A novel video recommendation system for algebra: An effectiveness evaluation study

Walter L. Leite

University of Florida, walter.leite@coe.ufl.edu

Samrat Roy

University of Florida, samratroy@ufl.edu

Nilanjana Chakraborty

University of Florida, nchakraborty@ufl.edu

George Michailidis

University of Florida, gmichail@ufl.edu

A. Corinne Huggins-Manley

University of Florida, amanley@coe.ufl.edu

Sidney K. D'Mello

University of Colorado Boulder, sidney.dmello@gmail.com

Mohamad Kazem Shirani Faradonbeh

University of Georgia, mohamadksf@uga.edu

Emily Jensen

University of Colorado Boulder, Emily.Jensen@colorado.edu

Huan Kuang

University of Florida, huan2015@ufl.edu

Zeyuan Jing

University of Florida, jingzeyuan@ufl.edu

This is the author version of the manuscript. The citation is:

Walter L. Leite, Samrat Roy, Nilanjana Chakraborty, George Michailidis, A. Corinne Huggins-Manley, Sidney K. D'Mello, Mohamad Kazem Shirani Faradonbeh, Emily Jensen, Huan Kuang and Zeyuan Jing. 2022. A novel video recommendation system for algebra: An effectiveness evaluation study. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21-25, 2022, Online, USA*. ACM, New York, NY, USA, 12 Pages. **https://doi.org/10.1145/3506860.3506906**

ABSTRACT

This study presents a novel video recommendation system for an algebra virtual learning environment (VLE) that leverages ideas and methods from engagement measurement, item response theory, and reinforcement learning. Following Vygotsky's Zone of Proximal Development (ZPD) theory, but considering low affect and high affect students separately, we developed a system of five categories of video recommendations: 1) Watch new video; 2) Review current topic video with a new tutor; 3) Review segment of current video with current tutor; 4) Review segment of current video with a new tutor; 5) Watch next video in curriculum sequence. The category of recommendation was determined by student scores on a quiz and a sensor-free engagement detection model. New video recommendations (i.e., category 1) were selected based on a novel reinforcement learning algorithm that takes input from an item response theory model. The recommendation system was evaluated in a large field experiment, both before and after school closures due to the COVID-19 pandemic. The results show evidence of effectiveness of the video recommendation algorithm during the period of normal school operations, but the effect disappears after school closures. Implications for teacher orchestration of technology for normal classroom use and periods of school closure are discussed.

CCS CONCEPTS •Human-centered computing~Human computer interaction (HCI)~Interactive systems and tools•Computing methodologies~Machine learning~Learning settings~Online learning settings•Computing methodologies~Machine learning~Learning paradigms~Reinforcement learning

Additional Keywords and Phrases: recommender system, engagement detection, item response theory, algebra, effectiveness study

1 Introduction

Virtual learning environments (VLE) frequently use learning analytics to provide learning resource recommendations and personalized content sequencing, which can accomplish a variety of objectives, including remediation, selection of learning resources with optimal level of challenge to the student, and maintaining student engagement. Learning resource recommendation systems for e-learning are very diverse, but most commonly use content-based filtering, collaborative filtering, or hybrid approaches [1, 2]. Content sequencing systems have used reinforcement learning (RL) methods [3], such as partially observable Markov decision processes (POMDP) and Multi-armed bandits (MAB) [4, 5]. In contrast to these systems, the current study presents an innovative video recommendation system which combines the use of sensor-free student engagement detection and formative assessment based on item response theory (IRT). Therefore, each video suggested to a student during his/her classroom time or home-based use of the VLE is personalized to match the student's learning needs and engagement state. Few VLE have incorporated student engagement, which is critical because lack of engagement negatively relates to student achievement [6].

The current study is an effectiveness study, which in contrast with efficacy studies, was conducted in natural settings instead of controlled laboratory settings. The strength of the effectiveness study is that it supports the scalability of the system and generalizability of effects to diverse populations of students. The study consisted of a multi-site randomized control trial with assignment at the student level. We conducted the study both before and during the COVID-19 pandemic, which allowed for the examination of video recommender effects in regular school settings and in emergency remote instruction settings. The goals of the study were to estimate both the intent to treat effect (ITT), which is the effect of offering video recommendations regardless of compliance, and the complier average causal effect (CACE) [7], which is the effect for those in the treatment group who watched videos when offered.

2 BACKGROUND AND RELATED WORK

2.1 Content sequencing

The current study shares similarities with content sequencing research using POMDP [8] and MAB [9] for intelligent tutoring systems (ITS), because there is a focus in optimizing learning, and the use of multiple formative assessments, and tracking student knowledge. Also, we anchored the development of the recommender system on Vygotsky's theory of Zone of

Proximal Development (ZPD) [10], which is the learning theory used to support many ITS (e.g., [11]). However, there are some important differences: 1) Although content sequencing systems can be used for any type of learning object (e.g., [12]), research for ITS frequently aim to define an optimal order of problems for students to solve (e.g., [13, 14]), while the system presented here provides a video recommendation; 2) In content sequencing systems for ITS, the learning path is usually fixed by the system and students are required to follow the sequence chosen by the system in order to continue using it. In contrast, the system presented here is driven by the student and teacher, who may decide to skip the next video recommended or the next assessment; 3) The system presented here can leverage short-horizon data, as opposed to existing significantly "data-hungry" RL algorithms 4) In ITS, student mastery/non-mastery of concepts are usually modelled with Bayesian Knowledge Tracing [15], while in the current system we use IRT to update continuous ability estimates and define the students' ZPD.

2.2 Student Engagement

The importance of engagement to learning has been recognized and investigated for decades. The research broadly supports the following general conclusion: a student who is engaged is primed to learn; a student who is disengaged is not. For example, a recent meta-analysis based on 29 studies (N = 19,052 students) found an overall significant negative mean Pearson r = -.24 of boredom on academic outcomes [6]. Much of the research on engagement has focused on traditional learning which occurs in the classroom and school settings [16]. However, with the advent of mobile devices, much of learning currently occurs via digital media. This poses a challenge since it is particularly difficult to engage students when they interact with digital learning technologies, often in isolation. Whereas a gifted human teacher or an expert tutor can design collaborative activities to increase engagement and can adapt the lesson when engagement appears to be waning, it is difficult for current digital learning technologies to promote and sustain meaningful engagement for all learners. Even when a learning technology is successful at initially capturing students' attention, it has little recourse when novelty fades, the student gets stuck, or boredom eventually sets in. Thus, there might be benefits to accounting for engagement in VLE.

