Investigating the Usage Patterns of Algebra Nation Tutoring Platform

Sahba Akhavan Niaki
Department of Statistics, University of Florida
sahbaakn@ufl.edu

Clint P. George[*]
Informatics Institute, University of Florida
clintpg@ufl.edu

George Michailidis[†]
Department of Statistics & Informatics Institute,
University of Florida
gmichail@ufl.edu

Carole R. Beal
School of Teaching and Learning, College of Education,
University of Florida
crbeal@coe.ufl.edu

ABSTRACT

We study the usage of a self-guided online tutoring platform called Algebra Nation, which is widely by middle school and high school students who take the End-of-Course Algebra I exam at the end of the school year. This article aims to study how the platform contributes to increasing students' exam scores by examining users' logs over a three year period. The platform under consideration was used by more than 36,000 students in the first year, to nearly 67,000 by the third year, thus enabling us to examine how usage patterns evolved and influenced students' performance at scale. We first identify which Algebra Nation usage factors in conjunction with math overall preparation and socioeconomic factors contribute to the students' exam performance. Subsequently, we investigate the effect of increased teacher familiarity level with the Algebra Nation on students' scores across different grades through mediation analysis. The results show that the indirect effect of teacher's familiarity with the platform through increasing student's usage dosage is more significant in higher grades.

CCS CONCEPTS

• **Applied computing → Computer-assisted instruction**;

KEYWORDS

Math education, Online tutoring platform, Hierarchical linear models, Mediation analysis

ACM Reference Format:

Sahba Akhavan Niaki, Clint P. George, George Michailidis, and Carole R. Beal. 2019. Investigating the Usage Patterns of Algebra Nation Tutoring Platform. In *The 9th International Learning Analytics & Knowledge Conference (LAK19), March 4–8, 2019, Tempe, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3303772.3303788

1 INTRODUCTION

The use of online tutoring platforms as a component in the math curriculum or as a supplemental technology-based educational tool to improve students' math achievement has grown considerably over the last decades. Given the proliferation and availability of these platforms, it is of great interest to assess their effectiveness in improving student achievement. Impact is inevitably influenced by implementation, referring to the frequency and context in which a platform is used. In the case of online platforms that have been designed by education experts in light of a theory of learning, better or more consistent implementation is usually associated with better end-user outcomes [3].

Efficacy studies considering implementation quality are more informative and can provide useful insights to developers for further improvements. Implementation data on educational platforms are usually collected through teacher and student surveys, interviews, self reports, classroom visit observations, or system log data. Interviews, surveys and self reports may be biased and require procedures to ensure validity and reliability of the reported data [8], while classroom visits are only feasible for a subset of the classes, especially in case of large scale platforms. On the other hand, monitoring implementation through analytics of system logs is objective, has lower cost and faster turnaround time [4].

Implementation dosage, referring to a measure that quantifies the implementation activities (e.g. amount of time using the platform, number of sections completed, number of answered questions) can be used to compute implementation fidelity. Fidelity refers to the degree to which an intervention is implemented as intended by developers and has been used in the literature to measure the implementation outcome of technology-based interventions with predefined usage guidelines [6, 8, 11, 17]. Professional development (PD) training is typically provided to familiarize teachers with the scope of the intervention and to provide support to help them use the technology in a manner that is consistent with what the intervention developers expect. For example, teachers implementing Cognitive Tutor Algebra I (CTAI) as part of a large-scale efficacy evaluation received at least 12 hours of PD training prior to platform implementation as well as two school visits from the developer staff per year.

Alternatively, an online tutoring platform may not have been designed under a model of ideal use. In this case, implementation data can be used to determine the possible impact of the platform

[*] Currently affiliated with Indian Institute of Technology Goa
[†] Corresponding author

LAK19, March 4–8, 2019, Tempe, AZ, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6256-6/19/03…$15.00
https://doi.org/10.1145/3303772.3303788

as well as the best ways of using it. For example, in the implementation study of Khan Academy, a self-guided video-based tutoring platform, [13] used student dosage levels (corresponding to the time spent on Khan Academy platform and the number of completed problem sets) and found a positive relation between dosage level and students' achievement test score for two research sites. The two sites involving 850 and 242 students, respectively, were selected to represent different facets of public school districts. Site 1 was a high-achieving suburban school district serving middle to upper middle class student population, while the second site was a public charter schools serving a low-income urban and predominantly Latino community.

In reality, actual technology integration in classrooms frequently seems to deviate from the ideal expectations even in the case of carefully-conducted Randomized Controlled Trial evaluation studies with PD for teachers [4, 8, 14]. However, studies indicate that increasing teachers' comfort and familiarity level with a technology-based platform would increase the effectiveness of its use. For interventions with usage guidelines, the familiarity is usually increased by PD training [12, 17], whereas for platforms with no usage guidelines it may be increased by using the platform over multiple years. In the Cognitive Tutor Algebra I study, even though teachers' fidelity level was still low by the second year of use, they had adjusted their instruction based on traditional and new teaching techniques which resulted in better platform effectiveness [8]. In the Khan Academy implementation study [13], teachers' attitude toward the platform positively improved even though the number of minutes their students used it actually declined from the first year to the second. These initial findings suggest that as teachers gain experience, they may select the components of an intervention that they view as most likely to benefit their students, which may or may not be consistent with developers' expectations. However, conclusions are limited because relatively few studies of technology-based interventions have been conducted over more than a single school year.

