Author Version of Paper presented at the 2022 International Learning Analytics and Knowledge Conference (LAK2022), online, March 12-15, 2022

Heterogeneity of treatment effects of a video recommendation system for algebra

Walter L. Leite†
University of Florida
Gainesville, FL
walter.leite@coe.ufl.edu

Huan Kuang
University of Florida
Gainesville, FL
huan2015@ufl.edu

Zuchao Shen
University of Florida
Gainesville, FL
zuchao.shen@coe.ufl.edu

Nilanjana Chakraborty
University of Florida
Gainesville, FL
nchakraborty@ufl.edu

George Michailidis
University of Florida
Gainesville, FL
gmichail@ufl.edu

Sidney D'Mello
University of Colorado Boulder
Boulder, CO
Sidney.Dmello@colorado.edu

Wanli Xing
University of Florida
Gainesville, FL
wanli.xing@coe.ufl.edu

ABSTRACT

Previous research has shown that providing video recommendations to students in virtual learning environments implemented at scale positively affects student achievement. However, it is also critical to evaluate whether the treatment effects are heterogeneous, and whether they depend on contextual variables such as disadvantaged student status and characteristics of the school settings. The current study extends the evaluation of a novel video recommendation system by performing an exploratory search for sources of heterogeneity of treatment effects. This study's design is a multi-site randomized controlled trial with an assignment at the student level across three large and diverse school districts in the southeast United States. The study occurred in Spring 2021, when some students were in regular classrooms and others in online classrooms. The results of the current study replicate positive effects found in a previous field experiment that occurred in Spring 2020, at the onset of the COVID-19 pandemic. Then, causal forests were used to investigate the heterogeneity of treatment effects. This study contributes to the literature on content sequencing systems and recommendation systems by showing how these systems can disproportionally benefit the groups of students who had higher levels of previous algebra ability, followed more recommendations, learned remotely, were Hispanic, and received free or reduced-price lunch, which has implications for the fairness of implementation of educational technology solutions.

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

KEYWORDS

Video recommendation system, content sequencing, randomized controlled trial, heterogeneity of treatment effect

1 INTRODUCTION

The last two decades have witnessed one of the most dramatic developments in human education: the emergence of widespread technology-based instructional resources. Students and teachers now have access to unprecedented amounts of information that is available online, as well as unique learning opportunities afforded by intelligent tutoring systems, simulations, educational games, and MOOCs. The wealth of electronic data that accumulates as students use virtual learning environments offers tremendous opportunities to explore new approaches to instruction, including the use of students' data to support decisions about the learning opportunities provided to them or to future students, and provide teachers with timely information about students' learning progress. This approach can personalize instruction to an unprecedented degree including individualized based on learners' individual needs, goals, aptitude, cultural background, motivation, and related characteristics [1, 2].

Yet despite its promise, research on the effectiveness of technology-based personalized learning systems is only emerging, especially in the context of large-scale, multi-site, randomized-controlled trials that aim to causally attribute improved education to a particular learning intervention compared to a control group [3]. Effectiveness research is similarly nascent even in the case of well-established personalized learning paradigms, such as the sequencing learning activities using reinforcement learning, where ongoing research dates back approximately 60 years [4]. When effectiveness studies do exist, they tend to focus on the average treatment effect (ATE) across a sample representative of a broad

population, which may hide differences in benefit of the systems across students of racial/ethnic groups, language majority/minority groups, and low ability/high ability groups. Therefore, it is also critical to evaluate whether the treatment effects are heterogeneous [5], and whether it depends on contextual variables that naturally vary in education, such as disadvantaged student status and characteristics of the school settings.

Heterogeneity of treatment effects (HTE) is the variation of effects of an intervention across subgroups. It is a broader term than moderation of effects, because moderation implies a priori specification of moderator variables [6], while HTE allows for exploratory discovery of treatment modification. In medical research, the concept of precision medicine has led to much interest in HTE [7], because the treatment effect for a patient is likely to differ from the ATE identified in clinical trials if HTE is large. In educational technology research, moderation of effects of intelligent tutoring systems (ITS) have been studied in meta-analyses [8-10], but most have looked at variation in effects due to characteristics of the system or characteristics of the study implementation (i.e., methodology). However, because meta-analyses are restricted to using subgroup information reported in published papers and do not have access to raw student data, these meta-analyses may have missed important sources of HTE related to disadvantaged groups and contexts.

The objective of the current study is to extend the evaluation of a novel video recommendation system for an online Algebra learning platform, Algebra Nation [11], by performing an exploratory search for sources of HTE using causal forests [12, 13]. Causal forests use regression forests to estimate the individual conditional average treatment effects (iCATEs) for student participants given a set of covariates, using the potential outcomes framework [14]. This study contributes to the literature on content sequencing systems and recommendation systems by showing how these systems may have varying effects for subgroups of students, which has implications for the fairness of implementation of educational technology solutions. This study's design was a multi-site randomized controlled trial with an assignment at the student level across three large and diverse school districts in the southeast United States. The study occurred in Spring 2021, when some students were in regular classrooms and others in online classrooms. The analysis first replicates the estimation of the effect of a previous study of the same video recommendation system that occurred in Spring 2020, at the onset of the COVID-19 pandemic [15], thereby contributing to the evidence on the effectiveness of personalized learning. Then, causal forests were used to investigate HTE of the video recommendation system, which is a new approach that has not yet been applied to studies of technology-based personalized learning systems. The following questions are addressed: 1) What are the effects of the recommendation system on learning outcomes both within the platform and from standardized tests compared to a control group; 2) Is there substantial HTE of the video recommendation system? and 3) What student characteristics predict the HTE?

2 BACKGROUND AND RELATED WORK

2.1 Heterogeneity of treatment effects of intelligent tutoring systems

HTE of ITS has been investigated in the literature, but most frequently focusing on moderation of effects by characteristics of the ITS or the research design, rather than characteristics of the students using the systems. For instance, VanLehn [16] examined moderation of the performance of ITS due to different types of tutoring (i.e., substep-based tutoring, step-based tutoring, answer-based tutoring), but did not examine whether these effects vary across subgroups of students. Kulik and Fletcher's [8] and Xu et al. [17] meta-analyses found that the amount of improvement attributed to ITSs depended to a great extent on whether the outcomes were measured on locally developed or standardized tests, but neither study looked at whether ITS effects varied according to the characteristics of the student samples.

Two meta-analyses have examined HTE due to student characteristics. Ma et al. [9] performed a meta-analysis of the effects of ITS that included moderation by continent where the sample was taken, grade level, and level of prior knowledge of the students. Their results showed higher effects of ITS with samples from Asia, followed by Europe, and lower effects for samples from North America and Oceania. They also found that ITS used with middle-school and post-secondary students had higher effects than with elementary and high-school students. They found no statistically significant moderation of ITS effects by prior knowledge, but pointed out that there were only two studies with a high level of prior knowledge, and a large number of studies did not report prior knowledge of students. Therefore, their conclusion of no HTE due to prior knowledge should be taken as preliminary.