2.3 Item Response Theory

While many VLE include formative assessments of student ability, most either use total scores from these assessments or Bayesian knowledge tracing (BKT) [15] to represent student knowledge states. The use of total scores ignores item differences and measurement error, while BKT requires substantial expert input for set up, and focuses on mastery/non-mastery of concepts. The use of item response theory (IRT) [17] to determine the current student ability [18, 19] is not common in VLE, despite being easier to set up than BKT, providing continuous scores, being sample independent, and accounting for variation in item discrimination, difficulty and measurement error [20]. The formal relationship between BKT and IRT was presented by Deonovic et al. [21].

Using IRT for assessment within VLE offers several benefits: 1) Pre-screening of tests for problematic items with very small or very large difficulty parameters, or very low or negative discrimination parameters; 2) Scoring of tests accounting for differences in item difficulty and discrimination; 3) Estimation of student-specific standard error of measurement for each test administration, and building of confidence intervals; 4) Selection of the best items for each student at their current development level to create personalized tests.

There have been few applications of IRT to designing learning resource recommender systems [18, 19]. For example, Baylari and Montazer [18] describe a multi-agent system that combines the use of IRT for a test agent and neural networks for a remediation agent. Liu and Yu [19] use person-fit statistics based on item response theory to detect aberrant response patterns, and a computer agent to deliver encouragement messages to students.

3 THE VIDEO RECOMMENDATION SYSTEM

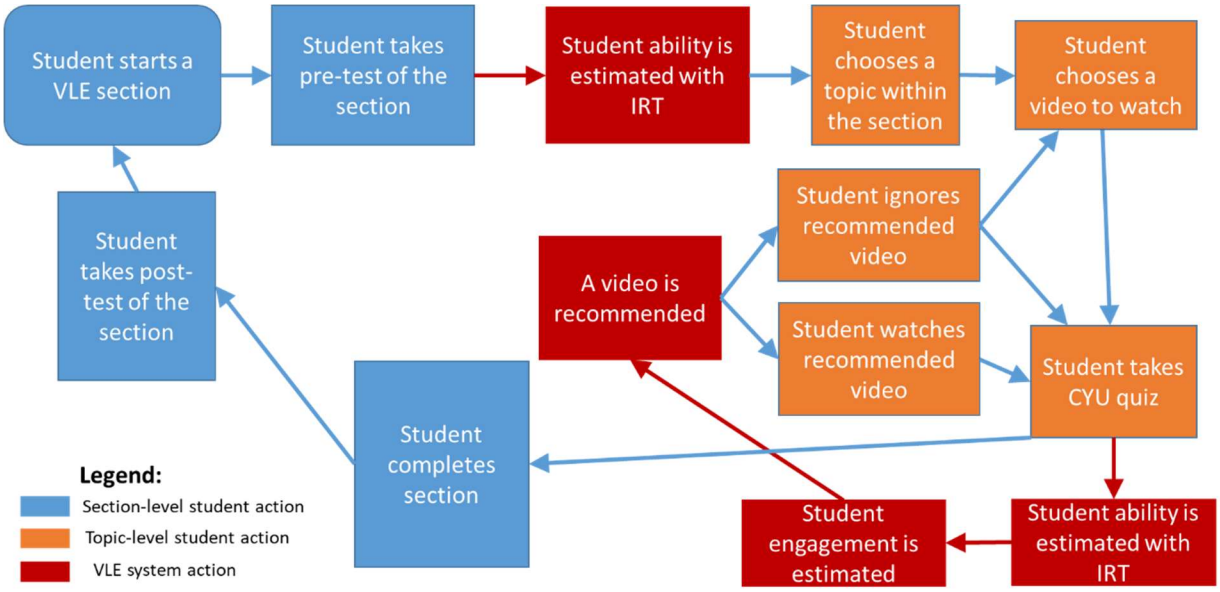

Figure 1. Flowchart of the video recommendation system

3. 1 The Virtual Learning Environment

A flowchart of the video recommendation system is depicted in Figure 1. Students received video recommendations within the Algebra Nation VLE [22], which is organized as a list of 93 algebra topics grouped into 10 sections (e.g., Section 1 – Expressions, Section 2 - Equations and Inequalities, etc.), with between 6 and 12 topics per section (e.g., Section 1 Topic 1: Using Expressions to Represent Real-World Situations; Section 1 Topic 2: Properties of Exponents, etc.) For each topic, students can choose to watch 6 versions of a video, each delivered by a different tutor using his/her own presentation style and cadence [23]. Each topic also has a fixed 3-question quiz referred to as Check Your Understanding (CYU). After the CYU is completed, students can review incorrect questions. The VLE also offers 10-item Test Yourself (TYS) assessments after each section. The items for each student's TYS are randomly selected from a large pool, and there is a solution video for each item. The VLE has other resources, such as a discussion forum [24], which will not be addressed in this study. Students use the VLE under the guidance of their teachers. Common ways that teachers incorporate the VLE into their instruction include showing videos to the whole class, asking students to watch specific videos at school or at home, and assigning CYU and TYS as warm-up activities, in-class assessment or homework [25]. Student use of the VLE has been shown to relate positively to student achievement [26, 27].

3.2 Learning theory

Vygotsky's theory of Zone of Proximal Development (ZPD) [10] has been extensively used to provide theoretical grounding for research on adaptive learning systems, resulting in the development of intelligent learning environments (ILE) **[28]** and ITS [5]. ZPD defines an area of development beyond the students' current ability that the student is able to attain with assistance. Therefore, intelligent VLE attempt to provide the needed assistance to move ability within the ZPD. However, this use of the ZPD requires formative measures of students' ability, so that a learning resource that probes for potential development can be provided. In the VLE for the current study, the CYU was the formative measure, and ability was estimated with IRT. The ZPD for each student is approximated in the recommendation policy presented in section 3.4 as the distance between a student's ability and the mean of peer ability estimates.

3.3 Engagement measurement

D'Mello, Dieterle and Duckworth [29] proposed the advanced, analytic, automated (AAA) approach to measure engagement for interactions with digital learning technologies. This approach focuses on a person-oriented operationalization of engagement as the momentary affective and cognitive states that arise throughout the learning process. The core idea of the AAA approach is to use machine learning to train a model that can estimate latent mental states associated with engagement (e.g., concentration, interest) from machine-readable signals.