The present case concerns a specific online tutoring platform called Algebra Nation (AN) that is an online video-based tutoring program launched in 2014-2015 school year to help prepare Florida students for the End-of-Course (EoC) test in basic algebra. The platform was similar to Khan Academy in being self-guided and easy to use on an ad hoc basis, without extensive PD. Informal adoption was rapid, and within three years there was at least some sporadic use in every district in the state. The resulting data set enables to address the following two issues: first, investigate the relation between the users dosage levels recorded through system log files and final student outcome (scores on the EoC). Further, the availability of multiple years of teacher log data enables us to locate instances of teachers who were using AN for the first time and compare them to similar teachers who had used the platform the previous years, and to determine the effect of teacher familiarity and exposure with AN in student usage and outcomes.

2 DATA

Participants. The data used in this study consist of records of more than 36,000 students in the 2014-15 school year, 39,000 students in the 2015-16 year and nearly 67,000 students in the 2016-17 year in grades 7 to 9. Students were considered to be users of the platform if they logged in five times or more during a school year. Table 1 provides more details on the number of participants in each grade for each school year. Platform usage of teachers of participating students is also recorded.

Demographic data. The collected student data contain information on personal student characteristics including gender, race, ethnicity, and free and reduced lunch status. In addition to student level data, we use publicly available data on [State] schools and districts such as percentages of economically disadvantaged students and minorities in the schools and overall rankings of schools and districts. Other collected student data are math scores at the beginning of each academic year (i.e. pre-score, a proxy for math preparedness) and at the End-of-Course Algebra 1 exam (i.e. EoC score). The EoC score is the main student outcome in this study.

Math pre-scores available for the 2015-2017 academic years and for the 2014-2015 academic year were based on two different state assessment standards tests, the Florida Standards Assessments (FSA) and the Florida Comprehensive Assessment Test 2.0 (FCAT). The EoC scores, ranging from 425 to 575, were based on FSA math tests, but they were scored on a different scale compared to pre-scores in both FSA and FCAT math exams. Thus, using pre-scores as the baseline knowledge at the beginning of the year requires an appropriate adjustment between pre-scores and EoC test scores. To that end, we use a quantile normalization method [1] with the distribution of EoC scores set as the reference distribution for adjusting the scales of the pre-scores.

A comparison between grade 7-9 students' EoC scores showed an ordering as indicated by the mean scores shown in the Table 1, with those in lower grades exhibiting a better performance. Further, pre-scores in each grade have different achievement level threshold scores.

Hence, we undertake a separate analysis for each grade and also consider only first time EoC test takers After removing missing values and extreme outliers we get 25,458 students in grade 7, 65,780 students in grade 8 and 53,477 students in grade 9 cumulatively for the three academic years under consideration. Table 1 depicts summary statistics on student, school, and district characteristics for different grades in each academic year.

Usage metrics. Data are recorded when students log in to the platform, and when they navigate to specific areas. These include viewing videos recorded by tutors who explain concepts and demonstrate how to solve problems, practice tests designed to resemble the state-administered EoC, and a monitored discussion area where students can pose questions to peers and volunteer tutors. Use of the discussion area was low, making logins, video views and attempting to solve practice problems the primary indicators of usage dosage. Table 2 provides a summary of usage features for each grade and for each academic year. The variables video, tys, session, and wall post indicate numbers of watched videos, test-yourself set answered, total sessions, and posted comments on the wall by students respectively.

Boxplots in Figure 1 show the weekly AN action counts of 2015-2016 students in grades 8 and 9. The overall trend shows that for both grades, students start using the platform in the last weeks close to the final exams, but this usage is more consistent and on

	Grade 7			Grade 8			Grade 9		
Academic year	2014-15	2015-16	2016-17	2014-15	2015-16	2016-17	2014-15	2015-16	2016-17
Counts	(7,133)	(8, 157)	(10,168)	(16,696)	(19,035)	(30,049)	(12,701)	(14,067)	(26,709)
Categorical (%)									
Gender									
F	52.31	52.29	51.13	55.68	55.73	54.20	52.49	51.99	50.83
FRL									
Y	41.22	43.58	43.28	49.87	54.82	50.77	69.51	72.55	66.32
Race									
W	73.20	71.99	69.51	72.35	72.80	71.84	62.37	63.20	65.53
B	13.71	14.23	15.68	17.93	17.46	17.19	31.45	30.19	26.59
P	0.46	0.48	0.56	0.47	0.40	0.39	0.28	0.28	0.58
I	2.09	2.40	2.15	4.01	3.87	4.35	3.58	4.01	4.52
A	10.54	10.90	12.10	5.24	5.46	6.24	2.32	2.31	2.78
Ethnicity									
Y	32.69	33.92	27.74	34.87	33.88	29.64	36.18	40.00	36.67
Quantitative (mean)									
Pre-score*	254.76	355.76	357.37	251.82	349.85	351.73	235.23	334.85	341.08
District Points (%)	60.92	56.40	58.77	61.05	56.13	58.89	60.24	55.57	58.69
School Points (%)	65.00	60.69	61.52	62.92	58.43	61.37	59.51	52.04	55.24
School EDS (%)	58.06	61.13	59.09	57.33	62.38	58.23	57.81	63.18	58.72
School MS (%)	64.63	66.37	62.09	62.66	62.19	57.83	64.41	66.41	61.77

Table 1: Comparing various features for grade 8 and grade 9 students in three consecutive years. *Pre-scores are based on FCAT-tests for academic year 2014-15 and FSA-tests for the next two years.

average higher for students in grade 8 with better backrgound. We observed similar trend for different action types for students in different grades and academic years.