Moderator analyses performed in the meta-analysis by Steenbergen-Hu and Cooper [18] showed that the effectiveness of ITS for helping students drawn from the general population was greater than for helping low achievers. However, there were only two studies that reported effect sizes for low achievers. They found no difference by grade level. In a later meta-analysis, Steenbergen-Hu and Cooper [10] found that the effectiveness did not significantly differ by different ITS, subject domain, or the manner or degree of their involvement in instruction and learning, but the study did not examine whether there was HTE due to characteristics of the students. Given the results of these meta-analyses, it is evident that there is a need for studies of ITS to report moderation by subgroups of students. Out of the six meta-analyses reviewed, the only two that looked at HTE by student characteristics reported that the number of studies available with effects by subgroups was very small. The current study contributed to reducing this gap by providing an HTE analysis by student characteristics using an innovative machine learning method to detect HTE.

2.2. Detection of heterogeneity of treatment effects

Detecting treatment effect heterogeneity can be part of a pre-analysis plan where, in a design phase, researchers clearly state a set of covariates that are hypothesized to have different treatment effects, e.g., through moderation analysis along with an average treatment effect estimation [19, 20]. In experimental studies without any covariates to test treatment effect variation, one simple approach is to test whether the outcome distributions are the same for the treatment and control groups. The null hypothesis is that the distributions of the treatment and control groups are identical, differing with a constant shift by the average treatment effect. This null hypothesis can be tested through standard Kolmogorov–Smirnov- (KS) type tests [21] when the average treatment effect is known. However, since the shift (i.e., the average treatment effect) is not known, it is a nuisance parameter that must be estimated from the data and the main focus here is any potential treatment effect variation across all observed and unobserved covariates.

Ding et al., [21] have shown that when the average treatment effect is unknown, the null hypothesis of no treatment effect variation is not sharp. The approach proposed by Ding et al. [21] is to first construct a confidence interval (CI) for the average treatment effect, then repeat the Fisher randomization test procedure pointwise over the interval and take the maximum p-value. This approach guarantees a valid test for potential treatment effect variation across treatment and control groups and such a test can be generalized for testing treatment effects beyond a hypothesized model [21, 22].

It is also vital to explore and identify potential heterogeneity in treatment effects that go beyond a pre-analysis plan [23]. Here, we focus on methods for identifying HTE that are exploratory in nature [23]. There are conventional approaches to perform nonparametric estimation of HTE, such as matching, kernel methods, and series estimation [24]. In cases where a relatively large number of pertinent covariates are available, machine learning methods to detect HTE such as regression forests [25], causal forests [13], and Bayesian additive regression trees (BART) [26, 27] may be useful. However, as pointed out by Wager and Athey [13], these methods have "lacked formal statistical inference results" (p. 1229). The current study will focus on causal forests to detect HTE because this method can construct a valid confidence interval for the test of heterogeneity [23].

2.4. Causal Forests

Athey and Imbens [23] proposed honest, causal trees for the test of heterogeneity, which does not have restrictions on model complexity and could handle many variables. This data-driven approach splits the training sample into two parts: one for constructing the tree and another one for estimating treatment effects within leaves of the tree. This approach differs from conventional classification and regression trees (CART) in two ways: First, it focuses on estimating conditional average treatment effects rather than predicting outcomes. Second, it uses separate samples for two tasks: constructing the partition and estimating the effects within leaves of the partition. More importantly, the method provides valid confidence intervals of average treatment effects estimated for the identified subpopulations (leaves).

However, this approach requires the estimation of the true treatment effects (which is again the nuisance parameter). Wager and Athey [13] developed an approach to estimate heterogeneous treatment effects using a nonparametric causal forest, which is an extension of the random forest algorithm. Starting from the potential outcomes framework with the unconfoundedness assumption, this causal forest approach is pointwise consistent for the treatment effect and provides confidence intervals for the true treatment effect estimation. When adopting the random forest algorithm, causal trees and forests can be established [13]. Simulation studies showed promising results of the method in terms of estimating heterogeneous treatment effects.

3. METHOD

3.1. Virtual Learning Environment

We investigated a video recommendation system implemented in Algebra Nation [11], which is a virtual learning environment (VLE) for Algebra. It is used extensively in the state of Florida as integrated with the school districts' system so that every student and teacher can log in with their school username and password. Algebra Nation has a series of instructional videos and formative assessments organized into ten domains (e.g., linear equations, quadratic equations, and exponential functions). These domains are aligned with the state's Algebra standards. Within each domain, there are several topics (ranging from 6 to 12) with a total of 93 topics across all ten domains. Each topic is associated with a video of an Algebra 1 focal question, and there are five versions of each video taught by different tutors who are ethnically diverse and have different instructional approaches. Students can choose their tutor, which launches the corresponding video pertaining to the focal question. Students use Algebra Nation under the guidance of their teachers [28], and access it both in the classroom and at home. Teachers have access to a dashboard showing student actions and assessment results, and can assign individualized homework from the platform. There is a monitored discussion forum where students can ask questions, which are answered by other students and study experts.

Algebra Nation includes formative assessments of students' ability. In total, there are ten pre-test and post-test measures. For each domain, there is a pre-test assessment with five questions and a post-test assessment with ten questions. Students must complete these assessments in a single domain. The questions are selected for each student based on a student model (i.e., two-parameter logistic item response theory [IRT] model) [29], which aims to maximize the amount of information about the ability estimate (i.e., reduce the standard error of measurement) given the

current estimate of the students' ability [30]. The questions' formats are similar to the state's high-stakes Algebra assessment. In addition to these assessments, each topic contains a short three-item check-your understanding (CYU) quiz, which students can voluntarily access at any time. There were 93 CYU quizzes in total and these CYU quizzes were used as one of the inputs for the recommendation system.

The video recommendation system (see Figure 1) integrated into Algebra Nation combines the video recommendation algorithm (presented later as Algorithm 1) and the implementation of a sensor-free measurement of engagement [31, 32]. To measure engagement for interactions with digital learning technologies, we adopted D'Mello, Dieterle [33] advanced, analytic, automated (AAA) approach [21]. Specifically, the recommendation that a student receives is based on 1) his/her score on the three-item CYU quiz, and 2) his/her current engagement score (high or low). The engagement score was derived from the student's interactions with the VLE during the past five minutes before taking the CYU quiz [15].

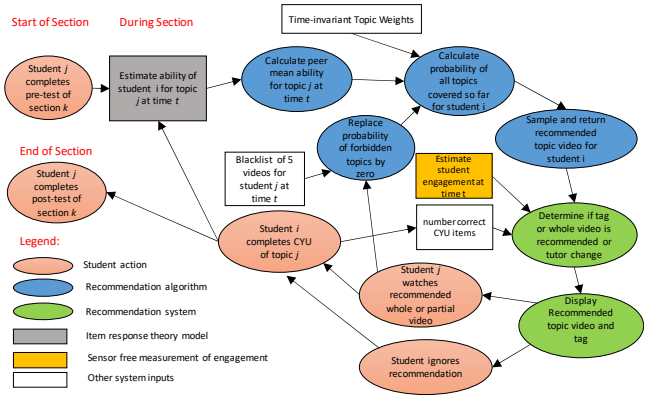

Figure 1. Video recommendation system

Students must complete the CYU quiz before receiving a video recommendation. Once responses to the quiz are submitted, the recommendation appears as a small floating screen at the top of the main screen. Students can either accept or dismiss the recommendation. If dismissed, the recommendation window moves to the bottom right corner of the window, where it remains until the student takes another quiz. To avoid repetition, the system blacklists the last five recommended videos the students watched by assigning a zero probability to them.

