RACE TO THE TABLET? THE IMPACT OF A PERSONALIZED TABLET EDUCATIONAL PROGRAM

Elizabeth Setren

Department of Economics
Tufts University
Medford, MA 02155
elizabeth.setren@tufts.edu

Abstract

The presence of tablets and laptops in schools has burgeoned in recent years, with $4.9 billion spent on over 10.8 million devices in 2015. Despite the large and increasingly prevalent monetary and time investments in education technology, little causal evidence of its effectiveness exists. I estimate the effect of a Math and English Language Arts tablet educational program that supplements core instruction using a randomized controlled trial in a Boston charter middle school. I find that the personalized learning technology can substantially increase end-of-year test scores by 0.202 standard deviation in Math, but find no effects for the summative English exam. For the quarterly formative exams, I find positive, but insignificant effects for Math and marginally significant effects for English. This paper demonstrates the potential of technology to enhance student learning in Math and could serve as a cheaper alternative to high-intensity tutoring for school districts without funding or labor supply for extensive tutoring programs.

https://doi.org/10.1162/edfp_a_00359

1. INTRODUCTION

Tablets, laptops, and other devices have a large and growing presence in U.S. classrooms. Elementary and secondary schools spend an estimated $8.38 billion on educational software and digital content and $4.9 billion on devices annually (Education Technology Industry Network 2015; Huang 2016). Educational technology companies claim that their programs target students' gaps in skills and improve student outcomes. Despite the increasing adoption of technology in the classroom, limited work on its effectiveness exists.

This paper analyzes the impact of a popular educational technology program, called eSpark, on students' academic outcomes using a randomized controlled trial (RCT). The experiment occurred in a Boston charter middle school for one class period during the school day as a supplement to core Math and English instruction. Students in grades 6 through 8 were randomly assigned to a treatment or control class. The classes met for 28 minutes a day, four days a week for three fourths of the school year. Students in the treatment class used a personalized learning tablet software that gave students interactive practice with the Common Core concepts in which they lagged most according to a pre-test. The school sorted the control group students into teacher-led, tracked classrooms based on ability.

The study is powered to detect substantial test score effects of around 0.2 standard deviation. I find that the personalized technology program increased students' end-of-year Math scores by 0.202 standard deviation. However, there were positive, but insignificant, effects on the quarterly formative Math exam. For English, the study estimates marginally statistically significant effects on the quarterly formative exam, but no significant effects on the end-of-year exam.

This study contributes to a growing literature on the effectiveness of technology in education (see Escueta et al. [2017] and Bulman and Fairlie [2016] for summaries). The limited research on the impact of computers and internet in classrooms has shown mixed results. Researchers find no test score effects from adding computers or Internet access to classrooms without guidance on how to use it for educational purposes (Angrist and Lavy 2002; Goolsbee and Guryan 2006; Machin, McNally, and Silva 2007). Banerjee et al. (2007) and Barrow, Markman, and Rouse (2009) find evidence that Math computer programs have positive effects on test scores in supplemental instruction and core instruction, respectively. Rouse and Krueger (2004) find no positive evidence of a supplemental instruction English computer program on English language skill growth. Muralidharan, Singh, and Ganimian (2019) find positive effects of a similar personalized learning technology in India during an after-school program on students' Math and language test scores and highlights the role of personalization in explaining the large effects. This study is one of the first to analyze modern technology, such as tablets and app-based learning tools, in a U.S. context.

The next section explains the technology intervention, data, and sample. Section 3 outlines the empirical framework and Section 4 reports the results. The final section concludes.

2. BACKGROUND AND DATA

Intervention Details

The eSpark program creates a personalized supplemental curriculum for each student based on a pre-assessment and adapts based on each student's progress. Students take a pre-assessment that captures their knowledge of the Math and English Language Arts (English) Common Core standards.[1] Based on the test results, eSpark creates an individualized curriculum of Math and English Common Core–aligned concepts for the student to learn and practice. The program has students first work on the most fundamental Common Core standard that they have not yet mastered based on the pre-assessment.[2] For each concept, students take a short pre-quiz and then watch an interactive video lesson. Next, they practice and reinforce what they learned in the video through an interactive app exercise. Then, students take a post-quiz on that concept. If they pass, they record a video where they teach what they learned and move onto the next skill.

If they do not pass, then they watch a different instructional video on the concept. Students are given the option to repeat the practice activities and are required to take a post-quiz. If students fails the post-quiz for the second time, then they start the same process for that domain in the Common Core, but for one grade-level lower.

Once students pass the post-quiz in a domain, they repeat the same process for the next most fundamental Common Core concept, which could be Math or English. The eSpark company curates educational apps, videos, and other resources to fit in their program, instead of creating the educational content themselves.

The curriculum is aligned with Common Core Math and English Language Arts standards for pre-kindergarten through eighth grade. Depending on their ability at the start of the program, students can work on below grade-level skills to catch up, practice grade-level skills, or continue to more advanced concepts.[3] At the time of the RCT, eSpark marketed toward grades pre-K through 8 and provided content for the Common Core standards in those grades. Today, eSpark markets to grades K through 5 so that students can practice at below and above grade-level standards. Efficacy of the program could vary with students' developmental capacity.