3 METHOD

3.1 Identifying important usage features

We used hierarchical linear models (HLMs, [16]) to identify usage features that are significantly associated with students' EoC scores. The clustered structure of education data has made HLMs a widely used modeling framework in their analysis. Figure 2 shows how student usage differs across grades, demographic variables, and district and school rankings. The single usage measure used in the plots corresponds to the first Principal Component based on student video views, test-yourself questions completed, and number of logins which explains more than 91% of the variation for these three usage features. Schools and districts are grouped into three equally sized groups based on the available district and school points with group 3 containing schools and districts exhibiting strong performance. The shifts in the distribution of this aggregate usage measure in the plots shows that as expected, better math background, coming from a more advantageous socioeconomic environment and attending more highly ranked schools positively impacts the utilization of the AN platform. Thus, we fit separate HLMs for each grade, while controlling for other demographic variables. Equation set 1 shows the two-level random intercept model employed in this study. The first level contains student variables, while the second level contains "class" related variables. Here, a "class" is defined as a group of students who attend the same school and are instructed by the same teacher. We also explored using HLMs with more than two levels and also adding random intercepts, but did not provide better data fits.

$$\begin{aligned}
\mathrm{EoC}_{ij} &= \beta_{0j} + \beta_{1j}\,\mathrm{video}_{ij} + \beta_{2j}\,\mathrm{tys}_{ij} + \beta_{3j}\,\mathrm{login}_{ij} \qquad (1)\\
&\quad + \beta_{4j}\,\mathit{wall\_post}_{ij} + \sum_{k=5}^{K} \beta_{kj} S_{kij} + \epsilon_{ij},\\
\epsilon_{ij} &\sim \mathbb{N}(0, \sigma^2)\\
\beta_{0j} &= \gamma_{00} + \sum_{j=1}^{T} \gamma_{0j}\,T_j + u_{0j},\ u_{0j} \sim \mathbb{N}(0, \tau_{00})\\
\beta_{kj} &= \gamma_{k0}, k = 1, 2, \ldots, K.
\end{aligned}$$

where EoC_{ij} refers to EoC score of student i in class j. Variables *video*, *tys*, and *login* are the main usage variables that indicate the number of videos watched, test-yourself questions completed, and total logins into AN, respectively. There are many other actions recorded in the platform as well, for example the number of searched videos, video pauses, wall posts, wall pages searched,

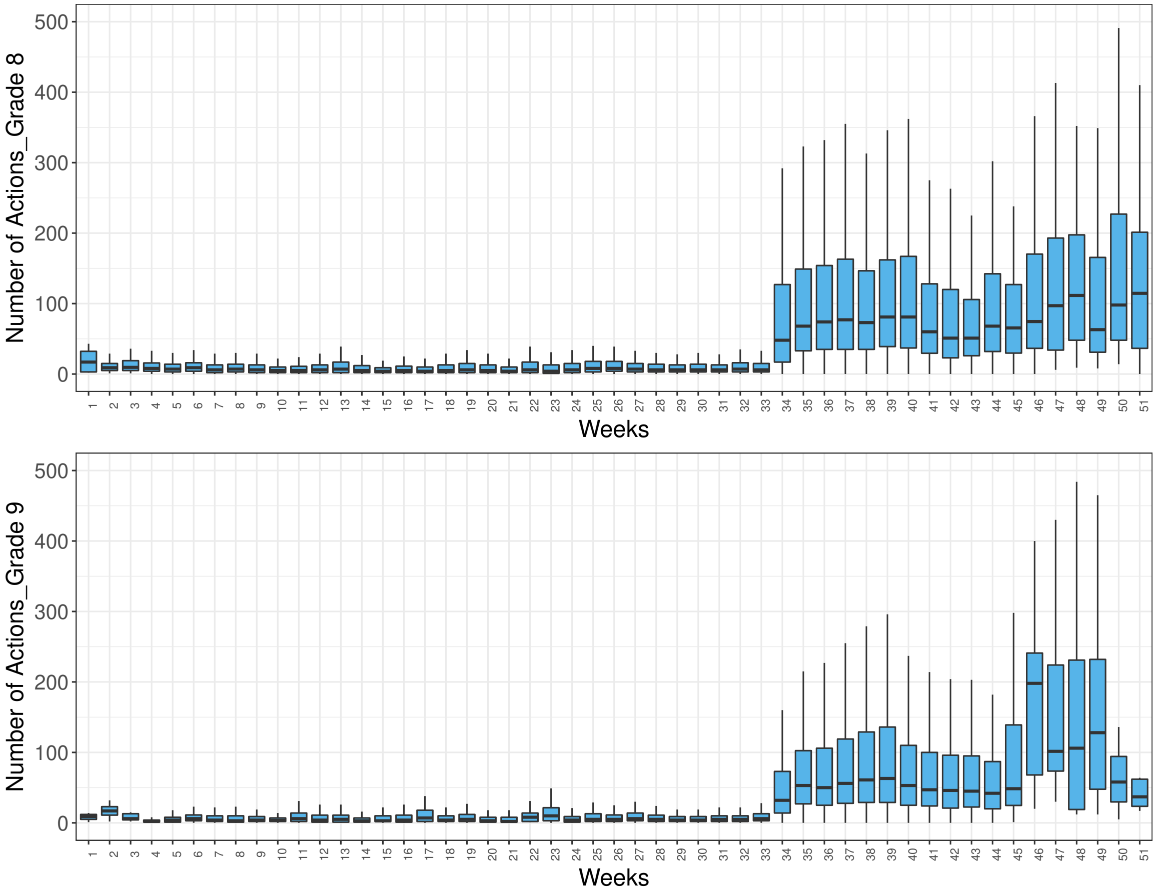

Figure 1: Weekly AN actions for grade 8 and 9 students in year 2015-2016.

