<0.001**)
65	-	-0.13 (0.89)	1.83 (0.06)	4.15 (**<0.001**)
74	-	0.50 (0.62)	1.34 (0.18)	2.70 (**0.008**)

Table 5: Usage-wise interaction effect of treatment and Previous Achievement level: scaled coefficients (p-values) for different usage levels

Usage Level	TC * Level 2	TC * Level 3	TC * Level 4	TC * Level 5
9	-0.47 (0.64)	-0.20 (0.84)	-0.40 (0.69)	-0.06 (0.94)
13	-1.08 (0.27)	-0.68 (0.49)	- 0.81 (0.42)	-0.34 (0.74)
27	0.62 (0.54)	1.19 (0.23)	1.05 (0.29)	1.60 (0.11)
33	0.78 (0.43)	1.47 (0.14)	1.32 (0.19)	2.13 (0.03)
48	-2.82 (**0.005**)	-1.21 (0.22)	-2.05 (**0.04**)	-
52	-2.85 (**0.004**)	-1.47 (0.14)	-2.13 (**0.03**)	-
55	-2.49 (**0.01**)	-1.18 (0.24)	-1.73 (0.08)	-
59	-2.40 (**0.02**)	-0.71 (0.48)	-1.68 (0.09)	-
65	-	2.56 (**0.01**)	2.01 (**0.04**)	2.85 (**0.005**)
74	-	-0.98 (0.32)	-1.28 (0.20)	-

Score for each of the $\sim$ 11,000 registered students, wherein the average is computed over the mini-scores for all the mini-assessments the student has completed.

Then, the following Analysis of Covariance model is fitted to the data. To control for the students math preparedness and school characteristics, we include the cluster information as a factor in the model.

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}, \quad (2)$$

wherein y_{ij} is the EoC score and x_{ij} is the Average Mini-Assessments Score for the j^{th} student in the i^{th} cluster. Further, μ is the overall mean effect and α_i is the additional effect due to the assignment of the student to the $i-$th cluster that accounts for prior math knowledge, grade and school characteristics of the students.

Table 6 depicts the estimated regression coefficients, their standard errors, together with the value of the test statistic and the p-value corresponding to the significance test for each of the coefficients. All p-values are significantly smaller than the nominal 0.05 (or 0.01) level, thus indicating that the corresponding predictor has a significant effect on the EoC test score. The estimated coefficient for the mini-assessment is 1.15. This small, but statistically significant coefficient indicates that an increase of one point in the average student score on the mini-assessment corresponds to an expected improvement in the EoC score of 1.15 (the corresponding scaled regression coefficient is 7.46) points. At first glance, this relationship between the average mini-score performance and the EoC test seems of limited practical significance. However, when examining the distribution of EoC scores across all students ($\sim 90,000$) that used the Math Nation platform at some point in time (not necessarily participants in the current study), we find that about 1.9% are within 1 point of the passing threshold. Hence, in light of this information, it is reasonable to posit that the recommendation algorithm would have been beneficial for a good number of students, if it were adopted and used by all platform participants.

Table 6: Results of the Analysis of Covariance model: Response EoC Score; categorical predictor cluster and numerical predictor Average Mini-Assessments Score

Coefficients	Estimate	Std. Error	t-value	p-value
Intercept	464.29	2.61	178.18	<2e-16
Cluster 3	29.79	3.63	8.21	3.9e-16
Cluster 4	29.61	2.77	10.70	<2e-16
Cluster 5	37.06	2.92	12.70	<2e-16
Cluster 6	38.99	3.13	12.46	<2e-16
Cluster 7	36.84	2.77	13.29	<2e-16
Cluster 8	49.74	2.92	17.02	<2e-16
Cluster 9	53.47	2.87	18.62	<2e-16
Cluster 10	73.48	2.90	25.33	<2e-16
Cluster 11	68.89	3.21	21.49	<2e-16
Cluster 12	75.34	2.88	26.13	<2e-16
Avg. Mini-Assessments	1.15	0.15	7.46	1.3e-13

5. DISCUSSION

The analysis of the data from the randomized control study provide a number of useful insights in designing recommendation strategies for IVLE. Firstly, the recommendation algorithm holds a lot of promise, but as it is well known in reinforcement learning, it requires adequate amount of usage to "explore" various possibilities in order to maximize expected reward. The adequate usage requirement is also discussed in the literature evaluating recommendation strategies for Massive Online Open Courses; see [6, 17, 18] and references therein. As mentioned in Section 3, an initial evaluation of the proposed algorithm during its development phase based on synthetic data indicated that it starts yielding satisfactory results, in terms of students improving their performance on the mini-assessments, once students follow its recommendations for over 45 times. The results of the analysis in Section 4 are in line with the aforementioned finding. As Table 3 indicates, the recommendation strategy shows significant impact starting from a usage level of 48. Further, note that in our study the primary outcome under consideration is the EoC test that takes place at the end of the academic year, as opposed to a more direct outcome related to the recommendation algorithm, such as performance over time on the mini-assessment tests. In many studies in the literature (e.g., [1, 22], assessment of a recommendation algorithm was based on more immediate outcomes (e.g., the mini-assessments in our setting), as opposed to a more distal outcome, such as the EoC. Nevertheless, the results of our experiment indicate that with stronger student engagement the developed algorithm could be more widely beneficial.