At the time of the intervention, the costs of this program were approximately $165 per student. The school already owned 60 iPads, so they did not incur additional hardware costs. Utilizing eSpark required a one-time purchase of a library of educational apps at $40 per device and annual costs: software and services at $90 per student, professional development, and technology services.[4]

One teacher supervised the classroom and distribution of the iPads for each of the treatment classrooms. The intervention required minimal time investment and preparation from the teacher. She participated in short virtual demonstrations of the program's interface. There was no lesson planning or other work outside of the classroom. During class, she managed the distribution and collection of the tablets, helped students with any technical issues (e.g., if they forgot their password), and circulated the room to check that students were actively engaged and to answer questions.

[1] In this study, all students took the i-Ready exam as the pre-test.

[2] For example, if a student incorrectly answered questions for a seventh-grade Math standard, but answered all sixth-grade Math and English standards and seventh-grade English questions correctly, then the student would start with that seventh-grade Math standard.

[3] Eighth graders cannot work on above grade-level material because there is no content for high school standards.

[4] Costs of eSpark and other adaptive learning technology vary by enrollment size and over time.

Experimental Design

In the 2013–14 school year, 438 middle school students at UP Academy Boston participated in the randomized controlled trial. We randomly assigned 60 students to the treatment group and the remaining students to the control group. The experiment compares the individual-level personalization of lessons in app-based lessons and practice sessions to a coarser, classroom-level ability tracking led by a teacher.

We stratified the individual student-level random assignment by grade and by the subject the student scored lowest on the pre-test. This assigned between 118 and 122 students to the control group in each grade. Due to the small sample size, prior to randomization we planned to re-sample until the treatment and control groups' baseline test scores were equal at the 90 percent confidence level (see Morgan and Rubin [2012] and Imbens [2011] for discussion of re-randomization plans).

The school assigned the treatment group to work with the eSpark program in a separate classroom during the school day for twenty-eight minutes a day, four days a week for the length of the school year. The experiment occurred during an open-block in the schedule and did not replace core instruction. In previous school years, that block was used for independent reading time. The school intended for this period to focus on the skills in which students lagged most through personalized learning technology in the treatment group and ability-tracked classrooms in the control group. The technology intervention targeted individual students' skill-level by ranking Common Core concepts by the student's level of mastery and working through each of them. The experiment did not change classroom assignment (and as a result classroom peers) for the rest of the students' schedules.

The control group experienced a teacher-led, coarser ability-tracking. Due to the larger size of the control group, the school split the students into several classrooms. Figure A.1 summarizes the placement guidelines for sorting students into each of the control group classrooms (available in a separate online appendix that can be accessed on *Education Finance and Policy's* Web site at https://doi.org/10.1162/edfp_a_00359). Students who performed relatively worse on the Math pre-test were supposed to enroll in one of two Math classrooms based on test performance.[5] One section reviewed basic computational skills from the previous grade, while the more advanced section reviewed grade-level procedural skills and Math facts to promote computational fluency. Students who performed relatively worse on the English Language Arts pre-test were sorted into one of three reading groups. Students with the lowest pre-test scores practiced reading comprehension strategies for passages at their reading level. The middle group practiced word and passage fluency. The most advanced group scored proficient or higher on Math and English. They spent the period reading the book of their choice independently.

Every two months, the control group students were reassigned to new groups based on their ability level at that time. If the control group teacher determined that a student had not yet mastered that level of reading or Math, that student remained in that section.

[5] In practice, the school did not perfectly follow the placement guidelines due to class size and other constraints. Of students who scored relatively lower on Math than English in the control group, 64 percent took a Math intervention in the first session (including the 15 percent who were missing session intervention data) and 85 percent had at least one Math intervention during the school year.

If students passed their previous section, they went to the next-lowest ability–level class based on their pre-test and a writing diagnostic.[6] In the third session, the independent reading class was replaced with a guided reading class where the class read short stories and analyzed them out loud.[7]

Over 60 percent of the control group students switched between subjects (Math and English) at least once during the year. Figure A.1 available in the online appendix shows movement of students between intervention types—both across subjects and levels of difficulty within subject. Of the 40 percent of the control group that focused on one subject the entire year, that vast majority focused on English: 35 percent of the control group focused just on English while 5 percent focused just on Math. The students who just received English interventions focused on English either because they were farther behind in English than Math throughout the year (22 percent of the control group) or scored proficient or higher on the baseline Math and English exams (13 percent of the control group).

Figure A.2 in the online appendix displays the proportion of time students spent in their initially assigned subjects. The figure shows that on average students received more time in English compared with Math, even for those who were initially assigned to Math. It also reveals that student assignment did not perfectly follow the placement guidelines outlined in Figure A.1. Eleven percent of students who scored lower on Math relative to English only had English interventions throughout the year. Likewise, 4 percent of students who scored lower on English relative to Math received only Math interventions throughout the year.