	2014-2015				2015-2016				2016-2017			
		Quantiles				Quantiles				Quantiles		
Variables	Mean	25%	50%	75%	Mean	25%	50%	75%	Mean	25%	50%	75%
Grade 7												
video	36.57	8	22	52	48.23	13	33	70	59.84	19	46	87
tys	13.06	0	4	18	12.11	0	5	16	6.74	0	2	9
session	23.46	10	18	29	25.43	11	20	33	12.95	8	11	16
wall post	8.61	0	0	1	5.88	0	0	0	5.44	0	0	0
Grade 8												
video	32.89	7	20	46	36.25	9	24	49	49.30	14	35	70
tys	10.71	0	2	13	9.42	0	2	12	6.07	0	1	8
session	21.23	9	16	27	21.07	9	15	27	11.69	7	10	15
wall post	3.98	0	0	0	3.33	0	0	0	3.08	0	0	0
Grade 9												
video	24.20	5	14	32	28.21	8	19	38	39.98	11	28	57
tys	5.78	0	1	6	4.86	0	1	5	3.40	0	0	4
session	16.28	7	12	20	17.40	8	13	22	10.85	6	9	14
wall post	1.98	0	0	0	1.13	0	0	0	0.91	0	0	0

Table 2: Summary of student usage variables recorded in AN.

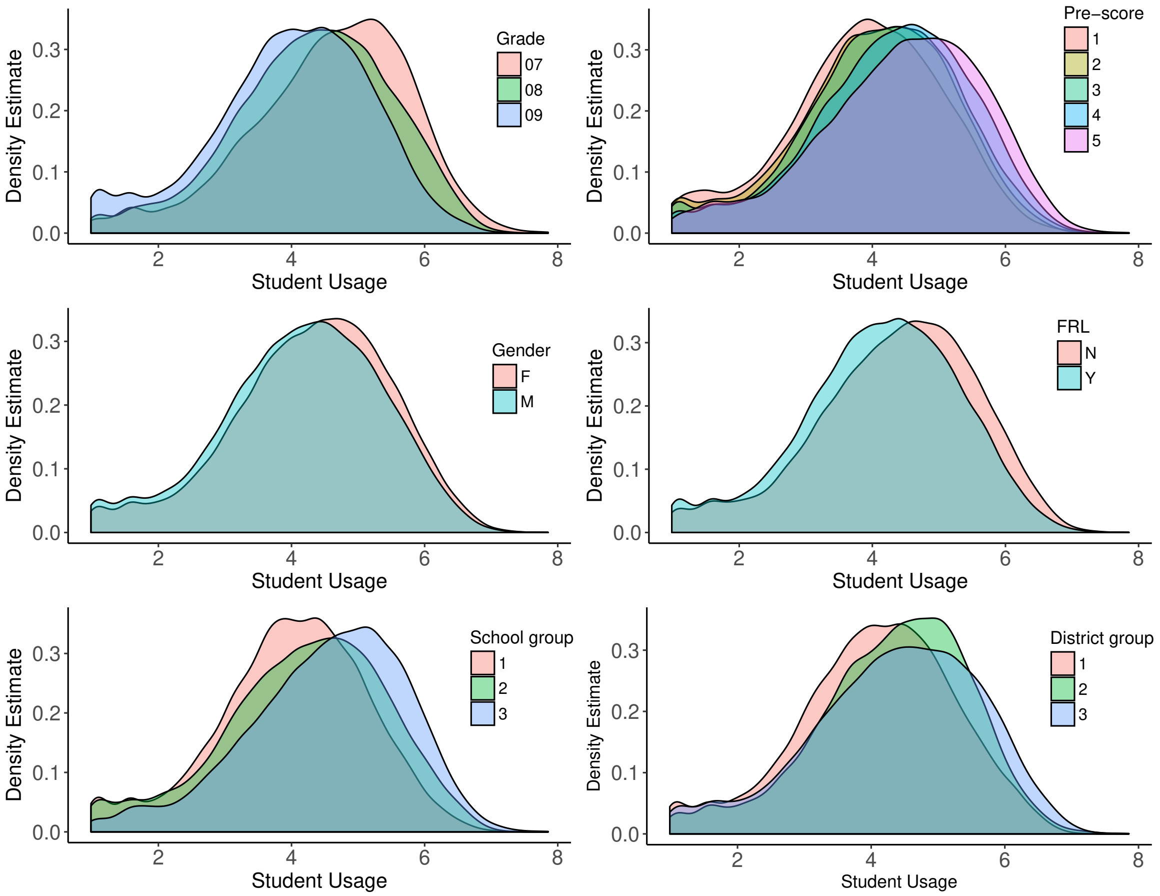

Figure 2: Distributions of student AN usage based on different demographic variables. School group 3 and district group 3 identify the best ranked schools and districts.

leader board loads, and review of incorrect test questions, which are only recorded in recent implementations of the platform. However, these fields are very sparse with most of them having zero value. S_{kij} refers to the k-th control variable for student i in class j and correspond to pre-score and other demographic variables introduced in Section 2. T_j refers to the class level variables for class j.

3.2 Identifying teachers' impact on AN utilization

We investigate the effect of teachers' previous AN-familiarity (i.e. familiarity with the AN platform) on students' usage of AN. To that end, we employed the framework of *mediation analysis* which in its simplest form involves three variables: the independent predictor variable, the mediator variable, and the outcome variable. A variable is called a mediator, when it accounts for all or part of the relation between the independent predictor and the outcome variable. The three paths in Figure 3 show the three essential conditions for a variable to act as a mediator [2]: 1) The independent predictor variable is a significant variable in modeling the mediator variable (a as the coefficient of independent variable in this model) , 2) The mediator is a significant variable in modeling the outcome variable (b as the coefficient of mediator), 3) After controlling for the independent and mediator variables, the significance of the relation between independent variable and the outcome variable is reduced (c' as the coefficient of independent variable after controlling for the mediator).