The model used in this work was trained on a large-scale dataset of 69,174 students who used the VLE as part of their regular algebra classes for a semester. We used experience sampling to collect 133,966 self-reports (on a 1 to 5 scale) of 18 mental states

related to engagement. The positive valence group consisted of Happiness, Hopefulness, Contentment, Relief, Pride, Pleasantness, Engagement, Interest, and Arousal. The negative valence group consisted of Frustration, Confusion, Disappointment, Anxiety, Sadness, Mind Wandering, and Boredom. Two additional states - Curiosity and Surprise – did not strongly align with either valence group. We computed 22 activity features (e.g., viewing a video, pausing a video, taking a quiz) extracted from the VLE log files in 5-minute windows prior to a self-report survey.

We originally trained student-independent supervised learning models to independently predict each affective state from the features, achieving correlations with self-reported affect ranging from .08 to .34 with a mean of .25 [30]. We also demonstrated that the models generalized across socio-demographics and usage patterns with the VLE [31]. However, we found that predictions for individual affective states were strongly associated within the above positive and negative valence groups (pairwise correlations ranging from 0.71 to 0.98), demonstrating a lack of discrimination [31]. Accordingly, we combined the data from the positive and negative groups to yield a 1 (low) to 5 (positive) engagement scale. Specifically, we inverted the responses to negative surveys so that responses of 5 on a negative survey (highest negative) corresponded to a 1 on the one-dimensional scale (not positive) engagement scale. Similarly, a response of 1 on a negative survey (lowest negative) corresponded to a 5 on the new scale (highest positive). We then trained Bayesian Ridge regression models to predict responses on our new one-dimensional scale from the 22 features using 5-fold student-level cross validation where a given student's data is either in the training or testing set in each fold. This unipolar model achieved an average Spearman correlation between folds of 0.22 (SD = 0.01) for individual surveys which is within the range reported for individual states as discussed above. The 0.22 correlation, which corresponds to a Cohen's d of 0.45 (medium effect), is admittedly modest, but is consistent with what can be achieved from using basic behaviors for engagement modeling in a large, heterogeneous dataset [29]. The pertinent question is whether it is sufficiently accurate to tailor recommendations that are sensitive to student engagement. Accordingly, we subsequently deployed the model in the current application, which entails providing an estimate of engagement for a given student at any given time. For this, we computed the 22 features for the 5-minute window preceding the specified timestamp and submitted them to the above regression model, which generates an estimated engagement score on the one-dimensional scale.

3.4 Recommendation Policy

Following the principle of staying within the students' ZPD when providing recommendations, but addressing low engagement and high engagement students separately, we created a system of five categories of video recommendation: 1) View new video; 2) Review current topic video with a new tutor; 3) Review segment of current video with current tutor; 4) Review segment of current video with a new tutor; 5) View next video in curriculum sequence. The category of recommendation that the student received was defined as shown in Table 1, based on the students score on the CYU and the current engagement estimate obtained.

Table 1. Video Recommendation System

CYU score	Engagement Threshold	Probability of Recommendation of Category C
0	< 3.5	p(C=1) = 0.7 p(C=2) = 0.3
0	>= 3.5	p(C=1) = 0.3 p(C=2) = 0.7
1	< 3.5	p(C=1) = 0.3 p(C=4) = 0.7
1	>= 3.5	p(C=1) = 0.3 p(C=3) = 0.7
2	Any	p(C=3) = 1
3	Any	p(C=5) = 1

For a Category 1 recommendation, a new video is selected as shown in Algorithm 1 below. In this algorithm, for a student *i* in a cluster of students of size *n*, a topic video *j* is selected at a time *t*, amongst the *r* available videos. Using VLE data from a similar experiment performed in the previous year with student users of the VLE [32], the importance weight w_j was estimated for video *j*. The estimation procedure will be explained shortly. Further, the input ability estimate a_{ij} is the estimated ability from the student's response to the CYU using a 2-parameter logistic IRT model. The importance weights w_j were estimated using the Orthogonal Greedy Algorithm (OGA) [33], which estimates w_j values using the data collected from the students taking the topic-specific CYU, as well as the state mandated End-of-Course (EOC) exam, in the previous educational year. The estimation procedure proceeds as follows: First, OGA selects the topic whose CYU has the highest correlation with the EOC score, among all topics $1,2,\dots,r$. Letting j_1 be the selected topic, the algorithm then fits a linear regression model with the CYU of $j_1$as the predictor and the EOC score as the response, and computes the residuals (i.e., the difference between the predicted and the actual

EOC scores). The resulting slope of this regression step is saved as the importance weight for topic j_1, i.e., w_{j_1}. Next, OGA proceeds by treating the above-mentioned residuals as the response, and selects the second topic, j_2, whose CYU has the highest correlation with the response among all remaining topics (i.e., all topics excluding j_1). Fitting a linear regression model to the predictor (CYU of j_2) and the response (residuals of the previous step), the slope estimate is saved as w_{j_2}. Then, the algorithm iterates finding the new residuals, treating them as the response of the next step, selecting the new topic j_3, and so forth.

ALGORITHM 1. New Video Recommendation Policy for Student i

Inputs: *initial ability estimates* $\{a_{ij}(0)\}, 1 \le i \le n, 1 \le j \le r.$

Output: *sequence of recommended videos* $\hat{j}\ (t) \in \{1, \ldots, r\}, t\ \ge 0$

for $t = 0, 1, \ldots$ **do**

Compute peer ability-estimates

$$b_j(t)\ = n^{-1} \sum_{i=1}^{n} a_{ij}\ (t).$$

Compute the probability distribution $\{p_j(t)\}, j = 1,2, \ldots, r,$

$$p_j(t) = \frac{exp[-\ w_j\ (a_{ij}(t) - b_j(t))]}{\sum_{j=1}^{r} exp[-\ w_j\ (a_{ij}(t) - b_j(t))]}.$$

Sample $\hat{j}\ (t)\ from\ the\ distribution\ \{p_j(t)\}_{,1 \le j \le r}.$

Read $\{a_{ij}(t+1)\}\ ,1 \le i \le n\ ,1 \le j \le r$ *from the database.*

end for

The OGA algorithm described above possesses some critical properties rendering it a scalable, reliable, and desired method, for estimating the importance weights. First, there are more than 90% missing values in the data. Thanks to the Greedy nature of OGA, the algorithm treats the topics one-at-a-time, and no additional step is required for missing data. That is, the method is remarkably robust to the issue of having large fractions of data being unavailable [33, 34]. Second, the procedure is dynamic in a sense that for any arbitrary number of videos that the user student might engage in, the selected subset of videos represents the most predictive subset among all possible subsets of videos of the same cardinality [33, 34]. Note that the latter issue of unpredictable number of videos exacerbates the first issue of missing data. Further considerations consist of the implementation feasibility for the fairly large number of involved parameters (large number of videos for every student, while the number of students is very large). Accordingly, a sequential adaptive decision-making algorithm needs to provide reliable recommendation as quickly as possible [33, 34].