I unfortunately do not have data on the balance of Math versus English content in the treatment group—individual-level app usage data were not available. Two factors suggest that the treatment group experiences a similar or higher level of balance between the two subjects. First, the treatment group received finer ability tracking. A student who received a proficient score on Math would not receive any Math instruction in the control group, but could receive Math instruction in the treatment group. Second, a student who scored proficient or higher on Math and English in the control group would only receive English interventions (independent and guided reading), but in the treatment group they could receive at or above grade-level instruction for either subject.[8]

It is important to note that the control group students' experiences varied by their baseline ability. Therefore, any differential effects by baseline ability could be due to differences in the counterfactual. Figure A.3 in the online appendix shows that students with higher baseline test scores spend relatively less time in Math and more time in English. Figure A.4 shows that students who scored relatively higher on the baseline

[6] The school added a writing class for the second through fourth sessions, where students practiced grammar, syntax, and voice.

[7] Student assignment to control group classrooms did not perfectly follow these guidelines. Teachers could use their own discretion in class placement.

[8] I started another similar randomized controlled trial in another school in the charter network, UP Academy Dorchester. Unfortunately, after randomization the school did not have the resources to implement the intervention (for reasons unrelated to this study and the perceived quality of the intervention). The results of that study are available at the request of the author. Students in the treatment group spent minimal time with the app-based technology and, not surprisingly, I find no significant difference in performance between the treatment and control groups.

Table 1. Descriptive Statistics

Baseline Characteristics	UP Academy Boston Study Participants (1)	Boston Charter Schools (2)	Boston Public Schools (3)	Massachusetts Public Schools (4)
Female	0.50	0.51	0.48	0.49
Black	0.50	0.51	0.34	0.08
Latino/a	0.33	0.35	0.41	0.17
White	0.09	0.10	0.13	0.66
Asian	0.06	0.02	0.09	0.06
Other race	0.03	0.02	0.03	0.03
Subsidized lunch	0.82	0.57	0.65	0.32
Special education	0.24	0.17	0.22	0.19
English language learner	0.23	0.10	0.28	0.07
Proficient or higher in Math	0.46	0.43	0.28	0.50
Proficient or higher in English	0.46	0.48	0.34	0.61
Math score	−0.33	–	–	0.00
English score	−0.58	–	–	0.00
N	438	8,211	28,480	578,043

Notes: This table shows student characteristics for the study participants, Boston charter schools, Boston Public Schools, and Massachusetts public schools. Study participant data comes from UP Academy records in the year of the intervention (2013–14). Math and English scores are standardized so that the state mean score for that grade is zero and the standard deviation is 1. The remaining charter and public school data come from the Massachusetts School District Profiles in 2013–14.

Math and English exams spend a larger portion of their year in the noninstructional intervention: Independent Reading. This means that for this higher proficiency level, the treatment group is being compared to an intervention that is less instructionally intensive and less Math-focused relative to the comparison for students with lower baseline levels of proficiency.

Data and Descriptive Statistics

UP Academy provided student-level data with demographics, class and teacher assignments, test scores, suspensions, and attendance. Students took an end-of-year state-standardized, summative exam[9] that tests student understanding of Common Core standards. In addition, the school administered four quarterly formative exams to assess students' progress and to adapt lesson plans.[10] Table 1 shows the demographic characteristics of the students in the study, Boston charters overall, Boston Public Schools, and Massachusetts public schools. Fifty percent of students in the study identify as black and one third identify as Latino. Representation of black and Latino students is similar to Boston charter schools' and larger than Massachusetts overall. Black students are more represented in the study than Boston Public Schools and Latino students are slightly underrepresented in the study relative to Latino representation in Boston Public Schools.

Special education students make up almost one fourth of the students in the study, slightly more than Boston charters, Boston Public Schools, and Massachusetts. English

[9] The Massachusetts Comprehensive Assessment System (MCAS).

[10] UP Academy uses the Achievement Network's formative ANet (Achievement Network) assessments.

Language Learners constitute 23 percent of the study sample. Boston Public Schools has more representation of English Language Learners (28 percent) and Boston charter schools have less (10 percent).[11]

Over 80 percent of students in the study come from economically disadvantaged families that qualify for free or reduced-price lunch. This proportion exceeds the prevalence of free or reduced-price lunch in Boston charter schools (57 percent), Boston Public Schools (65 percent), and Massachusetts overall (32 percent).

Despite the higher prevalence of economic disadvantage, a larger proportion of students in the study meet proficiency on their pre-study standardized Math and English Language Arts exam than Boston Public Schools students. Other work documents that students who apply to UP Academy Boston had low baseline test scores at the time of application. This is reflected in lower baseline test scores among new entrants to the school (sixth graders) relative to seventh and eighth graders who have already spent time in the school when they take their baseline exam. These higher proficiency rates in older grades and relative to the Boston Public Schools reflect the school's strong positive effect on lottery applicants' test scores and the likelihood that they reach proficiency (Angrist et al. 2016; Setren 2021).

While UP Academy Boston students' baseline test scores are higher than the average student in Boston Public Schools, they are lower than the state average for their grade. Table 1 shows that the average UP Academy Boston student scores 0.33 standard deviation below the state mean in Math and 0.58 standard deviation below the state mean for English.

I standardize the baseline state standardized test scores to the state mean by grade. The end-of-year and quarterly exams are centered to the school mean by grade. I only have quarterly exam data for UP Academy Boston, so I cannot standardize to a broader population. The end-of-year exam results are robust to standardizing to the state mean.