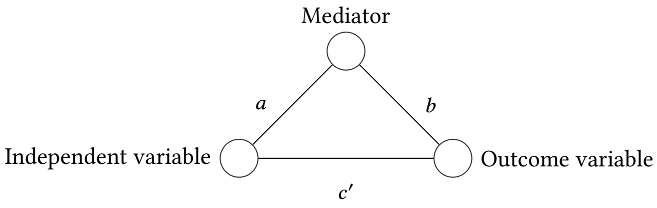

Figure 3: The basic mediation model

Mediation analysis can be easily extended to multilevel structured data. The multilevel mediation analysis follows the same framework as in the simple mediation analysis, with the addition of the following restrictions: the outcome variable should be at the lowest level of the hierarchy and each variable in the mediation chain can affect variables at the same level or lower [10].

In this part of our study, the independent variable of interest is the binary variable indicator of teachers' familiarity level in using AN, which is a second level variable. The other two affected variables, student usage and EoC_{ij} score, are measured at the lowest level of hierarchy, namely the student level. We hypothesize that student AN usage can act as a mediator variable between teachers' AN familiarity and student EoC scores being the outcome variable. In other words, part of the effect of having a highly familiar AN teacher on students' EoC score is through teacher's positive effect on increasing student usage in AN. Translating our hypothesis to a mediation analysis framework, we need to test the significant of the following three relations:

(1) teacher's AN familiarity level → student EoC score
(2) teacher's AN familiarity level → student AN usage
(3) teacher's AN familiarity level + student AN usage → student EoC score

The indirect effect of the independent variable (teacher's AN familiarity level) is transmitted to the dependent variable (student EoC score) through the mediator (student AN usage). Equations 2–4 translate the above three relations into the hierarchical linear modeling framework:

$$\begin{aligned}
\text{EoC}_{ij} &= \beta_{(1)0j} + \beta_{(1)1j}\, S_{ij} + \epsilon_{(1)ij} &(2)\\
\beta_{(1)0j} &= \gamma_{(1)00} + c\,\text{TF}_j + u_{(1)0j} &\\
M_{ij} &= \beta_{(2)0j} + \beta_{(2)1j}\, S_{ij} + \epsilon_{(2)ij} &(3)\\
\beta_{(2)0j} &= \gamma_{(2)00} + a\,\text{TF}_j + u_{(2)0j} &\\
\text{EoC}_{ij} &= \beta_{(3)0j} + \beta_{(3)1j}\, S_{ij} + b\, M_{ij} + \epsilon_{(3)ij} &(4)\\
\beta_{(3)0j} &= \gamma_{(3)00} + c'\,\text{TF}_j + u_{(3)0j} &
\end{aligned}$$

where i and j refer to student and teacher, S_{ij} is the vector of student variables (Pre-score, Gender, Race, FRL, etc.), TF is a binary variable indicator of teacher's AN familiarity level, and M_{ij} is the mediator variable, students AN usage.

Two popular approaches for estimating the mediation effect are the Linear Structural Equation Model (LSEM) [2] and Average Causal Mediation Effect (ACME) [7]. The LSEM-based estimate can be interpreted as an ACME estimator under a sequential ignorability assumption [7]. The LSEM approach in the multilevel framework gives two point estimates for the mediation effect: $a\,b$ and $c - c'$. In the presence of multiple mediators $c - c'$ estimates only the total mediation effect and cannot break the mediation effect of each mediator. In a simulation study, [9] show that the discrepancy between the two estimators gets larger, when the number of groups and group sizes are small and the true mediation effect is large.

In the single-level mediation analysis, the total effect of an independent variable can be decomposed into an indirect effect via the mediator and a direct effect. This decomposition does not hold for hierarchical models because of the multilevel structure of the models. Another important difference between mediation analysis in single versus multilevel data is that the two point estimators for mediation effect using the LSEM approach in the single-level framework give equal estimates, but the reformulation of these estimators in the multilevel framework do not give algebraically equivalent estimators. Further, there are multiple suggestions in the literature on how to compute standard error estimates of the mediation effect (first-order Taylor expansion/multivariate delta method for estimating the standard error of the product of two random variables [18], second-order Taylor expansion, exact variance under the condition of independence [5], or unbiased variance [5]). Based on the simulation study by [9], the unbiased estimates [5] would have the least bias in the presence of a large number of groups (more than 200) and large group sizes (20-30). In this paper we use the estimator ab and the estimate of unbiased standard error of $\sqrt{s_a^2\, b^2 + s_b^2\, a^2 - s_a^2\, s_b^2}$.

4 RESULTS

4.1 Identifying important usage features

Equation set 1 is used for modeling students EoC scores for each grade and each academic year. As previously mentioned, usage variables for years 2014-2015 and 2015-2016 were the number of OTP logins, video views, test-yourself sets completed, and wall posts. The wall post feature is rarely used by students as shown in Table 2 and is consistently not statistically significant; hence it is removed from any further consideration. Other usage variables recorded in the most recent version of the OTP implementation in 2016-2017 such as leader board check, watching solution videos, or number of karma awarded exhibited sparse patterns (an overabundance of zero values), and thus did not contribute to the HLMs examined. Therefore, we focused on the following three usage variables in our final analysis for each of the three academic years: logins, video views, and test-yourself sets completed. Results in Table 3 show the coefficient estimates for each of the HLM model. As an example, after controlling for all demographic variables for grade 8 students in year 2015-2016, an additional video view, test-yourself set, and login would result in an average increase of 0.036, 0.056, and 0.036 points in the EoC score. Based on Table 2, the average usage of this group of students is approximately 37 videos, 10 test-yourself sets, and 21 logins. The coefficient estimates are consistently higher for test-yourself sets completed compared to video watches, making an additional test set more beneficial in increasing the EoC score on average. However, the test-yourself sets are lengthy tests covering more material than what is presented in any single video and that is the primary reason that students tend to take these tests close to the EoC exam date.