CYU quiz score[1]	Engagement score[2]	Probability of recommendation Category		
		C1	C2	C3
0	low	0.9	0.1	0.0
0	high	0.7	0.3	0.0
1	low	0.7	0.3	0.0
1	high	0.5	0.5	0.0
2	low	0.5	0.5	0.0
2	high	0.3	0.7	0.0
3	low	0.3	0.0	0.7
3	high	0.1	0.0	0.9

Table 1. Video Recommendation System (Note: [1] CYU score ranges from 0 to 3; [2] The cutoff value of engagement score is 3, namely, 0 - 3 is low, and >= 3 is high.)

The recommendation system provided three categories of recommendations: Category 1 - view new video as determined by the recommendation algorithm; Category 2 - review segment of current video that is most related to the questions that the student answered incorrectly (by expert review) from a new tutor (but they could keep the same tutor); and Category 3 - view next video in curriculum sequence. The specific category recommended was derived from a set of probabilistic rules developed in conjunction with subject matter experts and learning scientists (Table 1).

The rationale for the rules is: the probability of showing students a recommended video based on the algorithm (i.e., category 1) should be higher for students who score lower in the CYU and have lower engagement. The video recommendation algorithm (Algorithm 1) is specifically used to provide the recommendation for Category 1. For highly engaged students, the system increased the probability of a review of the segment of the video directly related to the questions missed (i.e., category 2), under the assumption that highly engaged students are more likely to pursue reviewing their current work. For students who obtain a perfect score on the quiz, the system introduces a non-zero probability of moving to the next video of the curriculum sequence (i.e., category 3), under the assumption that this student is ready to proceed to new content. This video recommendation algorithm attempts to optimize learning based on student ability estimates from responses to short quizzes (i.e., formative assessments).

Algorithm 1. New Video Recommendation Policy for Student *i*

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

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

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

Compute peer ability estimates (within cluster)

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

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

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

Sample video $\hat{\jmath}\,(t)$ *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 video recommendation algorithm in Algorithm 1 attempts to optimize learning based on student ability estimates from responses to short quizzes (i.e., formative assessments). After

a student responds to a quiz, the current ability estimate a_{ij} is updated with a 2-parameter IRT model [29]. The algorithm is anchored in Vygotsky's theory of Zone of Proximal Development (ZPD) [34], because it attempts to approximate each student's ZPD by calculating the distance between the students' current ability estimate and the average estimate of the student's peers $b_j(t)$. We determined the student peer groups by clustering students into 20 clusters of equal size using quantiles of a Mahalanobis distance from the minimums of two measures of previous student achievement, teacher performance, and school performance (i.e., average student ability in previous CYU, average ability in 10-question quizzes, teacher value-added and school value-added [35] estimated by the state's department of education and available publicly). The topic importance weights w_j were estimated using the Orthogonal Greedy Algorithm (OGA) [36]. An extensive description of this application of OGA is presented in [15].

The algorithm is similar in objectives to content sequencing systems using partially observed Markov decision process (POMDP) [37] and multi-armed bandits (MAB) [38] for ITS. The algorithm's advantage over POMDB and MAB is that in situations where each student takes a small number of quizzes from a large number of available quizzes on different topics, the algorithm is robust to large amounts of missing data [36, 39]. Also, when each student watches a small number of videos from a large set of available videos, the selected subset of videos is the most predictive of an increase in student ability among all possible subsets of videos of the same cardinality [36, 39].

3.2. Data

Data for this study was obtained from a large-scale field experiment across 42 schools in three large school districts in Florida. This field experiment was conducted in the above-mentioned VLE and lasted one semester (January to June 2021). Students in the sample were randomly assigned to receive two types of video recommendations. The treatment group received personalized video recommendations, and the control group received generic recommendations that followed the sequence of algebra topics in the state's Algebra standards. Assignment of condition was blind to students and teachers. The structure of the data was naturally multilevel with students nested within classrooms, and classrooms nested within schools. The sample consisted of 2,995 students who enrolled in Algebra 1 or Algebra 1 Honors courses in the 2020-2021 academic year. These students were taught by 54 teachers in 42 schools. On average, there were 57 students per teacher (SD = 33, min = 4, max = 135) and 75 students per school (SD = 62, min = 13, max = 332).

3.3. Measures

The data consists of three parts, namely, Algebra I standardized test score (EOC score), the VLE variables (See Table 2), and students' background information (Table 3). In total, there were 18 variables with a missing rate at 4.1% of the total number of cases across all variables.

EOC scores. The Algebra I EOC assessment is a computer-based and criterion-referenced high-stakes assessment. The average EOC score in our sample was 513.22 (SD = 29.93, Min = 422, Max = 581).

VLE variables. There were five VLE variables derived from logs of Algebra Nation and related to this randomized controlled experiment (Table 2). Pre-test ability and post-test ability were computed by using a 2-parameter logistic IRT model based on the students' responses to all pre-/post-test items across the ten domains. The pre/post-test ability followed a standard normal distribution with a mean equal to zero and a standard deviation of one. In this study, pre-test and post-test abilities were standardized separately. Mean engagement was an aggregated variable. An engagement score was computed after students completed a single pre-test or a post-test. A higher value indicates that students were more engaged in interacting with the VLE before taking the test. For each student, multiple engagement scores (with a maximum of 20) were stored in the system which corresponded to different pre-/post-test in different domains. Therefore, we compute the average engagement scores across the ten domains for each student. Followed rate measured the rate that the students followed recommendations in all domains [15].

Variables	Mean	Missing	Note
Treatment indicator	0.51	0.00%	Binary (1 = treatment group, 0 = control group)
Post-test ability	0.15	13.22%	Continuous (SD = 0.93, Min = -2.54, Max = 3.27)
Pre-test ability	0.24	14.76%	Continuous (SD = 0.55, Min = -1.51, Max = 2.02)
Mean engagement	2.99	48.30%	Continuous (SD = 0.22, Min = 1, Max = 3.61)
Followed rate	0.14	0.00%	Continuous (SD = 0.19, Min = 0, Max = 1)

Table 2. Descriptive Statistics for VLE Variables

Background (Demographics) variables. Twelve background variables were shown in Table 3. Two binary variables were generated to indicate the three school districts. The students from the third school district served as the reference group, with 0 on both indicators. Likewise, four binary indicators were used to represent the ethnicities. White people served as the reference group, which were 0 on all four indicators.

3.4 Missing-data Imputation

Missing data were imputed using multiple imputations by chained equations using the *mice* package [40] of the *R* statistical software. A total of 10 imputed datasets were created using predictive mean matching because of its robustness to distributional assumptions and ability to handle both continuous and discrete

variables [41]. In the univariate imputation model for each variable, the clustered structure of the data was accounted for by adding fixed effects of teachers and school districts.

