Visualizing Student Problem-Solving Data

Yvonne Chen
evechen@uw.edu

Eleanor O'Rourke
eorourke@cs.washington.edu

ABSTRACT
Abstract goes here

INTRODUCTION
Intro here

RELATED WORK
A large body of research has explored methods of integrating technology into in-person classrooms and using technology to deliver education online. We review this research focusing on technology-delivered curriculums, systems that expose student data for teachers, and adaptive learning environments.

Exposing Student Data for Teachers
Education research shows that teacher behavior has a strong impact on student achievement [**?**, **?**, **?**, **?**], and that teachers can benefit from the availability of real-time student data [**?**, **?**, **?**]. For example, Koile found that when an instructor was given access to student problem solutions through tablet-based technology in real-time, the instructor devoted 75% of class time responding to student misunderstandings [**?**]. With access to real-time data, research suggests that instructors can intervene during a lesson when students are confused [**?**], alter the pace or content of instruction based on student engagement [**?**], immediately identify and assist students who are struggling [**?**], and choose topics of focus based on aggregates of student responses [**?**].

A number of technologies have been developed for exposing student data for teachers. One technology that is often used in lecture-based classes is the "student response system" or "clicker," which is used to poll students on multiple-choice questions during class [**?**, **?**]. A similar application designed for small classes is Plickers [**?**]. With the Plickers smartphone app, the teacher can scan the classroom while students hold up QR codes identifying a multiple-choice answer [**?**]. Researchers have also explored methods of providing instructors with access to student data outside of instruction time to monitor longer term academic progress [**?**, **?**]. Kim et. al. developed a system for compiling student responses to MOOC exercise problems, which teachers reported were useful for capturing student thought processes, identifying misconceptions, and engaging students with content [**?**].

In this work, we explore methods of exposing rich problem-solving data to elementary school teachers in real-time. Through a longitudinal study, we explore how teacher behavior is impacted by exposing information about student misconnects and progress through curriculum material.

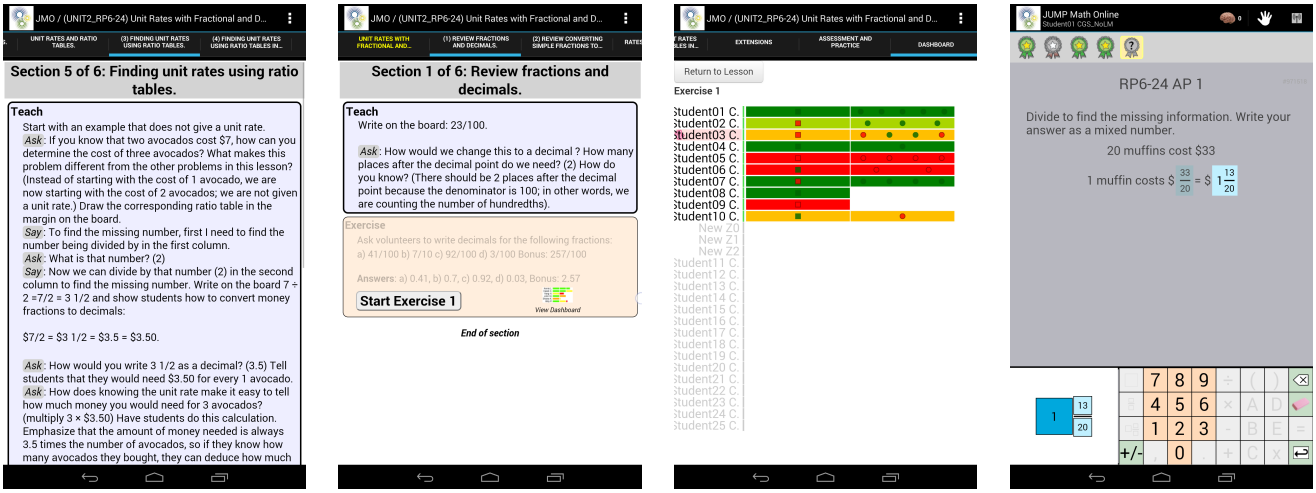
METHOD
We conducted a ten-week study of the software in the fall of 2014 to learn the strengths and weaknesses of the student and teacher software, and to observe the process of integrating a technology-delivered curriculum into real-world classrooms.

Participants
We recruited sixth-grade math teachers from four urban public schools in a northwestern city to participate in our trial. At each school, one teacher (or pair of co-teachers) participated. The teachers taught between two and four sixth-grade math classes, all of which were included in the trial. A total of 219 students participated in 11 classes. The schools represent a diverse set of communities; one serves a high-income neighborhood, two serve low-income neighborhoods, and one serves a low-income neighborhood but is co-managed by both the state and a non-profit organization. More information about each school is provided in Table **??**. We discuss school demographics in detail below. Parents, teachers and school principals gave informed consent for student participation.

Procedure
Our study had two conditions: a control condition using the original paper version of the JUMP Math curriculum, and an experimental condition using the tablet-based version. At each school, one class was assigned to the control condition and the remaining classes were assigned to the experimental condition. Since this was a long-term intervention that would strongly impact the dynamics of the participating classes, we allowed teachers to select which classes would receive the tablet-based curriculum. At all four schools, teachers chose to use the tablet-based curriculum with their weaker classes.

None of the teachers had used the JUMP Math curriculum previously, so at the beginning of the trial we conducted a 3-hour curriculum training session for the teachers. During this training, a JUMP Math representative described the curriculum philosophy and walked through example lessons. We also gave teachers separate training to learn the features of the tablet-based software. During the first week of school, we also conducted a training session for students in each class.

We designed the trial to last for ten weeks and cover four units of the JUMP Math curriculum. While we originally planned to create tablet-based versions of all four units, we were only able to complete units 2 and 4 due to time constraints. As a result, all classes began with a paper version of Unit 1, which was followed by either a paper or tablet version of Unit 2, a paper version of Unit 3, and either a paper or tablet version of



(a) (b) (c) (d)

Figure 1. Screenshots of the Teacher and Student software. Figure (a) shows the teacher lesson view. Figure (b) shows a prompt for the teacher to "Start Exercise 1" on all student tablets. Figure (c) shows the teacher Dashboard that displays student problem-solving progress. Figure (d) shows the student problem-solving interface, with a calculator input interface at the bottom of the screen and badges at the top that indicate problem correctness.