The random assignment makes it likely that students in the treatment and control groups have similar characteristics and baseline abilities. Table 2 shows no significant differences in the pre-randomization state-standardized test scores and demographics of the treatment and control groups. The *p*-value from the joint test is 0.838, which suggests that the observable characteristics in the treatment and control groups are similar.

3. EMPIRICAL FRAMEWORK

I use random assignment to the treatment group as an instrument to estimate the causal effect of the eSpark program in a two-stage least squares analysis. The second-stage equation links exposure to the treatment with outcomes as follows:

$$y_i = \alpha + \beta X_i' + \gamma T_i + \epsilon_i,$$

where y_i is the outcome of interest for student i, including test scores, attendance, grades, and behavior. The vector X_i' captures student-level characteristics, including

[11] The underrepresentation of special education and English Language Learner students could be driven by differences in classification between UP Academy and Boston Public Schools. Setren (2021) finds that Boston charter schools remove special education and English Language Learner classifications at higher rates than Boston Public Schools.

Table 2. Covariate Balance

Baseline Characteristics	Treatment Mean (1)	Control Mean (2)	Difference (3)
Female	0.467	0.505	−0.039 (0.070)
Black, Latino, or Other	0.800	0.854	−0.054 (.055)
White	0.100	0.093	0.007 (.042)
Asian	0.100	0.053	0.047 (.04)
Subsidized lunch	0.833	0.823	0.011 (.052)
Special education	0.250	0.241	0.009 (.06)
English language learner	0.183	0.235	−0.052 (.055)
Math score	−0.250	−0.340	0.090 (0.110)
English score	−0.568	−0.578	0.010 (0.105)
N	60	378	438
Joint F-test			0.838

Notes: This table shows descriptive statistics for treatment and control groups. Column 3 reports coefficients from regressions of observed characteristics on random assignment to the treatment group. Test scores are centered to the state's average score in the grade and year. *p*-values come from tests of whether all the coefficients equal zero.

grade dummies, race, ethnicity, subsidized lunch status, gender, special education status, English Language Learner status, and baseline test scores. It also includes an indicator for whether they scored lower on English relative to Math in their baseline test since this determines the first subject the treatment and control students work on.

I estimate the impact of getting assigned to the treatment group on proportion of the school year spent in the eSpark classroom in the following first-stage equation:

$$T_i = \kappa + \mu X_i' + \pi Z_i + \eta_i,$$

where T_i represents the proportion of time spent in eSpark, Z_i indicates whether student i was randomly selected for the treatment group, and π captures the effects of assignment to the treatment group on exposure to eSpark. Like the second-stage equation, the first stage includes controls for grade, demographic characteristics, and baseline test scores.[12]

The intervention occurred in 162 class sessions across three quarters of the school year. The time in eSpark variable reflects the proportion of days a student is officially enrolled in the eSpark classroom. If a student switches their class schedule in the middle of the year and as a result no longer attends the eSpark classroom, time in eSpark

[12] Because the random assignment occurred at the individual student-level, the main specifications cluster standard errors at the student-level or not at all. Abadie et al. (2017) show that it is not appropriate to cluster at the classroom-level when the assignment mechanism is not clustered at that level and the sampling process is not clustered.

Table 3. Test Score Effects

	Control Mean	First Stage	OLS	Reduced Form	2SLS
	(1)	(2)	(3)	(4)	(5)
Panel A: End-of-Year Exam (MCAS)					
Math	−0.017	0.897***	0.178*	0.181*	0.202**
		(0.017)	(0.098)	(0.094)	(0.103)
N					394
English	0.005	0.900***	−0.037	−0.025	−0.028
		(0.016)	(0.103)	(0.098)	(0.107)
N					397
Panel B: Quarterly Exams					
Math	−0.047	0.936***	0.141**	0.112	0.119
		(0.019)	(0.070)	(0.071)	(0.075)
N					1,109
English	−0.045	0.938***	0.165*	0.161*	0.172*
		(0.019)	(0.095)	(0.093)	(0.098)
N					921

Notes: This table reports the effects of the eSpark intervention on students' test scores. Column 1 displays the mean test score for untreated students. All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-test. All test scores are centered to the school's average score in the grade and year. Panel A displays estimates for the state-standardized end-of-year exam (Massachusetts Comprehensive Assessment System; MCAS). Data are at the student year level and standard errors are not clustered. Panel B shows estimates for the quarterly exam (ANet [Achievement Network]). Data are stacked at the student by quarter level, with data from the second through fourth quarters (the time of the intervention). Standard errors are clustered at the individual student level. Random assignment to eSpark instruments for the length of the program (panel A) or the proportion of time in the program for the quarter (panel B). OLS = ordinary least squares; 2SLS = two-stage least squares.

*Significant at 10%; **significant at 5%; ***significant at 1%.

equals the number of days they were enrolled in eSpark before they switched divided by the number of intervention classes leading up to the relevant outcome (e.g., the date of the exam).