3.5 Analyses

3.5.1 Intention to Treat Analysis

To address the first research question, *what are the effects of the recommendation system on learning outcomes*, we estimated the average treatment effect of exposure to the personalized video recommendations. In this analysis, the independent variable was the treatment indicator, and the outcome variables were post-test abilities and EOC scores (analyzed the two outcome variables separately).

Variables	Mean	Missing	Note
Sex	0.54	0.17%	Binary (1 = male, 0 = female)
Course type indicator	0.19	0.00%	Binary (1 = Algebra 1, 0 = Algebra 1 Honors)
Learning mode	0.35	3.54%	Binary (1 = on campus, 0 = distance learning)
Percent distance learning	60.10	31.79%	Continuous (SD = 44, Min = 0, Max = 100)
Free or reduced-price lunch	0.35	31.79%	Binary (1 = receive free or reduced-price lunch, 0 = not receive)
Absent days	3.17	0.23%	Continuous (SD = 6.93, Min = 0, Max = 120)
School district 1	0.20	0.00%	Binary (1 = in the first school district, 0 = not in)
School district 2	0.04	0.00%	Binary (1 = in the second school district, 0 = not in)
Ethnicity (Hispanic)	0.29	0.23%	Binary (1 = Hispanic, 0 = non-Hispanic)
Ethnicity (Black)	0.26	0.23%	Binary (1 = Black, 0 = non-Black)
Ethnicity (Asian)	0.07	0.23%	Binary (1 = Asian, 0 = non-Asian)
Ethnicity (Other)	0.04	0.23%	Binary (1 = Other, 0 = non-Other)

Table 3. Descriptive Statistics for Background Variables

We followed the intention to treat (ITT) analysis framework [21, 22] because it estimated the causal effect of students being randomly assigned to treatment or control conditions without considering the student compliance with recommendations. The ITT average treatment effect estimate is an unbiased causal effect of treatment, but it is conservative because it does not account for the compliance rate [21, 22].

We used two-level multilevel models to estimate the treatment effects with students nested within teachers [42]. Notice that we did not additionally account for school-level nesting because there was only one teacher per school for most of the schools. We did not control for the district-level effect as most of the students were in the same school districts. The multilevel model shown in Equations 1 to 3 was fit with the imputed datasets using the *lme4* package [43] in R (version 4.1.2 [44]).

The final estimates and standard errors were obtained using Rubin's rules [45]. Hedges *g* [46] was used to standardize the average treatment effect estimates. Specifically, the student-level model was:

$$y_{ij} = \beta_{0j} + \delta_j T_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \boldsymbol{N(0, \sigma_1^2)} \qquad (1)$$

where $y_{ij}$is the outcome (EOC or posttest abilities) of student *i* with teacher *j*; β_{0j} is the average score/ability of students with teacher *j*; T_{ij} is the treatment indicator for student *i* with teacher *j*; δ_j is the treatment effect that potentially varies across teachers; and ε_{ij} is the student-level error term.

The teacher-level model was:

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$
$$\delta_j = \delta + \mu_{1j} \qquad (2)$$

where $\gamma_{00}$is the average scores/abilities across all students accounting for the treatment status; $\mu_{0j}$is the random effect associated with teacher *j*; δ is the average treatment effect, and $\mu_{1j}$is the deviance of treatment effect with teacher *j* from the average treatment effect. $\mu_{0j}$and μ_{1j} are assumed to follow a multivariate normal distribution:

$$\begin{bmatrix}\mu_{0j}\\ \mu_{1j}\end{bmatrix} \sim \boldsymbol{MVN}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\tau_{00}} & \boldsymbol{\tau_{01}}\\ \boldsymbol{\tau_{01}} & \boldsymbol{\tau_{11}}\end{pmatrix}\right) \qquad (3)$$

where τ_{00} is the variance of intercept; τ_{11} is the variance of the treatment effect and it is called the treatment-by-site variance [42]; τ_{01} is the covariance between the intercept and treatment effect.

We also tested whether an alternative model that assumes zero treatment-by-site variance is true. When the model comparison (via ANOVA) results in a p-value greater than 0.05, we choose the more parsimonious model (which drops $\delta_j = \delta + \mu_{1j}$, such that$\delta_j = \delta,$and the distributional assumption becomes $\mu_{0j} \sim N(0, \tau_{00})$). A simpler model for the ITT analysis, can be interpreted as no treatment effect variation across teachers. It does not preclude investigating treatment effect heterogeneity [21, 22]. It is still possible to have variables that cause treatment effect heterogeneity in different directions [21, 22].

3.5.2 HTE: Causal Forest

To address the second research question, *is there substantial HTE of the video recommendation system*, we used a two-stage cluster-robust causal forest [13] to estimate the individual conditional average treatment effect (iCATE). Causal forest analysis also allowed us to evaluate the importance of each variable as a predictor. Aligned with the ITT analysis, we computed the HTE on two outcome variables (post-test abilities and EOC

scores) separately. The following twelve VLE variables and background variables were predictors for this analysis: 1) pre-test ability, 2) mean engagement, 3) followed rate, 4) sex, 5) learning mode, 6) percent distance learning, 7) free or reduced-price lunch, 8) absent days, and 9-12) ethnicity indicators.

We adopted a two-stage cluster-robust causal forest for two reasons. Firstly, the cluster-robust causal forest takes the multilevel data structure, where the students were nested in teachers in uneven size, into consideration. Cluster-robust causal forest assumes that teachers have some effect on the student's algebra ability, but without assuming the distribution of the effect. Secondly, the two-stage strategy emphasized the splits with the most important features in low-signal situations [12]. Concretely, the entire data set *S* was randomly divided into two subsamples S_1 and S_2. In the first stage, a pilot causal forest was trained on all variables with subsample S_1 to identify which variables are the most important ones. In the second stage, another forest was trained on the subsample S_2 to estimate iCATE, and only with those variables that were identified as important in the first stage.

In the first stage of the causal forest, variable importance was computed and averaged across the imputed data sets. First of all, a depth-weighted average of the number of splits on the twelve variables were calculated in each imputed data set. Next, we averaged the variable importance for each variable across the ten imputed data sets. Then, we found the median value of the average variable importance for all twelve variables. Last, a variable was considered important if its averaged variable importance value exceeded the median of the averaged variable importance for all variables. The variables which were labeled as important will be used in the second stage of the causal forest, and in the linear mixed-effects multilevel model (Equation 4). In the second stage of the causal forest, the iCATE for each student was computed with each imputed data set and then averaged across the ten imputed data sets. We computed two separated iCATE for each student. One was based on their EOC scores and the other based on their post-test abilities.

Regarding the tuning parameters in the causal forests, we set 1) the number of trees for a single forest at 10000, 2) the minimum size of the leaf node for individual trees at 5, and 3) the number of variables tried for each split at 5. We did not choose those tuning parameters by a cross-validation technique because we need to keep the tuning parameters consistent across the ten imputed data sets. Choosing based on cross-validation may lead to the situation that incomparable forests were built on different imputed data sets. The causal forest was fit with the *grf* package [47] in R (version 4.1.2 [44]).