School	Teaching	# Classes	# Students	Free Lunch	ELL	Math Score	Reading Score	AI/AN	A/PI	B	H	W	2+
A	Co-Taught	4	72	10.0%	1.9%	88.2%	90.4%	0.4%	10.8%	2.3%	7.1%	72.1%	7.3%
B	One Teacher	2	39	53.0%	7.0%	55.0%	69.5%	0.6%	16.1%	12.1%	18.9%	37.8%	14.4%
C	One Teacher	3	63	63.6%	16.4%	56.7%	70.3%	1.0%	27.1%	41.3%	10.5%	11.2%	8.9%
D	One Teacher	2	45	50.8%	5.7%	64.0%	64.0%	2.5%	14.9%	14.6%	23.4%	31.4%	13.4%

ELL = English Language Learner, **AI/AN** = American Indian / Alaskan Native, **A/PI** = Asian / Pacific Islander, **B** = Black, **H** = Hispanic, **W** = White, **2+** = Two or More Races.

Table 1. Data for the four schools.

School	Quality	Org.	Support	Prep.	Rapport	Challenges
A	4	4	4.33	4.33	4.33	1.67
B	1.33	1	1.5	2	2.33	3.67
C	3	3	1.67	3.25	4.25	4.75
D	4.75	5	4.5	5	4.5	2.75

Table 2. Results for the Likert-scale questions of our survey of Enlearn classroom observers. For each school, we collected data on the overall quality of the teacher, the level of class organization, the level of support provided by school administrators, the level of teacher preparedness, the rapport between teachers and students, and the level of challenge presented by students (issues at home, English language learners, etc).

Unit 4. Students were given pre- and post-tests designed by JUMP before and after each unit.

Measures

Our study has a mixed design with both qualitative and quantitative measures. Our qualitative measures captured high-level feedback on the Enlearn Platform from participating teachers as well as the impressions of staff from the non-profit who spent extensive time observing teachers in the classroom. Our quantitative measures captured information about teacher behavior in the classroom, student problem-solving behavior, and student engagement. We discuss each measure in detail and present results in the following sections.

SCHOOL CHARACTERIZATION

The schools that we worked with in this study serve diverse communities and the teachers and students at each school face a different set of daily challenges. School demographic factors alone do not capture the environment at each school; teacher behavior, classroom dynamics, and other social factors also have a strong impact. Extensive research in education shows that teachers greatly influence student behavior and achievement [**?**, **?**, **?**, **?**]. As a result, we were interested in viewing the process of integrating a technology-delivered curriculum in each school through the lens of the specific classroom and teacher environment.

To capture qualitative information about each classroom, we conducted a survey of the four staff members of hte non-profit who spent time in all four schools. These staff members observed and supported teachers weekly throughout the trial. The survey included both Likert-scale questions and free-form questions. The questions focused on quality of teacher instruction, level of support principals provide for teachers, and level of challenge presented by students (issues at home, English language, stress). Table **??** shows results of the Likert-scale questions and Table **??** show school demographic data. Below, we discuss the dynamics at each of the four schools, which we refer to as Schools A, B, C and D.

School A

This school serves a high-income, predominantly white neighborhood. Two teachers co-taught, leading lessons in turn. Observers reported that both teachers were prepared, kept the classroom organized, and supported each other effectively. They had a strong grasp on lesson content, and changed teaching tactics throughout the day if a particular approach was not effective. One observer noted that "*the teach-*

ers have a strong rapport with the students, leading to greater connection and willingness to pursue help during free periods and even lunch time." Students were generally well behaved, with little need for disciplinary intervention. Although a few students had behavioral or learning challenges, they often had personal aides. Teachers were well-supported by the school's principal, who visited classes occasionally.

School B

This school serves a racially diverse, low-income neighborhood. Observers reported that this teacher was ineffective and unable to restore order once students started to go off-task. The classroom was disorganized, and the teacher seemed inattentive. One observer noted that "*on every count, the instruction was weak.*" Although the teacher and students seemed to get along, his disciplinary approach did not work; consequently, both teacher and students would often raise their voices to draw attention, resulting in a chaotic learning environment. The students faced a moderate amount of home challenges, with a significant portion of students also being English Language Learners (ELL), but external help was not present. The teacher did not effectively support students in figuring out how to be more engaged. Likewise, he did not receive support or feedback to improve his teaching.

School C

This school serves a racially diverse low-income neighborhood. The teacher was typically prepared for class, although"*not ready to teach in any engaging way.*" Call-and-response or fill-in-the-blank teaching was common, which failed to retain some students' focus. However, the teacher was liked and respected by students. The students at this school presented many challenges; "*most of them have some sort of trauma that they are dealing with*" such as homelessness, unstable home environments, trauma and language barriers. Despite these issues, the school only offered a part-time counselor. Observers reported that the principal was rarely on campus, and teachers saw her as"*having abandoned the children.*" Classroom activities were often disrupted by last minute events like school assemblies and schedule changes. Overall, the classroom environment was erratic, but the teacher had a strong handle on managing her students.

School D

This school serves a racially diverse low-income neighborhood, but is co-managed by the state and a non-profit organization with a STEM-education focus. The teacher was highly prepared, providing strong support and high quality instruction. The classroom was"*unbelievably organized,*" with students obedient at all times while still having fun. One observer said that "*students would rush to do whatever she asked.*" This may have been a result of training held for students at the beginning of the year on staying focused in class. Students showed respect, and the teacher reciprocated. The school principal dropped by occasionally, and the teacher kept in regular contact with both the principal and technology lead of the school. Although many students came from immigrant families and faced some home challenges, there were no obvious signs of issues based on the effective classroom management. One observer noted that since this was a specially funded school, the students likely had families who were invested enough in education to enroll their children.