Random assignment to the eSpark treatment significantly increases time spent with the eSpark app technology. Column 2 of table 3 shows that students randomly assigned to treatment have about a 90-percentage-point higher enrollment rate in the eSpark intervention classroom compared with those assigned to the control group. This reflects strong adherence to the random assignment. The few exceptions are due to the small number of students who withdrew from the school, switched class schedules, or were assigned to repeat a grade after random assignment occurred. The two-stage least squares methodology accounts for these nonrandom changes.

4. RESULTS

Test Score Effects

The experiment is powered to find substantial effect sizes on the order of 0.2 standard deviation. I find that participation in the eSpark program boosted students' end-of-year Math test scores by 0.202 standard deviation relative to the control group (see column 5 of table 3, panel A). Because test scores are standardized to the grade-level average in the school, the treatment causes students to score on average about 0.2 standard deviation higher than their peers on the Math exam.[13] The ordinary least squares (OLS) and two-stage least squares estimates are similar, suggesting that the small number of changes in enrollment after random assignment do not bias the OLS results.

The formative quarterly exam estimates also suggest positive Math gains, though the two-stage least squares results are not statistically significant. Table A.2, column 5, shows imprecise, positive point estimates ranging from 0.1 to 0.15 for the second-, third-, and fourth-quarter exams. The intervention started in the second quarter, so we expect no effect in the first quarter. To determine the average effect of the intervention on the quarterly exam, I reshape the data to the student-by-quarter level and cluster the standard errors by student. I find that the intervention generates an average 0.119 standard deviation gain on the quarterly Math exam (see table 3, panel B). This is smaller than the minimal detectable effect size of 0.18 standard deviation for significance at the 10 percent level from the power calculations. Point estimates are similar for OLS and two-stage least squares, but only the OLS estimates are statistically significant.[14]

The results for the end-of-year English exam are inconclusive. From the power calculations and the size of the standard errors, I can rule out effect sizes of 0.2 standard deviation at the 10 percent significance level in the end-of-year English exam. The quarterly formative exam results suggest that eSpark boosts English scores. Combining all quarters, the effects are 0.172 standard deviation and significant at the 10 percent level. Estimates for each individual quarter are similar, but less precise (see table A.2).[15]

While the formative and summative assessments measure the same Common Core standards, they approach measuring student progress differently. The formative assessment is low stakes for students and teachers. It serves as a diagnostic to help teachers cater lessons to the concepts that students have not yet mastered. As such, the questions cover more detail to pinpoint why students may struggle with a topic. In contrast, the summative assessment is high stakes for both the teachers and students and tests them on mastery of the material. Therefore, while the tests cover the same standards, they serve different purposes and measure different things. That can explain why the point estimates are different across the formative and summative exams for both Math and English. However, given the confidence intervals, I cannot rule out that the effect sizes are the same for both the formative and summative exams for both subjects.

To put the size of the Math effects in the context, we can compare them to lottery estimates of attending a Boston charter school, which range from 0.2 to 0.4 standard deviation (Abdulkadiroğlu et al. 2011; Angrist, Pathak, and Walters 2013; Angrist et al. 2016; Cohodes, Setren, and Walters 2021; Setren 2021). To make a direct

[13] Approximately 10 percent of the sample do not have end-of-year exam scores and therefore are excluded from this estimation. Over 65 percent of these missing scores are due to students withdrawing from the school before the end-of-year exam. The remaining missing scores come from absences or illnesses on the day of the exam. Table A.1 in the online appendix shows that attrition rates are similar in the treatment and control groups.

[14] Table A.1 in the online appendix shows that there is no differential attrition in quarterly test taking by treatment status. Similar to the end-of-year exam, the attrition is due to students withdrawing from the school or being absent on exam day. The key exception is that the fourth quarter English exam was not given to eighth graders.

[15] Test score effects are robust to various checks, including controlling for which teachers and peers students have in their core academic subjects (using classroom fixed effects) and the class size of students' core subjects.

Table 4. Behavioral Outcome Effects

	Control Mean	First Stage	OLS	Reduced Form	2SLS
	(1)	(2)	(3)	(4)	(5)
Tardy days	2.778	0.938***	−0.608	−0.890*	−0.949*
		(0.018)	(0.545)	(0.527)	(0.561)
N					1,185
In school suspension days	0.206	0.938***	−0.104**	−0.100**	−0.107**
		(0.018)	(0.043)	(0.043)	(0.046)
N					1,185
Attendance rate	85.100	0.938***	−2.884	−2.334	−2.487
		(0.018)	(1.968)	(1.886)	(1.991)
N					1,185

Notes: This table reports estimates of the effects of the eSpark intervention on behavioral outcomes. Random assignment to eSpark instruments for proportion of time in the program for the quarter. Data are stacked at the student by quarter level, with data from the second through fourth quarters (the time of the intervention). All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-est. Standard errors are clustered at the individual student level. OLS = ordinary least squares; 2SLS = two-stage least squares.

*Significant at 10%; **significant at 5%; ***significant at 1%.

comparison, I re-center the end-of-year exam scores to the state's average score in the grade and year (the same centering as the Boston charter estimates). With this re-centering, the two-stage least squares estimate for Math is 0.154 standard deviation (see table A.3). In other words, use of personalized learning technology for less than two hours a week boosted students' Math scores nearly as much an intensive intervention that changed the school model, school culture, educational services, and amount of instructional time students experienced.