3.5.3 HTE: Mixed-Effects Multilevel Models

To address the third research question, *what student characteristics predict the HTE*, we built two linear mixed-effects multilevel models predicting iCATE using the important features. The linear mixed-effects multilevel models were built for interpretability. The outcome variable for the first model was iCATE estimated from the causal forest based on post-test scores. Similarly, the outcome for the second model was iCATE estimated from the causal forest based on EOC scores. The predictors for each model were selected separately based on the corresponding causal forest (see the first stage of the causal forest). The selected variables were the independent variables for the model. The linear mixed-effects multilevel model fits with the *lme4* package [43] in R. Concretely, the student-level model was:

$$y_{ij} = \beta_{0j} + \sum_{k=1}^{d} \delta_{jk} X_{ijk} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2}) \qquad (4)$$

where y_{ij} is the iCATE of student *i* with teacher *j*; β_{0j} is the average iCATE of all students with teacher *j*; *k* is the index of the predictors; X_{ijk} is student *i* with teacher *j* on the *k*th predictor (notice that predictors are generated by the causal forest); $\delta_{jk}$is the effect of *k*th predictor with teacher *j*; and ε_{ij} is the student-level error term.

The teacher-level model was:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \mu_{0j} \\ \delta_{jk} &= \delta_k + \mu_{1j} \end{aligned} \qquad (5)$$

where $\gamma_{00}$is the average iCATE across all students regarding all predictors; μ_{0j} is the random effect associated with teacher *j*; δ_k is the average effect of *k*th predictor; and μ_{1j} is the deviance of the effect of *k*th predictor with teacher *j* from the average effect of *k*th predictor.

4. RESULTS

4.1 ITT Results

Regarding the model with post-test abilities as the outcome, the standardized ITT average treatment effect was statistically significant (Hedges g = 0.330, p <.0001), indicating that students who were presented with video recommendations from the system obtained post-test abilities that were on average 0.330 standard deviations higher than students exposed to the control recommendations. The intraclass correlation (ICC) was 0.341, indicating that 34% of the variance of the post-test was due to teachers, thus justifying the use of multilevel models. Moreover, the ANOVA model comparison results show that a zero treatment-by-teacher variance model ($\delta_j = \delta$ in Equation 2) was chosen across nine out of ten imputed data sets. It indicates no treatment effect heterogeneity due to teachers.

Regarding the model with EOC scores as the outcome, the standardized ITT average treatment effect of the recommendation system was also statistically significant (Hedges g = 0.170, p <.0001), showing that students in the treatment group obtained Algebra 1 EOC scores that were higher than the control group by 0.170 standard deviations on average. It was expected that the average treatment effect on the Algebra 1 EOC would be smaller than on the post-test ability, because the former is a distal outcome measured close to the completion of the semester, while the latter is a proximal outcome measured at the end of each section. The ICCs was 0.582 indicating substantial clustering at the teacher

level, which justified the use of multilevel modeling. Moreover, the ANOVA model comparison results show that the model without treatment-by-teacher variance ($\delta_j = \delta$ in Equation 2) was chosen across all ten imputed data sets, again showing that the treatment effect did not vary across teachers.

4.2 HTE Result

Substantial HTE of the video recommendation system was found based on the causal forest analyses. We tested the calibration of the casual forest via two regressors: the forest prediction and the mean forest prediction [12]. We first computed the best linear fit of post-test abilities based on these two regressors. Then, we computed the best linear fit of EOC scores with these two regressors. All four tests (two forest prediction tests and two mean forest prediction tests) rejected the null that there is no heterogeneity. Therefore, we conclude that there was substantial HTE of the video recommendation system.

We investigated the relative importance of the twelve VLE variables and background variables based on the first stage of the causal forest. For these two casual forest models, importance values were normalized such that the sum of the variable importance values across all variables is one, and higher values indicate the corresponding variable was more important in discovering the HTE. Worth noticing that the importance values were scaled separately for the two causal forest models. The results for variable importance from the causal forest were presented in Table 4, and the variables were ordered by importance. It suggested that the most important variables were pre-test, and the followed rate was the second important variable to examine the HTE. It suggested that a large number of trees located the split with pre-test abilities or followed rate. Specifically, seven variables were selected for both models. These variables had a variable importance value higher than the median of the average variable importance for all variables in both causal forests. They were 1) pre-test ability, 2) followed rate, 3) sex, 4) percent distance learning, 5) absent days, 6) free or reduced-price lunch, 7) ethnicity (Hispanic). Moreover, the variable importance result for the EOC scores causal forest model and the post-test abilities causal forest model was almost identical after normalizing the importance values.

Variables	Average Variable Importance (Post-test)	Average Variable Importance (EOC)
Pre-test ability	0.241	0.241
Followed rate	0.228	0.229
Absent days	0.116	0.115
Ethnicity (Hispanic)	0.081	0.081
Percent distance learning	0.075	0.079
Free or reduced-price lunch	0.068	0.068
Sex	0.045	0.045

Table 4. Variable Importance Result

Five student characteristics were found as statistically significant predictors of the heterogeneity in iCATE (See Table 5) based on both linear mixed-effects multilevel models. The results indicate that the conditional mean of the treatment effect was different within the subpopulations for the significant predictors.

Variables	Post-test		EOC score	
Pre-test ability	0.077***	(0.013)	0.071***	(0.009)
Followed rate	0.152***	(0.038)	0.455***	(0.027)
Sex indicator	-0.006	(0.013)	-0.033	(0.009)
Absent days	-0.003	(0.001)	0.002**	(0.001)
Free or reduced-price lunch indicator	0.144***	(0.014)	-0.004	(0.011)
Percent distance learning	0.001***	(0.001)	0.001***	(0.001)
Ethnicity indicator (Hispanic)	0.040**	(0.015)	0.107***	(0.010)

Table 4. Summary for predicting estimated iCATE with post-test ability and EOC score (Note. Standard errors are in parentheses, * indicates p < .05, ** indicates p < .01, *** indicates p < .001)

The differences in estimated iCATE with respect to each significant predictor were presented in Figure 2. The results show that students who had one standard deviation unit higher in pre-test abilities tended to have about 0.07 standard deviation units higher in the conditional mean of treatment effect (estimated from both the EOC score and post-test abilities models). Likewise, students who followed more recommendations tended to have a higher conditional mean of treatment effect. Specifically, one standard deviation unit higher in followed rate was associated with 0.152 standard deviation units higher in iCATE with the post-test ability model and 0.455 standard deviation units higher in iCATE with the EOC scores model. There was small but significant heterogeneity in iCATE regarding learning remotely during Spring 2021. For students who attended school remotely, one standard deviation unit higher in percent distance learning was associated with 0.001 standard deviation units higher in iCATE from both EOC scores and post-test models. Hispanic students were found to have higher iCATE than non-Hispanic students. Receiving free or reduced-price lunch was only significant in predicting iCATE with the post-test abilities model. Receiving free or reduced-price lunch was associated with 0.144 standard deviation units higher in iCATE based on post-test ability. Absent days were only significant in predicting iCATE with the EOC score model. One standard deviation unit higher in absent days was associated with 0.002 standard deviation units higher in iCATE. Sex was not a significant predictor of the heterogeneity in iCATE from both models. There was no iCATE difference between male and female students.