The algorithm for Category 1 recommendation required two preparatory data analyses: First, we used CYU responses from the previous year to estimate the item difficulty and discrimination parameters of all items in the CYU item pool. This was accomplished with a 2-PL IRT with bias correction by neural networks [35]. This step was needed so that IRT ability estimates could be obtained immediately after a student completed a CYU. Second, we clustered students into 20 clusters of equal size using quantiles of a Mahalanobis distance from the minimums of a matrix containing three measures of previous student achievement (i.e., average ability in previous CYU, average ability in previous TYS, and score on the EOC assessment administered by the school district on the previous year). The clusters were used in the algorithm for Category 1 recommendation to determine which students were similar with respect to previous ability.

For Category 3 and 4 recommendations, the segment of the current topic video that were most related to the questions that the students answered incorrectly were determined by expert review. The recommendation for a new tutor for Categories 2 and 4 was implemented by showing the students the list of tutors and prompting to select a tutor, but students had the option of keeping the same tutor. Category 5 recommendations were the next video in the curriculum sequence of the VLE.

4 RESEARCH QUESTIONS

As mentioned in the introduction, this study investigated the ITT and CACE of the video recommendation system. Therefore, the first analysis undertaken addresses the following research question: "Did the students, who were offered video recommendations perform better on the post-test assessments than the students who were not offered such recommendations?". This research question focuses on a comparison of treatment and control groups with respect to student ability. To that end, the

outcome is the ability estimated with IRT based on student responses to the post-test (Post-test ability). We estimate the effect of the treatment/control assignment on the ability, while controlling for differences in student engagement, pre-test ability, the VLE section and cluster. This is an ITT analysis [36], because it estimates the causal effect of the policy of offering students a video recommendation, without examining whether the student complied with the recommendation. In effectiveness studies, an ITT analysis is important because it indicates whether the intervention will have an effect in the population of interest without any incentives for participation. To investigate how the ITT effect was impacted by school closures due to COVID-19, we addressed this research question for both the period of normal school operations as well as after school closures.

Very often the subjects fail to comply with their experimental assignments. In such situations, the CACE is the treatment effect for those who comply with their original assignments to the treatment/control groups [7]. To that end, our second analysis addresses the following research question: "What is the causal effect of video recommendations on the achievement of those students who watched the recommended videos when offered?" This research question was addressed for both pre and post school closure periods.

5 METHODS

5.1 Participants

For the effectiveness study, a field experiment was implemented in three large school districts in the southeast United States. A total of 18,925 middle and high school students in from 152 teachers in 149 schools were randomly assigned to treatment or control groups with equal probability, with blocking by teachers[1]. A video recommendation was presented to students as a pop-up screen at the top center part of the window every time a student completed a CYU. Students could click on the recommendation, or ignore it, in which case the recommendation moved to a smaller screen on the bottom right corner of the window. Students in the treatment group were presented with a pop-up video recommendation from Category 1 to 5 according to probabilities shown in Table 1. Students in the control group were always presented with a pop-up video from Category 5. Treatment assignment status was blind to students and teachers. The study lasted for 17 weeks during the Spring 2020 semester (i.e., February 3rd to May 31st), but had a transition on March 17th when all schools were closed due to the COVID-19 pandemic and instruction resumed online.

5.2 Measures

Five-question pre-test and ten-question post-test algebra achievement measures were created by selecting items from the TYS item pool for each domain, so there were a total of 10 pre-test and post-test measures. The optimal pre-test and post-test items were selected for each student based on IRT such that they maximized the amount of reliable information that can be gleaned from a student, given the current estimate of the students' ability when they started the test (i.e., maximized Fisher information method) [20]. The ability estimates for the post-test obtained with IRT using the estimation method described in [35], were used as the outcome for the analyses reported in this study.

5.3 Analysis

5.3.1 Intent to Treat Analysis

To estimate the ITT effect, we fit a regression model with cluster-robust standard errors of following form:

$$post\ test\ ability_{kl} = \beta_{0l} + \beta_{1l} * \ treatment_{kl} + \ \beta_{2l} * \ pre\ test\ ability_{kl} + \beta_{3l} * engagement_{kl}$$

$$+ \sum_{c=2}^{20} \beta_{c\,4l} * (cluster_c)\_kl \ + \ \sum_{s=2}^{10} \beta_{s\,5l} * (section_s)\_kl \ \ + \ \ \varepsilon_{kl} \qquad (1)$$

1 Because of disruptions due to the COVID-19 pandemic, student demographic information could not be obtained from the school districts.

In the above model, $post\ test\ ability_{kl}$ is the ability calculated after the k^{th} post-test assessment by the students under the l^{th} teacher. $\beta_{1l} = \gamma_{10}$ is the effect corresponding to the binary variable that takes the value '1' if the student taking the assessment is under the treatment group and '0' if he or she is in the control group. Similarly, $\beta_{2l} = \gamma_{20}$ and β_{3l} are the effects corresponding to the pre-test ability and the engagement respectively. For c = 1,2,...,20, the dummy-coded indicator variables $(cluster_c)\_kl$ take the value 1 when a student comes from cluster *c*. Similarly, for s = 1,2,...,10, the indicator variables $(section_s)\_kl$ takes the value 1 when the k^{th} assessment taken under the l^{th} teacher is based on section *s*. Thus $\beta_{c\,4l}$ for c = 2,3,...,20 and $\beta_{s\,5l}$ for s =2,3,...,10 are the additional effects (as compared to the baselines: cluster 1 and section 1) corresponding to the clusters and the sections respectively. Given these notations, the estimate of the ITT is the estimated coefficient β_{1l} . The residuals ε_{kl} are i.i.d. from N $(0,\sigma^2)$. The model was estimated with the survey package [37] in R, and standard errors were estimated adjusting for clustering of students by teachers.

The dataset contained missing data because some students only took either the pre-test or the post-test for a section. To account for missing data under the assumption of a missing at random (MAR) mechanism, we used multiple imputation by chained equations, with predictive mean matching (PMM) as the univariate imputation method [38]. We generated 10 imputed datasets by imputing treatment and control groups separately [39], which leads to 10 ITT estimates, along with the corresponding standard error. Using Rubin's rules [40] to combine estimates from imputed datasets, the final estimate of the ITT (*ITT final*) was obtained as the average of the 10 estimates. The standard errors were obtained by combining the within-imputation variance (Var_w) and between-imputation variance (Var_b), as follows:

$SE\ final = \sqrt{Var_w + Var_b + \frac{Var_b}{10}}$, where Var_w is the average of the 10 ITT variances and Var_b is the variance of the 10 ITT estimates. Once the *ITT final* and *SE final* were obtained, we performed the 2-tailed Z test for the hypothesis $H_0: ITT = 0\ vs.\ H_1: ITT \neq 0$.