FINDINGS: TEACHER SURVEYS

At the end of the study, we asked each teacher to complete a survey about their experiences with our tablet-based software. The survey included both Likert-scale and free-form questions. The questions focused on strengths and weaknesses of the Enlearn software and the most positive and negative changes caused by introducing the software into classrooms.

Overall, teachers viewed the software positively; all responded that they would recommend the software to a colleague. They felt that students were more motivated, engaged and focused when using the tablets. Teachers noted that "*students were catching their misunderstandings quickly and asking for help*", and students were "*accountable for their learning, which...seemed to make them more receptive to receiving help*". These responses suggest that students were assisted by the real-time feedback provided by the tablet software, and that this gave them a sense of ownership over their learning.

The software also helped teachers to monitor individual student understanding. "*It was very easy to see what kids were struggling with the problems*," said one teacher. Another said "*It is a great way to see how EVERY student is doing.*" While providing individual assistance to particular students, the tablets kept other students on task. On the professional development side, one teacher answered, "*We learned a lot about our teaching practice, as well as how our students function to try and hide misconceptions and struggles in math*."

While the real-time data provided to students and teachers was effective, teachers also suggested some ways to improve the student software. One noted that "*the fill-in-the-blanks aspect seemed to prevent students from making important connections when building on concepts. I think a stylus would help with entering the answers, and would eliminate the frustration that students felt with the way they currently have to input their answers.*" Another noted that "*a few students would just mash random answers in an attempt to see how many problems they could do*." To help monitor student progress, another suggested adding a feature to the Dashboard "*to show if students got the problem wrong initially, then went back to change it to the correct answer*.

Another common concern was that the tablet was not locked, so students would sometimes go off-task to play with other apps or settings. One of the teachers at School A, which serves a high-income neighborhood, commented that since most of their students had personal electronic devices, they would often approach the exercises as if "*playing a game on their phone*". The teacher from School D, which is co-managed by a STEM-focused non-profit, also noted that some students were distracted by other apps and features because the tablets felt novel in contrast to other classroom technologies that were used more frequently.

Overall, teachers and students appeared to benefit from the availability of real-time data and feedback. While there are some important areas for future design iteration, the intro-

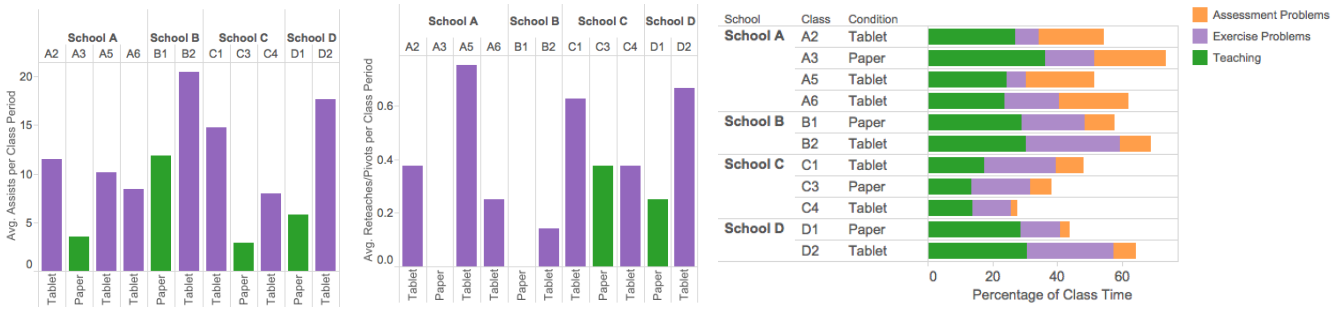


Figure 2. Figure (a) shows the average number of assists per class period, broken down by school and class. Figure (b) shows the average number of reteaches and pivots per class period, broken down by school and class. Figure (c) shows a breakdown of the percentage of time teachers spent on assessment problems, exercise problems, and teaching.

duction of a tablet-based curriculum into the classroom did not appear to disrupt the classroom workflow.

FINDINGS: TEACHER AND STUDENT BEHAVIOR

In this section, we present results from statistical analyses of a set of quantitative measures that capture teacher behavior, student problem-solving behavior, and student engagement during the study. For each outcome measure, we are primarily interested in analyzing differences based on *Condition* (either tablet or paper). However, since the schools included in our study are substantially different from each other, we also analyzed differences based on the covariate *School*. Finally, we expect that the tablet-based intervention may have different effects at different schools, so we analyzes the interaction between *Condition* and *School* to capture these differences.

Before analyzing our continuous outcome measures, we evaluated the Shapiro-Wilk test to assess the normality of our data, and found that it was statistically significant for all of our measures, meaning none of them were normally distributed. As a result, we applied the Aligned Rank Transform [**?**, **?**] procedure on each of our continuous outcome measures. This procedure aligns and ranks non-parametric data so that a standard ANOVA model can be used to perform a factorial analysis. For each main effect or interaction, the ART procedure aligns the data such that only that main effect or interaction remains, and then ranks the aligned data. A standard ANOVA model can then be used on the ranked data to measure the effect for which it was aligned. We used the ARTool program to align and rank our data [**?**].

In the cases where there is a significant *Condition***School* interaction, we perform a follow-up analysis for each school to determine how the outcome measure differs based on *Condition* at that school. For these comparisons, we use a Kruskal Wallis test to analyze our continuous outcome measures. Since this produces a large number of statistical comparisons, we risk inflating alpha. To address this potential issue, we sum p-values across all comparisons for each outcome measure to ensure that the combined alpha still falls below the 0.05 threshold. We also report effect sizes in addition to p-values to show the magnitude of the differences between our populations. We use an Eta-Squared (η^2) measure of effect size for our factorial analyses and an r measure of effect size for analyses within schools. For η^2, effects with values of 0.01 are *small*, 0.06 are *moderate*, and 0.14 or greater are *large*. For r, effects with values of 0.1 are *small*, 0.3 are *moderate*, and 0.5 or greater are *large*.