Further investigation shows that the eSpark Math effects arise from gains in the treatment group and not from declines in test performance of the control group. Before the intervention, UP Academy Boston students in the sample score 0.33 standard deviation below the state mean for their grade in Math (see table 1). After the intervention, both the treatment and control groups' Math test scores surpass the state mean. The control group scored 0.145 standard deviation above the state mean (see column 1 of table A.3) and the treatment group scored 0.460 standard deviation higher than the state mean in Math. These suggest large gains on top of the substantial growth due to general school quality.[16]

Behavioral Outcomes

Next, table 4 investigates whether the education technology intervention affected student behaviors such as attendance, tardiness, or suspensions. It is possible that the presence of technology in the classroom could positively improve behavior. If students feel more motivated or engaged working with interactive technology, it could improve attendance, reduce tardiness, or lower behavioral issues, which could in turn reduce suspensions. Results suggest that participating in eSpark may lead to very small

[16] The increasing test scores of the control group are consistent with Angrist et al. (2016), who find that UP Academy Boston has a positive effect on student exam scores.

improvements in student behavior. The treatment reduced tardiness by approximately one day per quarter, though the estimates are marginally significant. The eSpark classroom reduced in-school suspensions by 0.1 day per quarter. In contrast, eSpark had a negative, noisy, and insignificant effect on attendance rates.[17]

Subgroup Effects

Table 5 shows the point estimates by gender, race, free lunch status, special education status, and English Language Learner status. The study is underpowered to detect subgroup effects or differences in effects sizes by subgroup. The estimates for the largest subgroup, students who qualify for free or reduced-price lunch, which constitute 83 percent of the sample, match the main results, while the other subgroup estimates are too noisy to be conclusive. Table 5 also shows that effects generally appear stronger for the seventh and eighth graders, though the estimates are not statistically significantly different across grades.

A subgroup analysis by baseline test scores in table 6 shows no clear evidence of differential effects across baseline ability, though the estimates are noisy. I also find no evidence of significant differential effects by the initial subject assignment (see table A.4).

5. CONCLUSION

As schools spend an increasing amount of time and financial resources on educational technology, it's important to understand the impact on student learning. This paper estimates the impact of a popular personalized tablet learning technology using a randomized controlled trial in a Boston middle school. Students in the treatment group spent twenty-eight minutes a day, four days a week for three fourths of the school year, with an adaptive iPad program that targeted the skills in which they lagged most. The treatment and control instruction provided supplemental instruction during an open block in the school day. The intervention combined personalized curriculum based on pre-testing, interactive app-based exercises and lessons, and demonstrating mastery before moving onto the next concept.

The experiment was powered to find large effects of approximately 0.2 standard deviation for Math and English. I find effects of that size for the summative Math exam, but no significant effect for English. The treatment has a marginally significant positive effect for formative English exams, but positive and not statistically significant effects for Math. Findings demonstrate the potential of technology to enhance student learning in Math. Personalized learning technology could serve as a cheaper alternative to high-intensity tutoring for school districts without funding or labor supply for extensive tutoring programs.

The demographic and baseline ability subgroup analysis yielded imprecise, mostly insignificant results, but future work with a larger sample size could be better powered

[17] To increase precision, I estimated the effect of the intervention on quarterly behavioral outcomes using the same methodology as the stacked quarterly test scores. Annual effects have the same direction and proportional magnitudes, but are less precise for tardy and in-school suspension and significant at the 5 percent level for attendance rate. The sample includes all students who did not withdraw from the school.

Table 5. Demographic Subgroup Test Score Effects

	End-of-Year Exam		Quarterly Exam	
	Math	English	Math	English
	(1)	(2)	(3)	(4)
Male	0.247 (0.159)	−0.112 (0.162)	0.119 (0.101)	0.152 (0.160)
N	203	204	552	451
Female	0.134 (0.129)	0.015 (0.138)	0.125 (0.105)	0.162 (0.117)
N	191	193	557	470
Black	0.059 (0.158)	−0.180 (0.144)	0.007 (0.109)	0.033 (0.124)
N	195	197	546	446
Latino/a	0.212 (0.185)	0.071 (0.191)	0.170 (0.131)	0.186 (0.185)
N	128	128	356	305
Free lunch	0.207^{*} (0.111)	−0.032 (0.115)	0.166^{*} (0.088)	0.220^{**} (0.110)
N	321	323	903	741
Special education	0.305 (0.243)	0.190 (0.258)	0.142 (0.185)	0.231 (0.259)
N	91	91	244	201
English language learner	0.173 (0.258)	0.128 (0.270)	0.058 (0.167)	0.087 (0.160)
N	87	87	251	206
Grade 6	−0.164 (0.222)	−0.219 (0.216)	−0.136 (0.147)	0.106 (0.176)
N	127	127	358	341
Grade 7	0.397^{**} (0.195)	0.081 (0.214)	0.311^{*} (0.161)	0.473^{***} (0.165)
N	127	128	343	334
Grade 8	0.317^{**} (0.134)	−0.051 (0.146)	0.195^{**} (0.096)	0.017 (0.119)
N	140	142	408	246

Notes: This table reports the two-stage least squares estimates of the effects of the eSpark intervention on student quarterly and end-of-year test scores by demographic subgroups. All test scores are centered to the school's average score in the grade and year. All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-test. Columns 1 and 2 show estimates for the state-standardized end-of-year exam (Massachusetts Comprehensive Assessment System; MCAS). Data for these columns is at the student year level and standard errors are not clustered. Columns 3 and 4 display estimates for the quarterly exam (ANet [Achievement Network]). Data for quarterly exam estimates are stacked at the student by quarter level, with data from the second through fourth quarters (the time of the intervention). Standard errors are clustered at the individual student level. Random assignment to eSpark instruments for proportion of time in the program for the quarter or the length of the program.