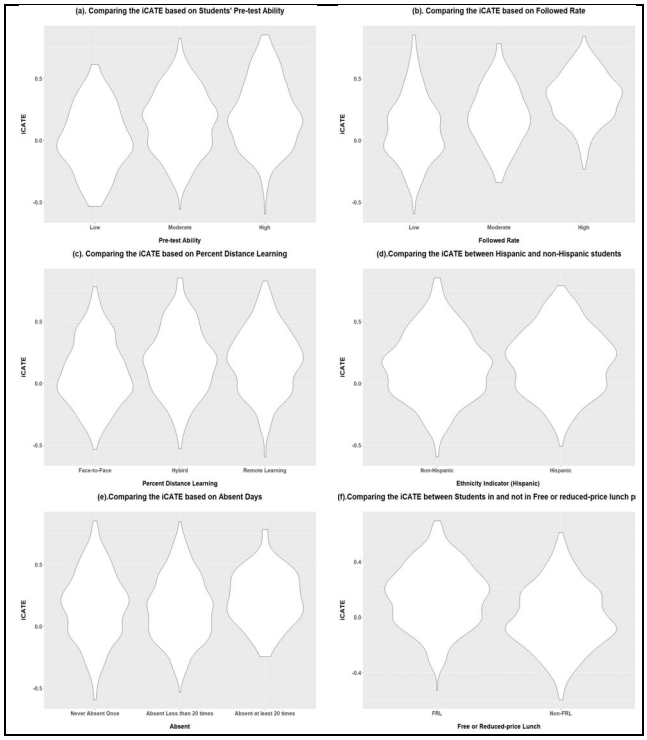

Figure 2. Difference in iCATE by Each Predictor.

Figure 2 shows violin plots of the distributions of iCATE across variables that are related to HTE. The iCATE with the post-test ability and EOC scores had a very similar distribution by each predictor, therefore, only the set of plots based on iCATE with EOC scores is presented. Likewise, the estimated iCATE had a very similar distribution by each predictor among the 10 imputed data sets, therefore, only the set of plots based on the first imputed data set is presented.

5. DISCUSSION

Many ITS have been used extensively prior to the COVID-19 pandemic to improve student achievement, but the pandemic brought new challenges for educational technology in teaching and learning that are just starting to be addressed in the literature [48, 49]. The current study examines a VLE used by students with substantial teacher involvement (e.g., [50, 51]), which shows promise for helping students overcome achievement gaps due to the pandemic as well as other social and economic factors that disadvantage certain subgroups of students.

The previous large-scale field experimental study of this recommendation system [52] was implemented in Spring 2020 during the start of the COVID-19 pandemic. The study lasted for 17 weeks, but was divided into a 6-week period of normal school operations and 11 weeks where schools were closed and instruction was delivered online. The analysis results showed that there was a significant average treatment effect of the recommendation system on the post-test before schools closed, but the effect disappeared during the period of school closure. The authors attributed the disappearance of the effect to the disruption of teachers' teaching strategies and students' learning routines caused by the onset of the pandemic. The sudden transition from physical classrooms to online learning could influence instructors' preparedness and confidence in teaching, which could compromise teachers' role as facilitators to support students' use of technologies for learning [53, 54]. Meanwhile, the sense of disconnection and discordance from such an abrupt transition could negatively influence students' self-regulated learning (SRL) behaviors and strategies in technology-enhanced learning environments [55].

In contrast, the current study was implemented in Spring 2021 during a period when the pandemic was still present, but schools had been re-opened by order of the state government since Fall 2020. However, as shown in the measures section of this paper, only 35.1% of students were attending school campuses in person. Interestingly, the results of Spring 2021 study show significant ITT average treatment effects on both the post-test ability estimates and the Algebra 1 EOC scores. One possible explanation is that teachers and students in the Spring 2021 had more experience with their current learning model (either in-person or online) than in Spring 2020, which facilitated students' SRL [56] with respect to using the recommendation system, and teacher orchestration [57-59] of instruction with the VLE. It is particularly important that the recommendation system had a significant effect on the Algebra 1 EOC scores, because this is a high-stakes test required for high school graduation.

The results of the HTE analysis showed some interesting predictors of the iCATE of students. The relationship between the iCATE and pre-test ability indicates that students with higher previous achievement benefit more from the recommendation system. This could be due to students with higher previous achievement having better SRL skills, an important factor to influence students' learning in VLE [60], which allowed them to use the recommendation system better. In the meantime, students with higher prior achievements might have higher self-efficacy, which could help them better adapt to and engage in VLE supported by learning analytics [61]. However, future research including measures of SRL skills and self-efficacy of students would be needed to investigate this hypothesis. Unsurprisingly, individuals with a higher followed rate of videos of the recommendation system were associated with higher iCATE, because the followed rate is a proxy for the dosage of intervention, and previous meta-analyses of ITS have shown that the duration of exposure to an ITS explains the size of the effect [9, 18].

The results showed an association between iCATE on the post-test and free-and-reduced lunch eligibility of the students, indicating that economically disadvantaged students benefited more from the recommendation system. This suggests that the recommendation system could contribute to narrowing knowledge gaps between disadvantaged and advantaged students, implying a possibility of improving equity in educational intelligent systems.

However, this result was not replicated for the iCATE on the EOC. The results also show that Hispanic students had higher iCATEs on both post-test and the EOC than non-Hispanic students, which may be associated with cross-cultural differences, differences in teacher orchestration of technology in classrooms, or school-level contextual differences. In the study of Olaussen and Bråten [62], the researchers discussed how SRL strategies could differ from a cross-cultural perspective (e.g., ethnicity, East vs. West). In an empirical study, Wan et al. [63] found differences in SRL strategies could have various impacts on students' learning, with students having higher social and goal orientation skills achieving better declarative knowledge acquisition. Additional research is needed to understand how disadvantaged students and Hispanic students may have benefited more from the video recommendation system.

We also found that the percent distance learning measure was positively associated with iCATE. This is interesting because the VLE where the recommendation system was implemented was not specifically designed for distance learning. In fact, previous studies [52] have shown strategies such as showing videos to the entire class, and creating centers to watch specific videos and work on quizzes are common strategies with this VLE. Therefore, the increase in iCATE for students with a higher percent of distance learning may be related either to teachers changing their orchestration strategies for the VLE, or students acquiring a higher level of independence on the use of the VLE than they would have in an in-person classroom. It was unexpected that the number of absent days was positively associated with the iCATE of the EOC. In a previous survey of the same population of teachers during the pandemic [64], teachers reported that they found it difficult to reliably control attendance for students in distance learning, which may be related to the finding in the current study.

In terms of absent days, which was a significant predictor for iCATE with the EOC score model. This finding echoed the previous research that teachers assigned the learning content in this VLE as make-up for students who were unable to present in the classroom learning [65]. It seemed that the more days those students were absent from school, the more heavily they relied on this VLE to prepare for Algebra 1 EOC assessment during Spring 2021.

5.1 Limitations

The HTE analysis shows a nuanced picture of the effects of the recommendation system, but it does not allow an explanation of the effects. It is important to understand the specific mechanisms by which some subgroups of students (e.g. Hispanic students and free-reduced lunch eligible students) benefited more from the recommendation system. Qualitative studies may be used to probe the specific mechanisms.