Teacher Behavior Changes With Access to Data

In addition to our surveys of the teachers experiences working with the our platform, we wanted to study how their behavior differed in the tablet and paper classes. We expected teacher to be impacted by the availability of real-time student data, and would spend more time re-teaching concepts and assisting individual students as a result. We also wanted to study how teachers chose to spend class time in each type of class, and whether tablets influenced classroom activities.

To compare teacher behavior in the two types of classes, we observed teachers regularly throughout the study, a total of 8 observation days per class. We developed a tablet application for logging teacher observations called the "Teacher Taplog." This application provides buttons that the observer can press to log the current classroom activity, such as "Teaching," "Demo," or "AP Problems." Actions that occur during a classroom activity, such as "Assisting Student," can also be logged. All observers were staff members at the non-profit organization that we partnered with, and all were trained in advance and tested for inter-rater reliability.

In our analysis of the Teacher Taplog data, we treat each class as a subject in our experiment. For each outcome measure, such as the time spent on teaching, we aggregate across all observation days to calculate the average amount of time spent teaching during that class period. While we have relative rich data about how the teacher spent time during each class, we only have a total of eleven classes in our study, which is a small number for statistical analysis. As a result, we are unlikely to see significant effects for many of our measures.

Number of Student Assists

Using the Teacher Taplog, our observers recorded each time a teacher assisted an individual student and the duration of the assist. We analyzed the average amount of time that teachers spend assisting students, and while *Condition* did not have a significant effect, there was a trend towards teachers assisting students more in the tablet classes ($F(1,3)=8.459$, $p=0.0621$).

Figure 3. Histogram of exercise problem completion for Unit 4. Each bar represents the percentage of students who fall into the related bin.

Neither *School* (F(3,3)=1.462, *n.s.*) nor the *Condition*School* interaction (F(3,3)=0.131, *n.s.*) had a significant effect on assist time. We also analyzed the average number of times teachers assisted students. Again, *Condition* did not have a significant main effect, but there was a trend toward teachers assisting students more often in the tablet-based classes (F(1,3)=7.493, p=0.0715). Neither *School* (F(3,3)=1.652, *n.s.*) nor the *Condition*School* interaction (F(3,3)=0.283, *n.s.*) had a significant effect on number of student assists.

Figure **??** shows the average number of assists for each of the eleven classes. The visualization clearly suggests that teachers assist students more often in the tablet classes. This suggests that teachers respond to the real-time student data, a finding that is supported by the results of the teacher survey. Teachers reported that the software helped them identify struggling students and encouraged students to ask for help.

Number of Re-Teach Events

A "re-teach" event is when a teacher stops the class to clarify a concept or present it in a new way. Our observers were trained to recognize and log re-teach events using the Teacher Taplog. While our analysis showed that emphCondition did not have a significant effect on re-teaches, there was a trend towards more re-teach events in the tablet classes (F(1,3)=8.459,p=0.0621). Neither *School* (F(3,3)=1.131, *n.s.* nor the *Condition*School* interaction (F(3,3)=0.444, *n.s.*) had a significant effect on the number of re-teach events.

Figure **??** shows the average number of re-teach events during each class. The data show that re-teaches happen very infrequently, less than once per class, but that teachers are re-teaching more often in the tablet classes. This suggests that teachers are using real-time student data to determine when students are struggling and need additional teaching on a concept, an observation that mirrors teacher survey comments.

Class Time Usage

In addition to studying specific classroom behaviors that we expected to change as a result of introducing real-time student data, we also wanted to study how teachers used their time in the tablet and paper classes. In particular, we thought that the time spent teaching and on exercise and AP problems might be influenced by the introduction of tablets. To evaluate this question, we first analyzed the average time spent teaching during each class. We found that *Condition* did not have a significant effect, but there was a trend towards spending less time teaching in tablet classes (F(1,3)=7.493, p=0.0715). *School* also had no significant effect, but we measured a trend toward teachers at Schools B and D spending more time teaching than those at Schools A and C F(3,3)=7.248, p=0.0690). There was no significant *Condition*School* interaction. We also analyzed the amount of time spent on exercise and AP problems, but found no significant differences based on *Condition*, *School*, or their interaction.

Figure **??** shows the average amount of time spent on teaching, exercise problems, and AP problems in each class. While we thought the tablet-delivered curriculum might strongly impact how teacher spend class time, especially problem-solving time since problems were delivered to students through tablets, but this expectation was not substantiated. We did see a trend towards teachers spending less time on teaching in the tablet-based classes, which could be an indication that more time was spent on other activities such as assisting students or classroom management.

Student Problem-Solving Behavior is Impacted by Tablets

We expected student problem-solving behavior to differ in the paper and tablet classes. We hypothesized that the adaptive, self-paced problem progressions would allow students to complete more Exercise and AP problems in the tablet classes, and that students would perform better on the AP problems and unit post-tests. To study problem-solving behavior, we recorded two measures: the number of problems completed, and problem correctness. In the tablet classes, we logged student interactions with the tablet to capture these measures. In the paper classes, we counted and graded AP problems, and counted Exercise problems. The AP problems were clearly labeled in the JUMP AP notebook, but exercise problems were completed in students' personal notebooks, making them difficult to interpret and grade. To count Exercise problems, graders used a well-organized notebook from each class as a template for counting the rest of the notebooks.

We analyze problem-solving behavior for Units 2 and 4 separately. The units cover different content that could influence problem-solving behavior, and the teachers at Schools B and C ran out of time for Unit 4 and rushed some of the content. As a result, we treat these units separately in the analysis.