*Significant at 10%; **significant at 5%; ***significant at 1%.

to investigate distributional effects. Similar effects across the skill distribution would support the hypothesis that personalization of content is a key mechanism.

As with any educational intervention, the quality of the content and the implementation matter. This program provided supplemental instruction and interactive practice on the concepts in which students lagged most. Other programs with lower quality, worse implementation, or different features could yield different results. Future work

Table 6. Exam Effects by Baseline Test Scores

	End-of-Year Exam		Quarterly Exam	
	Math	English	Math	English
	(1)	(2)	(3)	(4)
Panel A: Above and Below Median				
Below median	0.069	−0.039	0.059	0.088
	(0.178)	(0.184)	(0.140)	(0.153)
N	190	191	516	429
Above median	0.177*	−0.009	0.073	0.219**
	(0.105)	(0.122)	(0.084)	(0.108)
N	204	206	593	492
Panel B: Terciles				
Bottom tercile	0.010	−0.005	−0.058	−0.059
	(0.210)	(0.218)	(0.185)	(0.172)
N	129	130	349	284
Second tercile	0.044	−0.158	0.105	0.172
	(0.197)	(0.233)	(0.116)	(0.222)
N	127	127	352	303
Top tercile	0.123	−0.131	0.034	0.194
	(0.109)	(0.125)	(0.099)	(0.120)
N	138	140	408	334

Notes: This table reports the two-stage least squares estimates of the effects of the eSpark intervention on student quarterly and end-of-year test scores by students' baseline test scores. All test scores are centered to the school's average score in the grade and year. All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-test. Columns 1 and 2 show estimates for the state-standardized end-of-year exam (Massachusetts Comprehensive Assessment System; MCAS). Data for these columns is at the student year level and standard errors are not clustered. Columns 3 and 4 display estimates for the quarterly exam (ANet [Achievement Network]). Data for quarterly exam estimates are stacked at the student by quarter level, with data from the second through fourth quarters (the time of the intervention). Standard errors are clustered at the individual student level. Random assignment to eSpark instruments for proportion of time in the program for the quarter or the length of the program.

*Significant at 10%; **significant at 5%.

could investigate how the impact of the program varies with different parameters. For example, the success of the program could depend on how long students are expected to focus, how long they can stay on task, and how engaging the content is. The length of time students can stay on task partially depends on their developmental stage, so the optimal time spent on the intervention could vary by age group.

Follow-up studies would ideally have detailed information about which Common Core standards students practiced and outcomes data on below-grade-level material. Students generally practiced below-grade-level Common Core standards during the intervention, but were tested on grade-level standards. Tests that include the material students practiced in the treatment and the control group would be helpful to understand the impact of the program. The lack of large effects on the summative English exam could be due to the treatment being ineffective at improving English ability. Alternatively, the treatment could have improved English ability, but the below-grade-level skills students gained did not translate to higher performance on grade-level exams.

It is important to note that this paper does not analyze replacing core instructional time with learning technology. Instead, both the treatment and control groups worked on supplemental review during the experiment's class period. The findings suggest that personalized education technology could serve as a cheaper alternative to more expensive learning supplements, such as high-intensity tutoring. Technology may also be a viable option in school districts without dense labor markets to hire a large tutoring staff. We can compare these results to an RCT of a high-intensity tutoring intervention in Chicago conducted by Cook et al. (2015). They found large effects of 0.19 to 0.31 standard deviation in Math (up to 0.50 standard deviation depending on the standardization method) for 1 hour of Math tutoring per day. The personalized learning technology has similar Math effects to the lower end of that range with about half as much time per week and less than a tenth of the cost.[18] As a result, this study suggests the promise of personalized education technology to boost learning for students, particularly those without access to other personalized education programs like tutoring.

ACKNOWLEDGMENTS

I am grateful to Joshua Angrist, Parag Pathak, and Amy Finkelstein for their guidance and support. I also thank the participants in the MIT Public Finance Lunch and MIT Labor Lunch. Thanks also to UP Academy Boston, Ryan Knight, and Jill Kahane for sharing data and sharing valuable insights. Annice Correia provided excellent administrative support. This work was supported by a National Science Foundation Graduate Research Fellowship.

REFERENCES

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeff Wooldridge. 2017. When should you adjust standard errors for clustering? NBER Working Paper No. w24003.

Abdulkadiroğlu, Atila, Joshua Angrist, Susan Dynarski, Thomas J. Kane, and Parag Pathak. 2011. Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics* 126(2): 699–748.