5.3.2 Complier Average Causal Effect Analysis

To briefly describe the complier average causal effect (CACE) analysis, let Z_i be the binary variable, that takes the value 1 when the i^{th} individual is assigned to the treatment group and 0 if he/ she is assigned to the control group. Let $D_i(z)$ be another binary variable that takes the value 1 (or, 0) if the i^{th} individual chooses to receive the treatment (or, chooses not to receive the treatment) when his /her original treatment assignment was $Z_i = z$. In our study, the students who were assigned to the control group (that is, z=0), were simply asked to watch the next video in the sequence (see Section 1.3). Thus $D_i(0) = 0$ for all i. However, the students who were assigned to the treatment group and were offered recommendations, may or may not actually follow the recommendations and thus $D_i(1)$ can be both 0 or 1. In our study, students in the treatment group receive a video recommendation every time they complete a CYU quiz (see Figure 1). Therefore, we considered a student to be complier if he/she was in the treatment group and watched at least one recommended video. Students in the treatment group who did not watch any recommended video were considered non-compliers. Therefore, in this study non-compliance can occur only among the treatment group (z=0), which is referred as "One-sided Non-Compliance", in the literature, to differ from the situation where the control group may also comply with the treatment.

The assumptions of the CACE analyses are [7]: 1) Stable Unit Treatment Value (SUTVA) assumption: The potential outcomes of one individual are not affected by the treatment assignment of other individuals; In our study, the fact that students and teachers could not tell which student was receiving the intervention helped with this assumption. 2) Monotonicity assumption: There are no defiers, which are individuals that choose to take the treatment because they are not assigned to receive it; in our study, because the VLE system controlled access to the recommended videos, the existence of defiers is not possible; 3) Exclusion restriction: to receive the benefit of the intervention, it is necessary to participate in it; Given these three assumptions, the CACE can be estimated as follows [7]:

$$CACE = \frac{Intent\ to\ Treat\ Effect\ (ITT)}{Proportion\ of\ Compliers\ (ITT_d)} = \frac{E[Y_i(z=1)] - E[Y_i(z=0)])}{E[D_i(z=1)]} \qquad (2)$$

where $Y_i(z=1)$ and $Y_i(z=0)$ are the potential outcomes of the i^{th} student (post-test ability) under treatment and control respectively. Note that, under one-sided non-compliance, $E[D_i(z=0)] = 0$ and thus, unlike two-sided Non-Compliance, this term does not appear in the expression of ITT_d. The estimate of the numerator is equivalent to the ITT estimate obtained in our

first analysis (estimate of the coefficient β_{1l}). On the other hand, ITT_d is estimated as the proportion of compliers among the students who were assigned to the treatment group, that is $\frac{\Sigma D_i(z) * I(Z_i=1)}{\Sigma I(Z_i=1)}$. Once the CACE is estimated, we performed the two-tailed Z-test for the hypothesis $H_0: CACE = 0 \; vs. H_1: CACE \neq 0$ to check the significance of the effect of compliance. As in the first analysis, in this case too, we estimate the CACE and the standard errors based on 10 different imputed datasets for both pre and post school closure periods and combined the results using Rubin's rules.

6 RESULTS

6.1 Intent to Treat Effects

Tables 2 displays the ITT estimates, their standard errors, and their p-values, based on combining 10 imputed datasets (see analysis section) for before and after school closures due to COVID19. It can be seen there was a statistically significant *ITT*= 0.05 (SE = 0.03, p = 0.043) for the period of normal school operation. Therefore, the treatment group had a larger gain in the post-test ability than the control group. The difference indicates a small average treatment effect of 0.05 standard deviations between treatment and control groups. As a percentage of mean gain (i.e., 0.57), this indicates that the treatment group had mean gains 8.7% higher than the control group. For the period of school closure and remote instruction during the COVID-19 pandemic, the estimated ITT was -0.009 (SE = 0.030, p = 0.775), which was not statistically significant. Therefore, there was no effect of offering video recommendations on student achievement during the period when schools were closed due to COVID-19.

The results of the ITT model (see Equation 1) indicate that engagement scores were not related to post-test ability either before (β_{3l}= -0.021, SE = 0.022, p = 0.349) or after (β_{3l}= -0.007, SE = 0.028, p = 0.805) schools closed. There was also no association between pre-test ability and post-test ability before (β_{2l} = -0.012, SE = 0.040, p=0.764) or after (β_{2l} = 0.025, SE = 0.112, p = 0.825) schools closed, which can be explained by the model also including the cluster dummy indicators.

As mentioned previously, clusters were defined based on previous student performance. Clusters were dummy-coded with cluster 1 as the reference category (see Equation 1). Therefore, the coefficients $\beta_{c\,4l}$ are post-test differences between Cluster *c* and cluster *1*. Before schools closed, these differences had a clear increasing trend from clusters 1 to 20. Furthermore, all differences were positive and statistically significant, except for the differences between clusters 2 and 1, and between clusters 4 and 1. However, after school closures, the post-test differences between the clusters shrank and there were no differences between cluster 1 and clusters 2 to 14. Only clusters 15 to 20 were significantly different from cluster 1 with respect to post-test, and the differences were approximately half of the size of the differences between these clusters before school closures. This indicates that students that had different performances on the VLE before school closures performed more similarly after school closures.

Sections were dummy coded so that Section 1 – Expressions was the reference category. Before school closures, the only sections with significant lower post-test scores than Section 1 were Sections 2 and 10. After school closures, there was no difference in post-test scores between the sections.