Number of Exercise Problems Completed

We analyzed the number of exercise problems completed, and found that *Condition* had a significant main effect in both Unit 2 (F(1,209)=141.669, p<0.0001, η^2=0.77) and Unit 4 (F(1,211)=373.031, p<0.0001, η^2=0.90), with students completing more exercise problems in the tablet classes. *School* also had a significant main effect in both Unit 2 (F(3,209)=160.007, p<0.0001, η^2=0.74) and Unit 4 (F(3,211)=53.950, p<0.0001, η^2=0.95). Finally, the *Condition*School* interaction had a significant effect in both Unit 2 (F(3,209)=62.511, p<0.0001, η^2=0.66) and Unit

School	Unit	Condition	# Exercise Problems			# AP Problems			% AP Correct	% Pre-Test	% Post-Test
A	Unit 2	Tablet	51.5	$N = 70$	$p < 0.05$	351.5	$N = 70$	$p < 0.005$	97.95%	33%	56%
		Paper	77.5	$Z = 2.181$	$r = 0.26$	399.5	$Z = 3.076$	$r = 0.37$	92.18%	35%	53%
	Unit 4	Tablet	334.8	$N = 72$	$p < 0.001$	277.5	$N = 72$	$p < 0.005$	98.39%	45%	73%
		Paper	149.5	$Z = -5.587$	$r = 0.66$	245	$Z = -3.251$	$r = 0.38$	87.84%	45%	81%
B	Unit 2	Tablet	669	$N = 39$	$p < 0.001$	70.5	$N = 39$	$p < 0.05$	88.46%	13%	35%
		Paper	284	$Z = 5.041$	$r = 0.81$	97.25	$Z = -2.478$	$r = 0.40$	79.45%	33%	43%
	Unit 4	Tablet	257	$N = 39$	$p < 0.001$	89	$N = 39$	*n.s.*	88.46%	9%	22%
		Paper	26	$Z = 5.115$	$r = 0.82$	89.5	$Z = -0.397$		75.36%	27%	45%
C	Unit 2	Tablet	503	$N = 63$	$p < 0.001$	230	$N = 63$	$p < 0.001$	95.29%	50%	60%
		Paper	157	$Z = -6.342$	$r = 0.80$	440	$Z = 5.552$	$r = 0.70$	85.69%	40%	47%
	Unit 4	Tablet	411	$N = 63$	$p < 0.001$	206	$N = 63$	$p < 0.005$	97.24%	45%	68%
		Paper	66.5	$Z = -6.342$	$r = 0.80$	246.5	$Z = 3.256$	$r = 0.41$	89.96%	59%	82%
D	Unit 2	Tablet	686.5	$N = 45$	$p < 0.001$	236	$N = 45$	*n.s.*	97.07%	35%	48%
		Paper	347	$Z = 5.733$	$r = 0.85$	262	$Z = -1.124$		89.29%	40%	46%
	Unit 4	Tablet	847.5	$N = 45$	$p < 0.001$	126	$N = 45$	$p < 0.05$	95.76%	36%	84%
		Paper	226	$Z = 5.734$	$r = 0.86$	165	$Z = -2.124$	$r = 0.32$	85.94%	36%	90%

Table 3. Results of the comparisons within each school. For each analysis, we show the median values to the left of the cell, along with results of the statistical tests. We also report median percent of AP problems correct and median scores on the unit pre- and post-tests for each school.

4 (F(3,211)=22.971, $p<0.0001$, $\eta^2=0.74$), suggesting that tablets had a different effects in different schools.

To understand the effect of *Condition* at each school, we ran a series of follow-up analyses. The results are shown in Table **??**. Students in the tablet classes at Schools B, C, and D complete a huge number of exercise problems, significantly more than those in the paper classes. However, at School A, students in the paper class complete significantly more problems during Unit 2. This trend is reversed in Unit 4; however the Unit 4 effect size is smaller, perhaps because School A spent less time on the exercise problems than the other schools.

The histograms in Figure **??** show that while students completed more exercise problems in the tablet classes on average, there was a large amount of variation. There is much less variation in the paper classes, where students completed exercise problems in lock-step. This shows the potential benefit of generative adaptivity: students can work through problems at their own pace, practicing basic concepts or advancing to more challenging concepts as needed.

Number of AP Problems Completed

We also analyzed the number of AP problems completed. We found that *Condition* had a significant effect in both Unit 2 (F(1,209)=41.296, $p<0.0001$, $\eta^2=0.92$) and Unit 4 (F(1,211)=14.500, $p<0.0005$, $\eta^2=0.74$), and that overall students completed *fewer* AP problems in the tablet classes. emphSchool also had a significant effect in both Unit 2 (F(3,209)=50.775, $p<0.0001$, $\eta^2=0.90$) and Unit 4 (F(3,211)=106.607, $p<0.0001$, $\eta^2=0.85$). Finally, the *Condition*School* interaction had a significant effect in both Unit 2 (F(3,209)=20.880, $p<0.0001$, $\eta^2=0.73$) and Unit 4 (F(3,211)=9.527, $p<0.0001$, $\eta^2=0.46$), again suggesting that tablets had different effects in different schools.

We ran follow-up analyses at each school, and the results are shown in Table **??**. At School A, students in the paper class completed more AP problems in Unit 2, while students in the tablet class completed more in Unit 4, mirroring the results for exercise problems. However, at Schools B, C, and D, students in the paper classes completed either the same number of AP problems or more than those in the tablet classes.

These results are surprising. Our analysis of the Teacher Taplog data shows that there is no difference in the amount of time teachers spent on AP problems based on *Condition*. However, at Schools B, C, and D, the data trends towards teachers spending more time on exercise problems and less time on AP problems in the tablet classes. It is possible that students are solving fewer AP problems than exercise problems due to the different adaptive policy used for the AP. For AP problems, students are advanced to the next problem set regardless of whether they have reached mastery, while with exercise problems students continue on a given problem set until mastery is reached. This suggests that a mastery learning policy may be more effective.

Percentage of AP Problems Correct

In addition to counting AP problems, we graded AP problem correctness. We found that *Condition* had a significant effect for both Unit 2 (F(1,207)=116.217, $p<0.0001$, $\eta^2=0.91$) and Unit 4 (F(1,208)=160.240, $p<0.0001$, $\eta^2=0.89$), with students in the tablet classes performing better than those in the paper classes. *School* also had a significant effect for both Unit 2 (F(3,207)=39.806, $p<0.0001$, $\eta^2=0.81$) and Unit 4 (F(3,208)=37.062, $p<0.0001$, $\eta^2=0.81$). However, the *Condition*School* interaction did not have a significant effect for Unit 2 (F(3,207)=4.125, *n.s.*) and had only a small effect in Unit 4 (F(3,208)=3.936, $p<0.0001$, $\eta^2=0.34$), suggesting that the effect of tablets did not differ much across schools.