Because the recommendation system studied here is used within a VLE whose use by students is orchestrated directly by teachers, it is possible that teachers may encourage or discourage watching recommended videos differently in in-person versus distance instruction. The current study is limited in that it did not include teacher variables in the prediction of HTE, such as survey variables indicating when and how teachers used the VLE with their students. Although a previous survey study of teacher use of this VLE exists [28], the study only addressed pre-pandemic teacher use of the VLE. Also, the HTE analysis did not include school contextual variables such as the percentage of minority students, expenditures per pupil, and the percentage of students in poverty.

The current study serves as the first investigation of HTE with one recently developed method (i.e. causal forests). However, a competition of HTE identification methods [65], including causal forests, demonstrated that there can be substantial differences across methods. For future research, the current data will be evaluated with a few HTE methods to understand the stability of the results. Future research will also address the complier average causal effect (CACE) [66] of students watching recommended videos when offered. The CACE is higher than the ITT estimate because it accounts for the frequency that videos are watched.

6. CONCLUSION

The majority of studies about personalized learning technologies such as ITS and other content sequencing systems have focused on supplements to regular classroom instruction, and studies that evaluate personalized learning in VLE used as part of regular classroom instruction are rare [67]. The current study demonstrated that a video recommendation system for a VLE whose use is primarily for classroom instruction had significant positive effects on student learning, adding to evidence obtained by a previous large-scale evaluation of the same system [52]. It also showed substantial HTE, which had significant relationships to previous ability, mode of learning, the rate that students following recommendations, ethnicity, and poverty. These relationships deserve examination in future research to understand their nature and allow proper targeting of subgroups given the context of the classroom and school. This is critical because, when personalization is used in schools without considering the economic and political context, it may face resistance from students and teachers [68], and actually disadvantage certain groups of students and increase inequality [69].

Finally, it is important to distinguish between personalized learning that is customizable by the learner and customized by the system, the teacher, or school administrations [69]. The VLE examined in the current study is an example of this distinction, as it is adopted by the school districts as the main curriculum, but teachers have considerable flexibility in how and when to use it in the classroom, and students have the flexibility of how much to engage with it outside of the classroom. It is critical to consider the distribution of agency between students, teachers, and school administrators, as systems that over-prescribe learning experiences using extensive student data may violate students' rights to privacy and have adverse ethical implications [69]. For

example, content sequencing systems based on the previous achievement may provide more educational content to high achieving students than low achieving students [15]. As research on personalized learning technologies progresses, it is important to create standards for personalization that address both validity, fairness, and ethics. Standard exists for educational testing [69] that can serve as inspiration for technology-mediated personalized learning standards.

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 the views of the Institute of Education Sciences or the U.S. Department of Education.

REFERENCES

1. Bernacki, M.L., M.J. Greene, and N.G. Lobczowski, *A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)?* Educational Psychology Review, 2021. **33**(4): p. 1675-1715.
2. Plass, J.L. and S. Pawar, *Toward a taxonomy of adaptivity for learning.* Journal of Research on Technology in Education, 2020. **52**(3): p. 275-300.
3. Zhang, L., J.D. Basham, and S. Yang, *Understanding the implementation of personalized learning: A research synthesis.* Educational Research Review, 2020: p. 100339.
4. Doroudi, S., V. Aleven, and E. Brunskill, *Where's the reward?* International Journal of Artificial Intelligence in Education, 2019. **29**(4): p. 568-620.
5. Kizilcec, R.F. and H. Lee *Algorithmic fairness in education*. 2021.
6. Sies, A., K. Demyttenaere, and I. Van Mechelen, *Studying treatment-effect heterogeneity in precision medicine through induced subgroups.* J Biopharm Stat, 2019. **29**(3): p. 491-507.
7. Lipkovich, I., A. Dmitrienko, and S. B. R. D' Agostino, *Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials.* Stat Med, 2017. **36**(1): p. 136-196.
8. Kulik, J.A. and J.D. Fletcher, *Effectiveness of Intelligent Tutoring Systems.* Review of Educational Research, 2016. **86**(1): p. 42-78.
9. Ma, W., et al., *Intelligent tutoring systems and learning outcomes: A meta-analysis.* Journal of Educational Psychology, 2014. **106**(4): p. 901-918.
10. Steenbergen-Hu, S. and H. Cooper, *A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning.* Journal of Educational Psychology, 2014. **106**(2): p. 331-347.
11. Lastinger Center for Learning and University of Florida. *Algebra Nation*. 2019 [cited 2019 9/20/2019]; Available from: http://lastingercenter.com/portfolio/algebra-nation-2/.
12. Athey, S. and S. Wager, *Estimating Treatment Effects with Causal Forests: An Application.* Observational Studies, 2019. **5**(2): p. 37-51.
13. Wager, S. and S. Athey, *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests.* Journal of the American Statistical Association, 2018. **113**(523): p. 1228-1242.
14. Rubin, D.B., *Estimating causal effects of treatments in randomized and nonrandomized studies.* Journal of Educational Psychology, 1974. **66**: p. 688-701.
15. Leite, W.L., et al. *A novel video recommendation system for algebra: An effectiveness evaluation study*. in *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22)*. 2022. Online, USA: ACM, New York, NY, USA.
16. VanLehn, K., *The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems.* Educational Psychologist, 2011. **46**(4): p. 197-221.
17. Xu, Z., et al., *The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis.* British Journal of Educational Technology, 2019. **50**(6): p. 3119-3137.
18. Steenbergen-Hu, S. and H. Cooper, *A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning.* Journal of Educational Psychology, 2013. **105**(4): p. 970-987.
19. Dong, N., et al., *Power analyses for moderator effects with (non)random slopes in cluster randomized trials.* Methodology, 2021. **17**(2): p. 92-110.
20. Raudenbush, S.W. and H.S. Bloom, *Learning About and From a Distribution of Program Impacts Using Multisite Trials.* American Journal of Evaluation, 2015. **36**(4): p. 475-499.
21. Ding, P., A. Feller, and L. Miratrix, *Randomization inference for treatment effect variation.* Journal of the Royal Statistical Society: Series B: Statistical Methodology, 2016. **78**(3): p. 655-671.
22. Ding, P., A. Feller, and L. Miratrix, *Decomposing Treatment Effect Variation.* Journal of the American Statistical Association, 2018. **114**(525): p. 304-317.
23. Athey, S. and G. Imbens, *Recursive partitioning for heterogeneous causal effects.* Proc Natl Acad Sci U S A, 2016. **113**(27): p. 7353-60.
24. Anoke, S.C., S.L. Normand, and C.M. Zigler, *Approaches to treatment effect heterogeneity in the presence of confounding.* Stat Med, 2019. **38**(15): p. 2797-2815.
25. *ACM computing surveys*. 1971, New York, N.Y.: Association for Computing Machinery.
26. Carnegie, N., V. Dorie, and J.L. Hill, *Examining treatment effect heterogeneity using BART.* Observational Studies, 2019. **5**(2): p. 52-70.
27. Hill, J.L., *Bayesian Nonparametric Modeling for Causal Inference.* Journal of Computational and Graphical Statistics, 2011. **20**(1): p. 217-240.
28. Mitten, C., Z.K. Collier, and W.L. Leite, *Online Resources for Mathematics: Exploring the Relationship between Teacher Use and Student Performance.* Investigations in Mathematics Learning, 2021: p. 1-18.
29. Lord, F. and M. Novick, *Statistical theories of mental test scores*. 1968, Reading, MA: Addison-Wesley.
30. Embretson, S.E. and S.P. Reise, *Item response theory for psychologists*. 2000, Mahwah, NJ: Lawrence Erlbaum Associates.
31. Hutt, S., J. Grafsgaard, and S.K. D'Mello, *Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire School Year*, in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2019).* 2019, ACM: New York.
32. Jensen, E., S. Hutt, and S.K. D'Mello, *Generalizability of Sensor-Free Affect Detection Models in a Longitudinal Dataset of Tens of Thousands of Students*, in *The 12th International Conference on Educational Data Mining*, M. Desmarais, et al., Editors. 2019. p. 324–329.
33. D'Mello, S.K., E. Dieterle, and A.L. Duckworth, *Advanced, Analytic, Automated (AAA) Measurement of Engagement during Learning.* Educational Psychologist, 2017. **52**(2): p. 104-123.
34. Vygotsky, L.S., *Mind in society*. 1978, Cambridge: Harvard University Press.
35. Raudenbush, S.W., *What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?* Journal of Educational and Behavioral Statistics, 2004. **29**(1): p. 121-129.
36. Ing, C.K. and T.L. Lai, *A stepwise regression method and consistent model selection for high-dimensional sparse linear models.* Statistica Sinica, 2011: p. 1473-1513.
37. Chen, Y., et al., *Recommendation System for Adaptive Learning.* Appl Psychol Meas, 2018. **42**(1): p. 24-41.
38. Mu, T., et al., *Combining adaptivity with progression ordering for intelligent tutoring systems*, in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 2018. p. 1-4.
39. Hsu, H.-L., C.-K. Ing, and T.L. Lai, *Analysis of High-Dimensional Regression Models Using Orthogonal Greedy Algorithms*, in *Handbook of Big Data Analytics*. 2018. p. 263-283.
40. van Buuren, S. and K. Groothuis-Oudshoorn, *mice: multivariate imputation by chained equations in R.* Journal of Statistical Software, 2011. **45**(3): p. 1-67.
41. Kleinke, K., *Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching.* Journal of Educational and Behavioral Statistics, 2017. **42**(4): p. 371-404.
42. Raudenbush, S.W. and X. Liu, *Statistical power and optimal design for multisite randomized trials.* Psychological Methods, 2000. **5**(2): p. 199–213.
43. Bates, D.M., et al., *Fitting Linear Mixed-Effects Models Using lme4.* Journal of Statistical Software, 2015. **67**(1): p. 1-48.