Table 2: Combined coefficients, Std. Errors and p-values across 10 imputed datasets for model in Equation 1, for before and after school closure periods

	Before school closure			After school closure		
	Coefficient	SE	p-value	Coefficient	SE	p-value
(Intercept)	**-0.752**	0.176	0.000	-0.339	0.602	0.573
ITT	**0.054**	0.027	0.043	-0.009	0.030	0.775
Pretest	-0.012	0.040	0.764	0.025	0.112	0.825
Engagement	-0.021	0.022	0.349	-0.007	0.028	0.805
Cluster 2	0.076	0.105	0.468	-0.095	0.161	0.556
Cluster 3	**0.350**	0.098	0.000	-0.021	0.196	0.914
Cluster 4	0.143	0.119	0.229	-0.144	0.258	0.576
Cluster 5	**0.376**	0.131	0.004	-0.229	0.222	0.301

Cluster 6	**0.587**	0.123	0.000	-0.053	0.188	0.777
Cluster 7	**0.474**	0.136	0.001	-0.228	0.299	0.446
Cluster 8	**0.438**	0.119	0.000	-0.034	0.173	0.843
Cluster 9	**0.308**	0.138	0.026	-0.133	0.220	0.547
Cluster 10	**0.610**	0.142	0.000	-0.073	0.179	0.682
Cluster 11	**0.765**	0.172	0.000	0.058	0.207	0.780
Cluster 12	**0.580**	0.143	0.000	0.043	0.195	0.825
Cluster 13	**0.776**	0.129	0.000	0.167	0.222	0.450
Cluster 14	**0.786**	0.146	0.000	0.245	0.170	0.148
Cluster 15	**0.916**	0.162	0.000	**0.400**	0.198	0.044
Cluster 16	**1.018**	0.146	0.000	**0.464**	0.203	0.022
Cluster 17	**1.128**	0.156	0.000	**0.564**	0.203	0.005
Cluster 18	**1.150**	0.140	0.000	**0.645**	0.207	0.002
Cluster 19	**1.285**	0.146	0.000	**0.614**	0.256	0.016
Cluster 20	**1.364**	0.153	0.000	**0.721**	0.262	0.006
Section 2	**-0.705**	0.280	0.012	0.092	0.281	0.744
Section 3	-0.267	0.268	0.320	-0.402	0.394	0.307
Section 4	-0.311	0.214	0.145	0.090	0.224	0.688
Section 5	-0.281	0.151	0.063	0.239	0.254	0.346
Section 6	-0.135	0.213	0.525	-0.013	0.290	0.964
Section 7	-0.099	0.143	0.486	-0.189	0.288	0.511
Section 8	0.121	0.151	0.425	0.208	0.294	0.479
Section 9	-0.034	0.199	0.865	-0.044	0.377	0.907
Section 10	**-0.521**	0.164	0.001	-0.097	0.320	0.761

***Note. Statistically significant coefficients are in bold**

6.2 Complier Average Causal Effects

Before schools closed, the proportion of compliers among the students who were assigned to the treatment group was ITT_d = 0.15921, SE = 0.0188, CI = [0.122, 0.196]. After schools closed, the proportion of compliers was ITT_d = 0.1123, SE = 0.0112, CI = [0.090, 0.134]. Although the observed proportion compliance after school closure was lower than before school closure, the confidence intervals overlap, so the proportions are not significantly different. For the before-closure period, the final CACE standardized estimate is 0.34 (SE = 0.17). The p-value for testing $H_0: CACE = 0\ vs. H_1: CACE \neq 0$ is 0.043. Therefore, the students who complied with the recommendations had a significantly larger gain in post-test ability than the ones who did not comply. As a percentage of mean gain (i.e., 0.57), this indicates that the compliers group had mean gains 60% higher than the control group. However, the CACE was not statistically significant for the period after schools closed (CACE = -0.076, SE = 0.266, p = 0.775).

7 DISCUSSION

The results show evidence of effectiveness of the video recommendation algorithm during a period of normal school operations. The video recommender was an add-on to an existing VLE, and implemented in a non-intrusive way where students could easily dismiss recommendations. Therefore, the results show the potential of non-intrusive machine-learning based

interventions to improve student achievement in e-learning. These results are aligned with literature on the promise of nudges during e-learning to improve student outcomes [41]. However, this evidence is based on a field study that was shorter than initially planned. The original goal was for the study to run for 17 weeks (i.e., one academic semester), but the period of normal school operation was cut to 5 weeks due to the COVID-19 pandemic. Although one would have expected that the transition to a fully online learning environment would have favored the recommendation algorithm, such an effect was not supported by the data. This could be due to the complete disruption of teacher strategies for orchestration of instruction [42] with the VLE, as well as disruption of student established self-regulated learning strategies [43] for mathematics learning. We found support for these explanations in a survey study of teachers who used the VLE during the school closures, and the majority of teachers who responded to the survey indicated that they had to reduce the number of assignments and make the assignments shorter[2].

We used estimates of engagement in the recommendation system to address the potential problem that students were disengaged by the content they were interacting with on the platform. Specifically, for low-engagement cases, the recommendation system tended to suggest a different tutor or a new video. Since behavior-based estimates from interaction logs provide a relatively weak signal of student engagement [44], the effectiveness of this approach is inherently limited by the accuracy of the models, which were admittedly modest. However, the current video recommender presented a marked improvement over a previous video recommender system for the same VLE based solely total tests scores and using a Markov Decision process, which resulted in a non-significant ITT in a field study [32]. Thus, the present results provide a useful baseline for what can be achieved with a set of generic activity features is used for engagement detection for the purpose of informing knowledge-based subsequent intervention strategies.

One pervasive difficulty in evaluating a video recommendation system in a natural field setting is that, although the system offered 62,617 video recommendations during the period of the study, the frequency of students watching recommended videos was low. This was also the case in a previous study of a video recommender for this platform [32]. A key finding from the previous study was that students with relatively high usage levels (over 45 VLE sessions) exhibited statistically positive effects in their assessment scores, whereas the treatment effect was not significant for the remainder of the students. To address this issue, we adopted the OGA algorithm that is robust to sparseness caused by students only occasionally watching recommended videos, which turned out to work well for the present field study. A limitation of the developed recommendation algorithm is that prior achievement data from the VLE were required to create student clusters to initiate recommendations. In the current study, because most students had used the VLE before, such data were available. Overcoming this limitation constitutes a topic of future research.

The current study randomized the availability of the recommendation and thus estimated the effect of offering the recommendation, which is known as an intent-to-treat (ITT) analysis. This type of analysis is useful for large-scale community-based interventions, where the interest is in the effectiveness of making a program available to the community, but it is known to provide conservative effect estimates [36]. Even though the ITT was small, the CACE results show that the effect for the compliers was much larger. Previous research on the VLE under consideration shows that its use is driven strongly by the teachers [25]. If teachers are shown the potential benefits of students watching recommended videos, they may be able to provide encouragement that increases usage of the video recommended system among their students. Therefore, the results of this study have implications for teacher professional development with respect to orchestrating the use of VLE in the classroom.