We report the median percentage of correct AP problems for each school and condition in Table **??**. Students in the tablet-based classes perform better on the AP problems across the board, with scores between 5 and 12 percentage points higher than students in the paper-based classes. We chose not to perform a follow-up analysis of this measure since we did not measure a strong *Condition*School* interactions and the median values suggest a similar trend across all schools.

This result is interesting. While students in the tablet classes are completing fewer AP problems, they are getting more AP

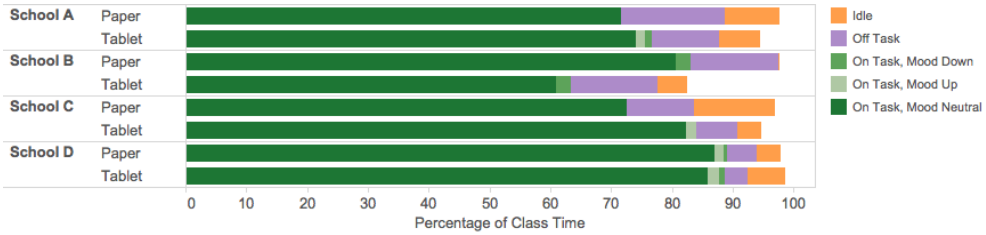


Figure 4. Average percentage of class time that students spent in each engagement state, broken down by condition and school.

problems correct than students in the paper class. It is possible that students in the tablet classes perform better because they had more practice on exercise problems before reaching the AP. It's also likely that students are reacting to the real-time correctness feedback provided by the tablet software by going back to fix mistakes. Since students in the paper classes have no feedback about problem correctness, it is not surprising that they perform more poorly. This is one of the strong advantages of a tablet-based curriculum; it allows students to correct mistakes as they happen.

Pre/Post Tests

In addition to the exercise and AP problems that were completed as part of class, students were given formal tests before and after each JUMP unit. We graded and analyzed student tests for both Unit 2 and 4. The JUMP curriculum includes a single test for each unit, but we wanted students to take a different version of the test before and after the unit to control for any learning effects. We asked JUMP to create a second version of each unit test that included the same item types with different numbers. Half of the students were given version A as their pre-test and version B as their post-test, and the other half were given the opposite. Many of the test items had multiple parts; we grade each part separate, and gave student partial credit. All tests were graded by the same researcher.

We analyzed test data for each unit with a repeated measures ANOVA, measuring the effect of *Condition* and *Test* (either pre-test or post-test). We found that *Condition* did not have a significant effect on in either Unit 2 ($F(1,210)=2.165$, *n.s.*) or Unit 4 ($F(1,203)=0.179$, *n.s.*). *Test* had a significant effect for both Unit 2 ($F(1,210)=417.807$, $p<0.0001$) and Unit 4 ($F(1,203)=259.621$, $p<0.0001$), showing that students performed better on the post-tests than on the pre-tests. We report median pre- and post-test scores for each school in Table **??**. Recall that students in the paper classes were stronger on average than those in the tablet classes.

While students in the tablet classes performed better on the AP problems, these learning gains did not transfer to the post-test. Gains may not have transferred because students solved all exercise and AP problems through the tablet interface, and were not used to working on paper. It is also possible JUMP Math tests may not have effectively covered the material teachers focused on in class.

Student Engagement Was Impacted by Tablets

We were interested in measuring how student engagement impacted by the introduction of a tablet-based curriculum. To measure engagement, we observed students during the study and logged information about their affect using the Baker-Rodrigo Ocumpaugh Monitoring Protocol (BROMP). BROMP is a protocol for capturing student affect through quantitative field observations [**?**] that has been used widely in educational technology research (e.g. [**?**, **?**]). Using this protocol, the observer works through students sequentially, assessing affect and logging this information with a smartphone application. By looping through students repeatedly, the coder captures a large number of observations for each student. We recorded two variables: student *attention* (on-task, off-task, idle, or unknown) and student *mood* (positive, negative, neutral, or unknown). All observers were trained in advance and tested for inter-rater reliability. We observed each class on three separate days, except at School B where we only observed twice due to scheduling constraints.

Time On Task

First we studied how the tablets impacted students' on-task time. For each student, we calculated the percentage of total observations for which the student was on-task. We found that *Condition* did not have a significant main effect on on-task time ($F(1,208)=0.490$, *n.s.*). However, both *School* ($F(3,208)=11.178$, $p<0.0001$, $\eta^2=0.82$) and the *Condition*School* interaction ($F(3,208)=8.505$, $p<0.0001$, $\eta^2=0.81$) had significant effects on the amount of time students spent on-task.

We performed follow-up analyses at each school, which are reported in Table **??**. The tablets had very different effects on student attention at the four schools. We measured no significant difference in on-task time at Schools A and D, where the teachers are highly organized and class time is carefully structured. At School B, students in the tablet class spent significantly less time on-task than those in the paper class. This teacher was the least organized and had trouble managing his tablet classes, which could explain the lower level of on-task time. At School C, students in the tablet class spent significantly more time on-task. At this school, the teacher had to manage a number of challenging students, and class time was often disrupted. In the tablet classes, students were able to continue working while the teacher engaged in classroom management activities, which could explain the higher level of on-task time in the tablet classes.