44. R Development Core Team, *R: A language and environment for statistical computing*. 2022, R Foundation for Statistical Computing: Vienna, Austria.
45. Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys*. 1987, New York: Wiley.
46. Hedges, L.V., *Effect sizes in cluster-randomized designs.* Journal of Educational and Behavioral Statistics, 2007. **32**(4): p. 341–370.
47. Tibshirani, J., et al., *grf: Generalized Random Forests. R package version 2.0.2*. 2021.
48. Yan, L., et al., *Students' experience of online learning during the COVID-19 pandemic: A province-wide survey study.* Br J Educ Technol, 2021.
49. Mahmood, S., *Instructional strategies for online teaching in COVID-19 pandemic.* Human Behavior and Emerging Technologies, 2021. **3**(1): p. 199-203.
50. Holstein, K., *Designing Real-time Teacher Augmentation to Combine Strengths of Human and AI Instruction*, in *School of Computer Science*. 2019, Carnegie Mellon University: Pittsburgh, Pennsylvania.
51. Segedy, J., B. Sulcer, and G. Biswas, *Are ILEs ready for the classroom? Bringing teachers into the feedback loop.* Intelligent Tutoring Systems, 2010: p. 405-407.
52. Authors, *Reference masked for blind review.* 2022.
53. Carrillo, C. and M.A. Flores, *COVID-19 and teacher education: a literature review of online teaching and learning practices* European Journal of Teacher Education, 2020. **43**(4): p. 466-487.
54. Hensley, L.C., R. Iaconelli, and C.A. Wolters, *"This weird time we're in": How a sudden change to remote education impacted college students' self-regulated learning.* Journal of Research on Technology in Education, 2021: p. 1-16.
55. Fong, C.J., *Academic motivation in a pandemic context: a conceptual review of prominent theories and an integrative model.* Educational Psychology, 2022: p. 1-19.
56. Broadbent, J. and W.L. Poon, *Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review.* The Internet and Higher Education, 2015. **27**: p. 1-13.
57. Dillenbourg, P., *Design for classroom orchestration.* Computers & Education, 2013. **69**: p. 485-492.
58. Drijvers, P., et al., *The teacher and the tool: instrumental orchestrations in the technology-rich mathematics classroom.* Educational Studies in Mathematics, 2010. **75**(2): p. 213-234.
59. Prieto, L.P., et al., *Orchestrating technology enhanced learning: a literature review and a conceptual framework.* International Journal of Technology Enhanced Learning, 2011. **3**(6): p. 583-598.
60. Wang, C.H., D.M. Shannon, and M.E. Ross, *Students' characteristics, self-regulated learning, technology self-efficacy, and course outcomes in online learning.* Distance Education, 2013. **34**(3): p. 302-323.
61. Alemayehu, L., & Chen, H. L., *The influence of motivation on learning engagement: the mediating role of learning self-efficacy and self-monitoring in online learning environments.* Interactive Learning Environments, 2021: p. 1-14.
62. Olaussen, B.S. and I. Bråten, *Students' Use of Strategies for Self-regulated Learning: cross-cultural perspectives.* Scandinavian journal of educational research, 1999. **43**(4): p. 409-432.
63. Wan, Z., D. Compeau, and N. Haggerty, *The effects of self-regulated learning processes on e-learning outcomes in organizational settings.* Journal of Management Information Systems, 2012. **29**(1): p. 307-340.
64. Leite, W.L., et al. *Teacher Strategies to Use Virtual Learning Environments to Facilitate Algebra Learning During School Closures*. 2022. DOI: 10.17605/OSF.IO/P2JRB.
65. Carvalho, C., et al., *Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge.* Observational Studies, 2019. **5**(2): p. 21-35.
66. Schochet, P.Z. and H.S. Chiang, *Estimation and Identification of the Complier Average Causal Effect Parameter in Education RCTs.* Journal of Educational and Behavioral Statistics, 2011. **36**(3): p. 307-345.
67. Major, L., G.A. Francis, and M. Tsapali, *The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis.* British Journal of Educational Technology, 2021. **52**(5): p. 1935-1964.
68. Holmes, W., et al., *Technology-enhanced personalised learning: Untangling the evidence*. 2018, Robert Bosch Stiftung GmbH: Stuttgart, Germany.
69. FitzGerald, E., et al., *A literature synthesis of personalised technology-enhanced learning: what works and why.* Research in Learning Technology, 2018. **26**(2095): p. 1-16.