4.2 Investigating the role of teachers in enhancing students usage

The first step in identifying the role of teacher familiarity level with the OTP platform on students' EoC score is to come up with a measure of their familiarity. We use the recorded teacher log data in the platform for this purpose. Similar to students, OTP records teachers' activities including number of logins, video views, test-yourself sets completed, and wall posts. In this part of the study, we only focus on the last year of the OTP implementation, namely 2016-2017, and investigate the teacher OTP familiarity effect by filtering the data set to only include those teachers who started using the OTP platform in 2016 and teachers who had high OTP familiarity level by using the platform extensively in previous years.

	Grade 7			Grade 8			Grade 9		
Fixed effects	Estimate	± S. E.	Pr (>|t|)	Estimate	± S. E.	Pr (>|t|)	Estimate	± S. E.	Pr (>|t|)
2014 − 2015									
video	0.039	± 0.007	0.000	0.045	± 0.005	0.000	0.026	± 0.008	0.001
tys	0.043	± 0.012	0.000	0.051	± 0.009	0.000	0.119	± 0.016	0.000
login	0.077	± 0.015	0.000	0.061	± 0.011	0.000	0.054	± 0.018	0.002
2015 − 2016									
video	0.032	± 0.004	0.000	0.036	± 0.004	0.000	0.021	± 0.006	0.001
tys	0.048	± 0.011	0.000	0.056	± 0.007	0.000	0.056	± 0.015	0.000
login	0.035	± 0.010	0.001	0.036	± 0.008	0.000	0.041	± 0.015	0.007
2016 − 2017									
video	0.024	± 0.003	0.000	0.039	± 0.002	0.000	0.037	± 0.004	0.000
tys	0.085	± 0.015	0.000	0.121	± 0.009	0.000	0.130	± 0.015	0.000
login	0.110	± 0.030	0.000	0.116	± 0.019	0.000	0.117	± 0.026	0.000

Table 3: Fixed effect estimates of the HLMs for usage variables in different years.

To measure teachers' OTP familiarity level in the previous two years, we used Principal Component Analysis (PCA) on teachers' activities in each year. The resulting first Principal Component captures more than 82% of variation in teachers' OTP usage in the two years and is used in our analysis to identify highly OTP-familiar teachers.

We define the binary teacher level variable TF as the indicator of a teacher's OTP-familiarity level. Highly OTP-familiar teachers take value 1 and newly joined teachers take value 0 for this variable. Comparing student and school characteristics for students of the two groups of teachers, we find significant differences between students' demographic information, pre-test scores, and school rankings, with OTP-familiar teachers having students with better average pre-test scores and higher socioeconomic status. To reduce the potential bias in estimating the teacher familiarity effect and the mediation effect of OTP student usage, we further create comparable groups by filtering the data set to consider only the new and OTP-familiar teachers who both teach in the same schools. By fixing the schools, we get students who are comparable across the previously mentioned characteristics for both groups of teachers. Table 4 shows summary statistics of the resulting data set. Comparing the average number of OTP students per teacher for both groups of teachers shows that on average highly OTP-familiar teachers in all grades have more OTP using students. This is the first observation which suggests that teacher's familiarity with OTP is positively correlated with students using OTP as part of their algebra learning.

Mediation analysis to test for indirect effect of teacher's familiarity with OTP on students' EoC scores through OTP student usage was first performed using total action counts of students in the platform as the single mediator. We further broke down the total action counts into more interpretable actions including video views, logins, and test-yourself sets completed to use them as the mediators in a multiple mediation framework. We should note that none of the action variables causally precede other actions, even for the login variable, as students do not log out after each action or they may choose to remain logged in after a session if they are using their own devices for accessing OTP. Table 5 reports the coefficient estimates and p-values of interest for each three steps of the mediation analysis based on [2] method and the average mediation effects estimated based on two methods by [2] (adjusted based on hierarchical model with multiple mediators) and [7] (i.e. ACME) for all usage variables except for variable test-yourself sets completed. Variables for the number of total actions, video views, and logins are approximately Normally distributed after the log transformation, but variable test-yourself sets completed are too sparse to be considered Normal. So the Normality assumption of linear regression in Step 2 of the mediation framework by [2] is violated in this case. Thus, we used a binary variable for the test-yourself mediator which is 1 if the student has tried at least one test-yourself set and 0 otherwise. For this binary variable, the Step 2 model would be based on a logistic mixed model. The usual approach of estimating the mediation effect does not work in this particular case as a result of the nonlinearity in Step 2. So, we only used the model-independent approach of [7] to estimate average mediation for the test-yourself set variable.

The first step of the mediation analysis (Equation 2) is independent of any mediator variable. The coefficient estimate and p-values of TF variable in Step 1 in Table 5 shows that the total effect of teachers familiarity with OTP on increasing students' EoC scores is higher for students in lower grades. Single mediation analysis for student's total action counts shows that the positive effect of having highly OTP-familiar teachers in higher grades is mediated through increasing students overall engagement in the platform. Multiple mediation analysis results using more detailed student actions further indicate that the positive effect of the high OTP-familiar teacher is mediated through increasing the number of times grade 8 students log in to OTP and encouraging grade 9 students to try the test-yourself sets more.