On-Task Excitement

We also studied students' moods while they were on-task, specifically looking for indications of positive affect or excitement. We found that *Condtion* had a significant effect ($F(1,208)=9.980$, $p<0.005$, $\eta^2=0.63$), with students in

Condition	% On-Task			% On-Task Excitement		
A Tablet	78%	$N = 71$	*n.s.*	1.5%	$N = 71$	$p < 0.001$
A Paper	73%	$Z = -1.430$		0%	$Z = -3.900$	$r = 0.46$
B Tablet	70%	$N = 38$	$p < 0.05$	0%	$N = 38$	*n.s.*
B Paper	83%	$Z = -2.424$	$r = 0.39$	0%	$Z = -0.416$	
C Tablet	86%	$N = 62$	$p < 0.005$	1.6%	$N = 62$	$p < 0.005$
C Paper	74%	$Z = -3.390$	$r = 0.43$	0%	$Z = -3.171$	$r = 0.40$
D Tablet	90%	$N = 45$	*n.s.*	1.9%	$N = 45$	*n.s.*
D Paper	90%	$Z = 0.204$		1.3%	$Z = 0.539$	

Table 4. Results of the comparisons within each individual school. For each analysis, we show the median values to the left of the cell, along with results of the statistical tests.

the tablet classes displaying more on-task positive affect than those in the paper classes. *School* did not have a significant main effect (F(3,208)=0.257, *n.s.*), however the *Condition*School* interaction did have a significant effect (F(3,208)=4.255, $p<0.01$, $\eta^2=0.77$).

We performed follow-up analyses at each school, which are reported in Table **??**. The results show that overall, there were very few observations of on-task excitement. However, at School A and School C, we observed significantly more on-task excitement in the tablet classes than in the paper classes. At School B there was almost no on-task positive affect observed in either class, perhaps an indication of the poor relationship that the teacher had with students. At School D, there was a relatively high rate of on-task excitement in both classes, which could be a result of the strong relationship that teacher had with students.

CONCLUSION

REFERENCES

1. Rahul Agarwal, Stephen H. Edwards, and Manuel A. Pérez-Quiñones. 2006. Designing an Adaptive Learning Module to Teach Software Testing. *SIGCSE Bull.* 38, 1 (March 2006), 259–263. DOI: http://dx.doi.org/10.1145/1124706.1121420
2. Erik Andersen, Sumit Gulwani, and Zoran Popović. 2013. A Trace-based Framework for Analyzing and Synthesizing Educational Progressions. In *CHI*. 773–782.
3. John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences* 4, 2 (1995), 167–207.
4. Richard Anderson, Ruth Anderson, K. M. Davis, Natalie Linnell, Craig Prince, and Valentin Razmov. 2007. Supporting Active Learning and Example Based Instruction with Classroom Technology. *SIGCSE Bull.* 39, 1 (March 2007), 69–73. DOI: http://dx.doi.org/10.1145/1227504.1227338
5. Kimberly E. Arnold and Matthew D. Pistilli. 2012. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge (LAK '12)*. ACM, New York, NY, USA, 267–270. DOI: http://dx.doi.org/10.1145/2330601.2330666
6. Ryan Shaun Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. 2004. Off-task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 383–390. DOI: http://dx.doi.org/10.1145/985692.985741
7. Madeline Balaam, Geraldine Fitzpatrick, Judith Good, and Rosemary Luckin. 2010. Exploring Affective Technologies for the Classroom with the Subtle Stone. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1623–1632. DOI: http://dx.doi.org/10.1145/1753326.1753568
8. B. S. Bloom. 1968. Learning for mastery. *Evaluation Comment* 1, 2 (1968), 1–12.
9. Benjamin S. Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher* 13, 6 (1984), pp. 4–16. http://www.jstor.org/stable/1175554
10. Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy Vanderwende. 2014. Divide and Correct: Using Clusters to Grade Short Answers at Scale. In *Proceedings of the First ACM Conference on Learning @ Scale Conference (L@S '14)*. ACM, New York, NY, USA, 89–98. DOI: http://dx.doi.org/10.1145/2556325.2566243
11. Eric Coopey, Ethan Danahy, and Leslie Schneider. 2013. InterLACE: Interactive Learning and Collaboration Environment. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion (CSCW '13)*. ACM, New York, NY, USA, 11–14. DOI: http://dx.doi.org/10.1145/2441955.2441959
12. A. Corbett, K. R. Koedinger, and J. R Anderson. 1997. Intelligent Tutoring Systems. In *Handbook of Human-Computer Interaction, Second Edition*, M. Helander, T. K. Landauer, and P. Prahu (Eds.). Elsevier Science, Amsterdam, 849–874.
13. Albert T. Corbett and John R. Anderson. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Model. User-Adapt. Interact.* 4, 4 (1995), 253–278.
14. ALEKS Corporation. 2015. ALEKS – Assessment and Learning, K-12, Higher Education, Automated Tutor, Math. (2015). http://www.aleks.com/ [Online; accessed 19-May-2015].
15. H. L. Dangel and C. X. Wang. 2008. Student response systems in higher education: Moving beyond linear teaching and surface learning. *Journal of Educational Technology Development and Exchange* 1, 1 (2008), 93–104.

16. James P. Gee. 2008. *What Video Games Have to Teach Us About Learning and Literacy* (revised and updated edition. ed.). St. Martin's Press. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1403984530

17. Arthor C. Graesser, Shulan Lu, George T. Jackson, Heather H. Mitchell, Mathew Ventura, Andrew Olney, and Max M. Louwerse. 2004. AutoTutor: a tutor with dialogue in natural language. *Behavior Research Methods, Instruments, and Computers* 36, 2 (2004).

18. Timothy J. Hickey and William T. Tarimo. 2014. The Affective Tutor. *J. Comput. Sci. Coll.* 29, 6 (June 2014), 50–56. http://dl.acm.org/citation.cfm?id=2602724.2602735

19. J. J. Higgins and S. Tashtoush. 1994. An aligned rank transform test for interaction. *Nonlinear World* 1, 2 (1994), 201–211.

20. Heather C. Hill, Brian Rowan, and Deborah Loewenberg Ball. 2002. Effects of Teachers? Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal* 42, 2 (2002), 371–406. http://aer.sagepub.com/content/42/2/371.full.pdf+html

21. Juho Kim, Elena L. Glassman, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. RIMES: Embedding Interactive Multimedia Exercises in Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1535–1544. DOI:http://dx.doi.org/10.1145/2702123.2702186

22. Kenneth R. Koedinger and John R. Anderson. 1997. Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education* 8 (1997), 30–43.