Teachers play a critical role in student use of educational technology, by demonstrating, recommending and rewarding use [25]. However, how a VLE is used by teachers varies considerably. Previous research [25] indicates that showing videos to the entire classroom, and assigning specific videos and CYU quizzes as homework is the most common strategy chosen by teachers to use the VLE. However, the predominance of these strategies may have contributed to the low frequency of recommended video views. This is because students are less likely to deviate from the assignment by watching a video that is recommended, but it is not part of the assignment's requirements.

This study showed treatment effect heterogeneity [45] across two very different learning settings: Student learning with a VLE during regular school operation; and student learning with a VLE when schools are closed and instruction is being delivered online. We found that this change in setting due to the COVID-19 pandemic removed the effect of the video recommendation system on student achievement. We also found that clusters of students with different achievement before

[2] Details on this survey study can be obtained by contacting the first author.

schools closed performed similarly after schools closed. However, our results suggest that students performed more similarly not because low achievement students did better, but because the distribution was compressed towards the bottom, due to difficulties students encountered with learning continuity after schools closed. New research about perceptions of students about their learning challenges during the pandemic offer support to this conclusion [46-48]. However, research on the effects of the pandemic on student achievement scores are scarce, and existing studies focus on student perceptions. The current study adds to the emerging research about the effects of the pandemic on student learning.

One limitation of the current study was the unavailability of demographic and economic data on students, which did not allow for an extensive examination of treatment effect heterogeneity across minority and economically disadvantaged groups. It may be that the there is substantial heterogeneity of treatment effects that is also related to economic disadvantage, which is an equity issue [49]. More specifically, it may be that students of higher socio-economic status, which may have more home support for using the VLE and feel more empowered to do so, may benefit more from interventions that branch-out from the learning path prescribed by teachers, such as the video recommendation system examined in this study.

The temporary closure of schools in 2020 due to the COVID-19 pandemic ushered a dramatic expansion of e-learning through the use of virtual learning environments (VLE), but one may hypothesize that the emergency nature of the adoption of VLE by students resulted in poorer outcomes than if e-learning was introduced as part of a regular school program [50]. The results of the current study provided some evidence in support of this hypothesis by showing that a recommender system that had significant ITT and CACE during normal school instruction had no effect when instruction was solely online. As research related to learning during the COVID-19 pandemic begins to emerge, we will know more about potential learning losses and which instructional strategies may have mitigated those losses.

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.

REFERENCES

1 Khanal, S.S., Prasad, P.W.C., Alsadoon, A., and Maag, A.: 'A systematic review: machine learning based recommendation systems for e-learning', Education and Information Technologies, 2019, 25, (4), pp. 2635-2664

2 Deschênes, M.: 'Recommender systems to support learners' Agency in a Learning Context: a systematic review', International Journal of Educational Technology in Higher Education, 2020, 17, (1)

3 Sutton, R.S., and Barto, A.G.: 'Reinforcement learning: An introduction (2nd edition)' (The MIT Press, 2018. 2018)

4 Clement, B., Oudeyer, P.-Y., and Lopes, M.: 'A Comparison of Automatic Teaching Strategies for Heterogeneous Student Populations', in Editor (Ed.)^(Eds.): 'Book A Comparison of Automatic Teaching Strategies for Heterogeneous Student Populations' (2016, edn.), pp.

5 Shen, S., Mostafavi, B., Barnes, T., and Chi, M.: 'Exploring Induced Pedagogical Strategies Through a Markov Decision Process Framework: Lessons Learned', Journal of Educational Data Mining, 2018, 10, (3), pp. 27-68

6 Tze, V.M., Daniels, L.M., and Klassen, R.M.: 'Evaluating the relationship between boredom and academic outcomes: a meta-analysis', Educational Psychology Review, 2016, 28, (1), pp. 119-144

7 Schochet, P.Z., and Chiang, H.S.: 'Estimation and Identification of the Complier Average Causal Effect Parameter in Education RCTs', Journal of Educational and Behavioral Statistics, 2011, 36, (3), pp. 307-345

8 Chen, Y., Li, X., Liu, J., and Ying, Z.: 'Recommendation System for Adaptive Learning', Appl Psychol Meas, 2018, 42, (1), pp. 24-41

9 Mu, T., Wang, S., Andersen, E., and Brunskill, E.: 'Combining adaptivity with progression ordering for intelligent tutoring systems'. Proc. Proceedings of the Fifth Annual ACM Conference on Learning at Scale2018 pp. Pages

10 Vygotsky, L.S.: 'Mind in society' (Harvard University Press, 1978. 1978)

11 Clement, B., Roy, D., Oudeyer, P.-Y., and Lopes, M.: 'Multi-Armed Bandits for Intelligent Tutoring Systems', Journal of Educational Data Mining, 2015, 7, (2), pp. 20-48

12 Spain, R., Rowe, J., Smith, A., Goldberg, B., Pokorny, R., Mott, B., and Lester, J.: 'A reinforcement learning approach to adaptive remediation in online training', The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 2021

13 Pardos, Z., and Heffernan, N.: 'Determining the significance of item order in randomized problem sets', Educational Data Mining, 2009

14 David, Y.B., Segal, A., and Gal, Y.a.K.: 'Sequencing educational content in classrooms using bayesian knowledge tracing', in Editor (Ed.)^(Eds.): 'Book Sequencing educational content in classrooms using bayesian knowledge tracing' (2016, edn.), pp. 354–363

15 Corbett, A.T., and Anderson, J.R.: 'Knowledge tracing: modeling the acquisition of procedural knowledge', User modeling and user-adapted interaction, 1995, 4, (4), pp. 253–278

16 Christenson, S.L., Reschly, A.L., and Wylie, C.: 'Handbook of research on student engagement' (Springer, 2012. 2012)

17 Lord, F., and Novick, M.: 'Statistical theories of mental test scores' (Addison-Wesley, 1968. 1968)

18 Baylari, A., and Montazer, G.A.: 'Design a personalized e-learning system based on item response theory and artificial neural network approach', Expert Systems with Applications, 2009, 36, (4), pp. 8013-8021