As implementors of the platform in classrooms, teachers guide students on how to use the OTP, by suggesting which videos to watch, what tests to take, or how often to use the platform. Overall, having OTP-familiar teachers can positively affect scores of younger students in grade 7 who have better background in math, more than students in grades 8 and 9. As teachers gain familiarity

	Grade 7		Grade 8		Grade 9	
	TF0	TF1	TF0	TF1	TF0	TF1
Student counts	493	625	2,676	3,095	3,524	3,253
Teacher counts	28	27	82	69	115	82
Average number of students per teacher	18	24	33	45	31	40
Gender(%)						
F	49.49	49.92	54.22	52.15	50.09	51.89
FRL(%)						
Y	49.90	54.88	51.38	51.37	72.02	69.97
Race(%)						
W	70.99	68.96	73.24	72.08	65.18	65.75
B	13.79	9.28	15.02	15.54	26.87	27.48
P	0.20	0.64	0.37	0.38	0.40	0.34
I	2.43	3.84	4.86	4.85	5.11	4.12
A	12.58	17.28	6.50	7.14	2.44	2.31
Ethnicity(%)						
Y	26.98	35.04	31.69	28.17	40.75	39.99
Pre-score(mean)	357.98	358.14	350.62	351.21	335.54	338.19

Table 4: Comparing various features of groups of students in grade 7 to 9 having highly OTP-familiar teachers (TF1) with teachers who are relatively new to OTP (TF0). Only the schools with both kind of teachers are included.

by using the platform and possibly incorporating them into their teaching, they will gradually detect the best way students of different background levels can take advantage of OTP. This familiarity indirectly increase students' scores through encouraging them to take more tests in grade 9 and log in to the platform more often in grade 8. Younger students in grade 7 seems to be more motivated in using the platform, as they have more logins, video views, and test-yourself sets completed (see Table 2) regardless of the teacher's familiarity level (TF not significant in all Step 2s for grade 7).

5 DISCUSSION AND CONCLUSION

In this paper, we studied the implementation of an online tutoring platform called Algebra Nation (AN) by examining usage patterns by students over a three consecutive years period, with nearly 69, 000 participants in the last year. As an important step in evaluating the AN, we identified important components of the platform that contribute significantly to students' EoC scores. By identifying the most and least effective components of the platform, one is able to provide the AN developers with insights on the components that are working as designed and on those that require revision or a better implementation in future upgrades. Findings show that higher number of test-yourself sets and videos are both significantly correlated with higher scores, but the overall dosage levels show low level of platform utilization especially amongst grade 9 students.

The results indicate that the test-yourself feature is an important, yet not frequently used, feature of the AN. Further investigation showed that the length of the test and coverage of a lot of material in it are hindering factors for students to take the tests. This brings the opportunity of modifying this feature, by designing shorter tests that could be directly tied to specific videos. Further, usage levels of many other features in the platform including the wall post section, the leader board section, etc., are low and not used as intended/designed and as a consequence do not contribute in improving students' EoC scores.

In addition, we investigated the role of teachers in increasing usage levels. Specifically, we examined whether teachers with greater familiarity with the platform increase usage levels of their students and hence positively contribute to the EoC scores. The results indicate that especially for grade 9, the effect of teachers' experience with the platform on EoC scores was mediated through increasing the usage in the test-yourself feature. On the other hand, the usage level of grade 7 students was higher than grades 8 and 9, irrespective of teachers experience and familiarity with the platform; these students showed the higher level of self motivation in using the platform. Overall, as indicated by [8], teachers tend to adjust their instruction over time as they get more familiar with an online platform, and the mediation analysis results render support to the fact that teachers tend to affect students in different grades with different math background levels differently.

The study took advantage of the availability of usage data for a very large number of students across three consecutive years to gain insights into how the platform was being used. In particular, the analyses revealed that some platform features that had been included on pedagogical grounds were not being utilized as intended. For example, the discussion wall provided students with the opportunity to seek individual assistance but was actually ignored by many students. The wall required continuous monitoring by adult instructors, so learning that it was rarely used prompted an assessment of whether it was the best use of expensive instructor resources. Similarly, finding that students often did not complete the full-length practice tests led to the re-design of this feature to make it shorter and more targeted to specific skills.

	Grade 7		Grade 8		Grade 9	
	Estimate	p-value	Estimate	p-value	Estimate	p-value
Step 1:						
TF	3.372	* 0.058	2.787	* 0.006	1.738	* 0.072
Step 2-total actions:						
TF	0.112	0.560	0.103	0.542	0.273	* 0.047
Step 3-total actions:						
total_actions	0.433	0.345	2.278	* 0.000	1.954	* 0.000
TF	3.327	* 0.061	2.580	* 0.016	1.230	0.226
ACME (average)	0.049	0.732	0.229	0.554	0.512	* 0.060
LSEM ($\hat{a}\,\hat{b}$)	0.049	0.251	0.235	0.541	0.534	* 0.049
Step 2-tys (0-1):						
TF	0.801	0.318	0.555	0.426	1.767	* 0.003
Step 2-video (log):						
TF	0.274	0.305	0.206	0.203	−0.009	0.526
Step 2-logins (log):						
TF	0.136	0.133	0.144	* 0.014	0.114	* 0.019
Step 3:						
tys	4.350	* 0.000	1.882	* 0.001	2.853	* 0.000
video (log)	0.083	0.879	0.805	* 0.000	0.839	* 0.001
login (log)	1.027	0.434	2.985	* 0.000	1.190	* 0.091
TF	2.704	0.113	2.053	* 0.064	1.037	0.321
tys ACME	0.358	0.346	0.087	0.426	0.438	* 0.000
video ACME	0.015	0.980	0.169	0.182	−0.078	0.510
video LSEM ($\hat{a}\,\hat{b}$)	0.022	0.614	0.166	0.211	−0.078	0.514
login ACME	0.133	0.552	0.433	* 0.010	0.134	0.110
login LSEM ($\hat{a}\,\hat{b}$)	0.139	0.391	0.428	* 0.023	0.136	0.143

Table 5: Results of the single mediation analysis using total actions and multiple mediation analysis with video, test-yourself set, and login counts as mediators of interest.