To address the issue of low usage, a new experiment has been designed, wherein the teachers are directly involved in the implementation of the recommendation system in the classroom, which is expected to yield higher levels of engagement of students with the IVLE platform. This experiment is under way at the time of this publication.

It is also worth mentioning that our first analysis was of "Intent-to-Treat" type, because it evaluated the effect of being randomly assigned to treatment or control groups without consideration of the extent that students used the recommendation strategy. On the contrary, traditional Complier Average Causal Effect analysis [21, 25] is based on "Treatment-on-the-Treated" principle, wherein one estimates the treatment effect for those who complied with the treatment. The latter constitutes a direction of future research.

Another issue of broader interest is that many IVLE recommendation algorithms are designed to assign test problems in an adaptive way, as opposed to assigning videos that Math Nation does. However, in the modified implementation of the algorithm currently under evaluation, the student can skip watching the recommended video and take the mini-assessment directly; in case, (s)he gets less than two of the questions correctly, the algorithm recommends to watch the segment of the video that covers the corresponding material and then retake the mini-assessment. This modification aims to enhance the emphasis of the recommendation algorithm on solving problems, but at the same time enable students to review relevant material to questions that they answered incorrectly.

6. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

7. REFERENCES

[1] J. Bassen, B. Balaji, M. Schaarschmidt, C. Thille, J. Painter, D. Zimmaro, A. Games, E. Fast, and J. C. Mitchell. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[2] A. C. Cameron and D. L. Miller. A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372, 2015.

[3] Y. Chen, X. Li, J. Liu, and Z. Ying. Recommendation system for adaptive learning. *Applied psychological measurement*, 42(1):24–41, 2018.

[4] M. Chi, K. VanLehn, D. Litman, and P. Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1-2):137–180, 2011.

[5] S. Doroudi, K. Holstein, V. Aleven, and E. Brunskill. Towards understanding how to leverage sense-making, induction and refinement, and fluency to improve robust learning. *International Educational Data Mining Society*, 2015.

[6] K. S. Hone and G. R. El Said. Exploring the factors affecting mooc retention: A survey study. *Computers & Education*, 98:157–168, 2016.

[7] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[8] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark. Intelligent tutoring goes to school in the big city. 1997.

[9] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[10] J. A. Kulik and J. Fletcher. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, 86(1):42–78, 2016.

[11] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, 106(4):901, 2014.

[12] S. Mojarad, A. Essa, S. Mojarad, and R. S. J. de Baker. Studying adaptive learning efficacy using propensity score matching. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK18)*, 2018.

[13] P. Morgan and S. Ritter. An experimental study of the effects of cognitive tutor algebra I on student knowledge and attitude. Pittsburgh, CA: Carnegie Learning Inc., 2002.

[14] R. Murphy, L. Gallagher, A. Krumm, J. Mislevy, and A. Hafter. Research on the use of Khan Academy in schools. Menlo Park, CA: SRI Education, 2014.

[15] S. A. Niaki, C. P. George, G. Michailidis, and C. R. Beal. The impact of an online tutoring program for algebra readiness on mathematics achievements; results of a randomized experiment. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 363–372. ACM, 2019.

[16] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144, 2014.

[17] J. Reich. Mooc completion and retention in the context of student intent. *EDUCAUSE Review Online*, 8, 2014.

[18] J. Reich and J. A. Ruipérez-Valiente. The mooc pivot. *Science*, 363(6423):130–131, 2019.

[19] S. Ritter, J. Kulikowich, P. Lei, C. McGuire, and P. Morgan. Big data comes to school: Implications for learning, assessment, and research. In *15th International Conference on Computers in Education: Supporting Learning Flow through Integrative Technologies, ICCE 2007*, pages 13–20, 2007.

[20] J. P. Rowe and J. C. Lester. Improving student problem solving in narrative-centered learning environments: A modular reinforcement learning framework. In *International Conference on Artificial Intelligence in Education*, pages 419–428. Springer, 2015.

[21] B. J. Sagarin, S. G. West, A. Ratnikov, W. K. Homan, T. D. Ritchie, and E. J. Hansen. Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological methods*, 19(3):317, 2014.

[22] S. Shen, M. S. Ausin, B. Mostafavi, and M. Chi. Improving learning & reducing time: A constrained action-based reinforcement learning approach. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 43–51, 2018.

[23] S. Shen, B. Mostafavi, C. Lynch, T. Barnes, and M. Chi. Empirically evaluating the effectiveness of pomdp vs. mdp towards the pedagogical strategies induction. In *International Conference on Artificial Intelligence in Education*, pages 327–331. Springer, 2018.

[24] S. Steenbergen-Hu and H. Cooper. A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106(2):331, 2014.

[25] E. A. Stuart, D. F. Perry, H.-N. Le, and N. S. Ialongo. Estimating intervention effects of prevention programs: Accounting for noncompliance. *Prevention Science*, 9(4):288–298, 2008.

[26] K. Vanlehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.

[27] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[28] L. S. Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.