23. Kimberle Koile and David Singer. 2006. Improving Learning in CS1 via tablet-PC-based In-class Assessment. In *Proceedings of the Second International Workshop on Computing Education Research (ICER '06)*. ACM, New York, NY, USA, 119–126. DOI:http://dx.doi.org/10.1145/1151588.1151607

24. Hristina Kostadinova, George Totkov, and Hristo Indzhov. 2012. Adaptive e-Learning System Based on Accumulative Digital Activities in Revised Bloom's Taxonomy. In *Proceedings of the 13th International Conference on Computer Systems and Technologies (CompSysTech '12)*. ACM, New York, NY, USA, 368–375. DOI:http://dx.doi.org/10.1145/2383276.2383330

25. Stefan Kreitmayer, Yvonne Rogers, Robin Laney, and Stephen Peake. 2013. UniPad: Orchestrating Collaborative Activities Through Shared Tablets and an Integrated Wall Display. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 801–810. DOI:http://dx.doi.org/10.1145/2493432.2493506

26. J. A. Kulik and C. C. Kulik. 1989. Meta-analysis in education. *International Journal of Educational Research* 13, 2 (1989), 221–340.

27. Chinmay Kulkarni, Julia Cambre, Yasmine Kotturi, Michael S. Bernstein, and Scott R. Klemmer. 2015. Talkabout: Making Distance Matter with Small Groups in Massive Classes. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1116–1128. DOI:http://dx.doi.org/10.1145/2675133.2675166

28. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* 20, 6, Article 33 (Dec. 2013), 31 pages. DOI:http://dx.doi.org/10.1145/2505057

29. Alina Lazar. 2007. Engaged Learning in a Computer Science Course. *J. Comput. Sci. Coll.* 23, 1 (Oct. 2007), 38–44. http://dl.acm.org/citation.cfm?id=1289280.1289288

30. IXL Learning. 2015. IXL Math and English — Online math and language arts practice. (2015). https://www.plickers.com/ [Online; accessed 19-May-2015].

31. Yun-En Liu, Christy Ballweber, Eleanor O'Rourke, Eric Butler, Phonraphee Thummaphan, and Zoran Popović. 2015. Large-Scale Educational Campaigns. *ACM Trans. Comput.-Hum. Interact.* 22, 2, Article 8 (March 2015), 24 pages. DOI:http://dx.doi.org/10.1145/2699760

32. JUMP Math. 2015. JUMP Math. (2015). http://jumpmath.org/

33. Merrilea J. Mayo. 2009. Video Games: A Route to Large-Scale STEM Education? *Science* 323 (2009), 79–82.

34. Beverley Murray. 2015. *Increased Math Achievement in Elementary Students Participating in JUMP Maths 2013-14 National Book Fund Program*. Technical Report.

35. J. Ocumpaugh, R.S. Baker, and M.M.T. Rodrigo. 2015. *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual*. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

36. Eleanor O'Rourke, Kyla Haimovitz, Christy Ballweber, Carol Dweck, and Zoran Popović. 2014. Brain Points: A Growth Mindset Incentive Structure Boosts Persistence in an Educational Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3339–3348. DOI:http://dx.doi.org/10.1145/2556288.2557157

37. Zachary A. Pardos, Ryan S. J. D. Baker, Maria O. C. Z. San Pedro, Sujith M. Gowda, and Supreeth M. Gowda.

2013. Affective States and State Tests: Investigating How Affect Throughout the School Year Predicts End of Year Learning Outcomes. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*. ACM, New York, NY, USA, 117–124. DOI: http://dx.doi.org/10.1145/2460296.2460320

38. New Classrooms Innovation Partners. 2015. New Classrooms Innovation Partners. (2015). http://www.newclassrooms.org/ [Online; accessed 19-May-2015].
39. Plickers. 2015. Plickers. (2015). https://www.plickers.com/ [Online; accessed 19-May-2015].
40. Johnmarshall Reeve, Hyungshim Jang, Dan Carrell, Soohyun Jeon, and Jon Barch. 2004. Enhancing Students' Engagement by Increasing Teachers' Autonomy Support. *Motivation and Emotion* 28, 2 (2004), 147–169. DOI: http://dx.doi.org/10.1023/B:MOEM.0000032312.95499.6f
41. K. C. Salter and R. F. Fawcett. 1993. The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation* 22, 1 (1993), 137–153.
42. Daniel Szafir and Bilge Mutlu. 2013. ARTFul: Adaptive Review Technology for Flipped Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1001–1010. DOI: http://dx.doi.org/10.1145/2470654.2466128
43. SMART Technologies. 2015. SMART Education - SMART Technologies. (2015). http://education.smarttech.com/ [Online; accessed 19-May-2015].
44. Kurt VanLehn. 2006. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education* 16 (2006), 227–265.
45. Kathryn R. Wentzel. 2002. Are Effective Teachers Like Good Parents? Teaching Styles and Student Adjustment in Early Adolescence. *Child Development* 73, 1 (Jan. 2002), 287–301. http://onlinelibrary.wiley.com/store/10.1111/1467-8624.00406/asset/1467-8624.00406.pdf?v=1&t=i9yfhr3p&s=a0e68e6a2d2c2bf6c579e860419e72d88966d6ab
46. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 143–146. DOI: http://dx.doi.org/10.1145/1978942.1978963
47. S. Paul Wright, William L. Sanders, and Sandra P. Horn. 1997. Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education* 11, 1 (1997), 57–67. DOI: http://dx.doi.org/10.1023/A:1007999204543
48. Meilan Zhang, Robert Trussell, Benjamin Gallegos, and Rasmiyeh Asam. 2015. Using Math Apps for Improving Student Learning: An Exploratory Study in an Inclusive Fourth Grade Classroom. *TechTrends: Linking Research & Practice to Improve Learning* 59, 2 (2015), 32 – 39. http://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=100711551&site=ehost-live