The analysis also revealed how usage changed as changed as teachers gained experience with the platform, and demonstrated that this appeared to be especially important for those most at risk for failure in algebra. More specifically, students who take Algebra 1 in Grade 9 have relatively weak math skills, and are also disproportionally male, of lower socioeconomic status, and attending schools in lower-achieving districts compared to students who take algebra in Grades 7 or 8. Here, Grade 9 students received a greater benefit from the platform if their teachers had used it the previous year. Although the outcome measure was performance on a single state-administered achievement test that may not have reflected meaningful learning of algebra concepts, performance on the exam had significant consequence: students who did not earn a passing score would not qualify for high school graduation. In 2017, the overall passing rate was 42%; those who failed were disproportionately African American and Hispanic, students with disabilities and students who were English Learners [15]. The exam thus functions as a barrier to inclusion and access to the benefits associated with completing high school. Thus, using analytics to identify ways the AN can be improved (e.g., redesigning the test yourself feature) and how it was most effectively implemented by teachers, especially for students at risk, should help to promote student success.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences through grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and are not necessarily the views of the Institute or the U.S. Dept. of Education. We would like to thank Study Edge and the Florida Department of Education for making the aligned and anonymized data used in the analysis available for the research.

REFERENCES

[1] Dhammika Amaratunga and Javier Cabrera. 2001. Analysis of Data From Viral DNA Microchips. *J. Amer. Statist. Assoc.* 96 (2001), 1161–1170. Issue 456. https://doi.org/10.1198/016214501753381814

[2] Reuben M. Barron and David A. Kenny. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51, 6 (1986), 1173–1182.

[3] Andrew U Dane and Barry H. Schneider. 1998. Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review* 18, 1 (1998), 23–45.

[4] Mingyu Feng, Jeremy Roschelle, Neil Heffernan, Janet Fairman, and Robert Murphy. 2014. Implementation of an Intelligent Tutoring System for Online Homework Support in an Efficacy Trial. In *Intelligent Tutoring Systems*. Springer International Publishing, Cham, 561–566. https://doi.org/10.1007/978-3-319-07221-0_71

[5] Leo A. Goodman. 2012. On the exact variance of products. *J. Amer. Statist. Assoc.* 55, 292 (2012), 708–713. https://doi.org/10.1080/01621459.1960.10483369

[6] Heewon Lee Gray, Isobel R. Contento, and Pamela A. Koch. 2015. Linking implementation process to intervention outcomes in a middle school obesity prevention curriculum: Choice, Control and Change. *Health Education Research* 30 (2015), 248–261. Issue 2. https://doi.org/10.1093/her/cyv005

[7] Kosuke Imai, Luke Keele, and Teppei Yamamoto. 2010. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statist. Sci.* 25, 1 (2010), 51–71. https://doi.org/10.1214/10-STS321

[8] Rita Karam, John F. Pane, Beth Ann Griffin, Abby Robyn, Andrea Phillips, and Lindsay Daugherty. 2017. Examining the implementation of technology-based blended algebra I curriculum at scale. *Education Technology Research and Development* 65 (2017), 399–425. Issue 2. https://doi.org/10.1007/s11423-016-9498-6

[9] Jennifer L. Krull and David P. MacKinnon. 1999. Multilevel Mediation Modeling in Group-based Intervention Studies. *Evaluation Review* 23, 4 (1999), 418–444. https://doi.org/10.1177/0193841X9902300404

[10] Jennifer L. Krull and David P. MacKinnon. 2001. Multilevel Modeling of Individual and Group Level Mediated Effects. *Multivariate Behavioral Research* 36 (2001), 249–277. Issue 2. https://doi.org/10.1207/S15327906MBR3602_06

[11] Heewon Lee, Isobel R. Contento, and Pamela Koch. 2013. Using a Systematic Conceptual Model for a Process Evaluation of a Middle School Obesity Risk-Reduction Nutrition Curriculum Intervention: Choice, Control and Change. *Journal of Nutrition Education and Behavior* 45 (2013), 126–136. Issue 2. https://doi.org/10.1016/j.jneb.2012.07.002

[12] Lindsay Clare Matsumura, Helen E. Garnier, and Lauren B. Resnick. 2010. Implementing Literacy Coaching: The Role of School Social Resources. *Educational Evaluation and Policy Analysis* 32 (2010), 249–272. Issue 2. https://doi.org/10.3102/0162373710363743

[13] R. Murphy, L. Gallagher, A. Krumm, J. Mislevy, and A. Hafter. 2014. Research on the Use of Khan Academy in Schools.

[14] Sahba Akhavan Niaki, Clint P. George, George Michailidis, and Carole R. Beal. 2017. The Impact of an Online Tutoring Program for Algebra Readiness on Mathematics Achievements; Results of a Randomized Experiment. Technical report.

[15] Florida Dept. of Education. 2017. Florida Standards Assessments: English Language Arts and Mathematics. http://www.fldoe.org/core/fileparse.php/5668/urlt/89FSAPacket.pdf

[16] S. W. Raudenbush and A. S. Bryk. 2000. . Sage Publictions, Thousand Oaks, CA.

[17] Louise Ann Rohrbach, Melissa Gunning, Ping Sun, and Steve Sussman. 2010. The Project Towards No Drug Abuse (TND) dissemination trial: implementation fidelity and immediate outcomes. *Prevention Science* 11 (2010), 77–88. Issue 1. https://doi.org/10.1007/s11121-009-0151-z

[18] Michael E. Sobel. 1982. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology* 13 (1982), 290–312.