19 Liu, M.-T., and Yu, P.-T.: 'Aberrant Learning Achievement detection Based on Person-fit Statistics in Personalized e-Learning Systems', Educational Technology & Society, 2011, 14, (1), pp. 107–120
20 Embretson, S.E., and Reise, S.P.: 'Item response theory for psychologists' (Lawrence Erlbaum Associates, 2000. 2000)
21 Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., and Maris, G.: 'Learning meets assessment', Behaviormetrika, 2018, 45, (2), pp. 457-474
22 http://lastingercenter.com/portfolio/algebra-nation-2/, accessed 9/20/2019 2019
23 Shin, J., Balyan, R., Banawan, M., Leite, W.L., and McNamara, D.: 'Pedagogical Communication Language in Video Lectures: Empirical Findings from Algebra Nation'. Proc. 2021 Meeting of the International Society of Learning Sciences.2021 pp. Pages
24 Banawan, M., Balyan, R., Shin, J., Leite, W.L., and McNamara, D.: 'Linguistic Features of Discourse within an Algebra Online Discussion Board'. Proc. The 14th Conference on Education Data Mining, Paris, France, June 29th – July 2nd 2021 pp. Pages
25 Mitten, C., Collier, Z.K., and Leite, W.L.: 'Online Resources for Mathematics: Exploring the Relationship between Teacher Use and Student Performance', Investigations in Mathematics Learning, 2021, pp. 1-18
26 Leite, W.L., Cetin-Berber, D.D., Huggins-Manley, A.C., Collier, Z.K., and Beal, C.R.: 'The relationship between Algebra Nation usage and high-stakes test performance for struggling students', Journal of Computer Assisted Learning, 2019, 35, (5), pp. 569-581
27 Leite, W.L., Jing, Z., Kuang, H., Kim, D., and Huggins-Manley, A.C.: 'Multilevel Mixture Modeling with Propensity Score Weights for Quasi-Experimental Evaluation of Virtual Learning Environments', Structural Equation Modeling: A Multidisciplinary Journal, 2021, pp. 1-19
28 Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., and Koper, R.: 'Recommender Systems in Technology Enhanced Learning', in Ricci, F., Rokach, L., Shapira, B., and Kantor, P.B. (Eds.): 'Recommender Systems Handbook' (Springer, 2011), pp. 387-415
29 D'Mello, S.K., Dieterle, E., and Duckworth, A.L.: 'Advanced, Analytic, Automated (AAA) Measurement of Engagement during Learning', Educational Psychologist, 2017, 52, (2), pp. 104-123
30 Hutt, S., Grafsgaard, J., and D'Mello, S.K.: 'Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire School Year'. Proc. Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2019). New York2019 pp. Pages
31 Jensen, E., Hutt, S., and D'Mello, S.K.: 'Generalizability of Sensor-Free Affect Detection Models in a Longitudinal Dataset of Tens of Thousands of Students'. Proc. The 12th International Conference on Educational Data Mining2019 pp. Pages
32 Chakraborty, N., Roy, S., Leite, W.L., Faradonbeh, M.K.S., and Michailidis, G.: 'The effects of a personalized recommendation system on students' high-stakes achievement scores: A field experiment', in Editor (Ed.)^(Eds.): 'Book The effects of a personalized recommendation system on students' high-stakes achievement scores: A field experiment' (2021, edn.), pp.
33 Ing, C.K., and Lai, T.L.: 'A stepwise regression method and consistent model selection for high-dimensional sparse linear models', Statistica Sinica, 2011, pp. 1473-1513
34 Hsu, H.-L., Ing, C.-K., and Lai, T.L.: 'Analysis of High-Dimensional Regression Models Using Orthogonal Greedy Algorithms': 'Handbook of Big Data Analytics' (2018), pp. 263-283
35 Xue, K., Huggins-Manley, A.C., and Leite, W.: 'Semisupervised Learning Method to Adjust Biased Item Difficulty Estimates Caused by Nonignorable Missingness in a Virtual Learning Environment', Educational and Psychological Measurement, 2021
36 Gupta, S.K.: 'Intention-to-treat concept: A review
', Perspectives in Clinical Research, 2011, 2, (3), pp. 109–112
37 Lumley, T.: 'Analysis of Complex Survey Samples', Journal of Statistical Software, 2004, 9, (8), pp. 1-19
38 White, I.R., Royston, P., and Wood, A.M.: 'Multiple imputation using chained equations: Issues and guidance for practice', Statistics in Medicine, 2011, 30, (4), pp. 377-399
39 Puma, M.J., Olsen, R.B., Bell, S.H., and Price, C.: 'What to Do When Data Are Missing in Group Randomized Controlled Trials', in Editor (Ed.)^(Eds.): 'Book What to Do When Data Are Missing in Group Randomized Controlled Trials' (National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education., 2009, edn.), pp.
40 Rubin, D.B.: 'Multiple Imputation for Nonresponse in Surveys' (Wiley, 1987. 1987)
41 A., M., M., G., A., P., and V., D.: ' Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching', in Editor (Ed.)^(Eds.): 'Book Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching' (Springer, 2019, edn.), pp.
42 Prieto, L.P., Dlab, M.H., Gutiérrez, I., Abdulwahed, M., and Balid, W.: 'Orchestrating technology enhanced learning: a literature review and a conceptual framework', International Journal of Technology Enhanced Learning, 2011, 3, (6), pp. 583-598
43 Broadbent, J., Sharman, S., Panadero, E., and Fuller-Tyszkiewicz, M.: 'How does self-regulated learning influence formative assessment and summative grade? Comparing online and blended learners', The Internet and Higher Education, 2021, 50, pp. 100805
44 Hutt, S., Grafsgaard, J., and D'Mello, S.K.: 'Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire Schoolyear': 'Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2019)' (ACM, 2019)
45 Bolger, N., Zee, K.S., Rossignac-Milon, M., and Hassin, R.R.: 'Causal processes in psychology are heterogeneous', J Exp Psychol Gen, 2019, 148, (4), pp. 601-618
46 Yan, L., Whitelock-Wainwright, A., Guan, Q., Wen, G., Gasevic, D., and Chen, G.: 'Students' experience of online learning during the COVID-19 pandemic: A province-wide survey study', Br J Educ Technol, 2021
47 Dwidienawati, D., Tjahjana, D., Abdinagoro, S.B., and Gandasari, D.: 'E-Learning Implementation during The COVID-19 outbreak:The Perspective of Students and Lecturers', Journal of the Social Sciences, 2020, 48, (4), pp. 1190-1201
48 Chakraborty, P., Mittal, P., Gupta, M.S., Yadav, S., and Arora, A.: 'Opinion of students on online education during the COVID -19 pandemic', Human Behavior and Emerging Technologies, 2020, 3, (3), pp. 357-365
49 Tsai, Y.-S., Perrotta, C., and Gašević, D.: 'Empowering learners with personalised learning approaches? Agency, equity and transparency in the context of learning analytics', Assessment & Evaluation in Higher Education, 2019, 45, (4), pp. 554-567
50 https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teachingand-online-learning.2020