Angrist, Joshua, Atila Abdulkadiroğlu, Peter Hull, and Parag Pathak. 2016. Charters without lotteries: Testing takeovers in New Orleans and Boston. *American Economic Review* 106(7): 1878–1920. 10.1257/aer.p20161080

Angrist, Joshua, and Victor Lavy. 2002. New evidence on classroom computers and pupil learning. *Economic Journal* 112:735–765. 10.1111/1468-0297.00068

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. 2013. Explaining charter school effectiveness. *American Economic Journal: Applied Economics* 5(4): 1–27. 10.1257/app.5.4.1

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics* 122(3): 1235–1264. 10.1162/qjec.122.3.1235

Barrow, Lisa, Lisa Markman, and Cecilia Rouse. 2009. Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy* 1(1): 52–74. 10.1257/pol.1.1.52

[18] At the time of the study the app and licenses cost between $100 and $150 per student. The intervention also required iPad technology and a teacher to monitor the classroom. The SAGA tutoring program cost $3,800 per student in the experiment and the authors estimate it would cost $2,500 per student at scale in a large district.

Bulman, George, and Robert W. Fairlie. 2016. Technology and education: Computers, software and the internet. In *Handbook of the Economics of Education*, edited by E. A. Hanushek, S. Machin, and L. Woessmann, pp. 239–280. Philadelphia: Elsevier.

Cohodes, Sarah R., Elizabeth M. Setren, and Christopher R. Walters. 2021. Can successful schools replicate? Scaling up Boston's charter school sector. *American Economic Journal: Economic Policy* 13(1): 138–167. 10.1257/pol.20190259

Cook, Philip J., Kenneth Dodge, George Farkas, Roland G. Fryer, Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg. 2015. Not too late: Improving academic outcomes for disadvantaged youth. Institute for Policy Research Northwestern University Working Paper Series No. WP-15-01.

Escueta, Maya, Vincent Quan, Andre Joshua Nickow, and Philip Oreopoulos. 2017. *Education technology: An evidence-based review*. Working Paper No. 23744, National Bureau of Economic Research.

Goolsbee, Austan, and Jonathan Guryan. 2006. The impact of internet subsidies in public schools. *Review of Economics and Statistics* 88(2): 336–347. 10.1162/rest.88.2.336

Huang, Linn. 2016. Managed CloudView Survey Report. International Data Corporation U.S. Educational Devices.

Imbens, Guido W. 2011. Experimental design for unit and cluster randomized trials. Paper presented at Initiative for Impact Evaluation. Cuernavaca, Mexico, 15–17 June 2011. http://cyrussamii.com/wp-content/uploads/2011/06/Imbens_June_8_paper.pdf.

Machin, Stephen, Sandra McNally, and Olmo Silva. 2007. New technology in schools: Is there a payoff? *Economic Journal* 117(522): 1145–1167. 10.1111/j.1468-0297.2007.02070.x

Morgan, Karl Lock, and Donald B. Rubin. 2012. Rerandomization to improve covariate balance in experiments. *Annals of Statistics* 40(2): 1263–1282. 10.1214/12-AOS1008

Muralidharan, Karthik, Abhijeet Singh, and Alejandro Ganimian. 2019. Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review* 109(4): 1426–1460. 10.1257/aer.20171112

Rouse, Cecilia, and Alan Krueger. 2004. Putting computerized instruction to the test: A randomized evaluation of a "scientifically based" reading program. *Economics of Education Review* 23(4): 323–338. 10.1016/j.econedurev.2003.10.005

Setren, Elizabeth M. 2021. Targeted vs. general education investments: Evidence from special education and English language learners in Boston charter schools. *Journal of Human Resources* 56(4): 1073–1112. 10.3368/jhr.56.4.0219-10040R2

Software & Information Industry Association. 2015. *Education Technology Industry Network of SIIA*. Available https://www.ena.com/affiliates/education-technology-industry-network-of-siia/. Accessed December 2015.

APPENDIX

Table A.1. Attrition

	Control Mean	Attrition Differential by Treatment Status
	(1)	(2)
Quarterly Math score	0.160	−0.016 (0.041)
N		1,314
Quarterly English score	0.289	0.039 (0.043)
N		1,314
End-of-year Math score	0.101	0.006 (0.048)
N		438
End-of-year English score	0.098	−0.032 (0.047)
N		438

Notes: This table reports the two-stage least squares estimates of the effects of the eSpark intervention on attriting from the sample. All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-test. Quarterly exam estimates use data stacked at the student by quarter level, with data from the second through fourth quarters (the time of the intervention). Standard errors are clustered at the individual student level. End-of-year exam (Massachusetts Comprehensive Assessment System; MCAS) estimates use data at the student year level and standard errors are not clustered. Random assignment to eSpark instruments for proportion of time in the program for the quarter or the length of the program.

Table A.2. Quarterly Exam Estimates

	Control Mean	First Stage	OLS	Reduced Form	2SLS
	(1)	(2)	(3)	(4)	(5)
			Panel A: First Quarter		
Math	−0.008	0.986^{***} (0.010)	0.030 (0.103)	0.043 (0.104)	0.043 (0.103)
N					395
English	0.032	0.984^{***} (0.011)	−0.180 (0.123)	−0.177 (0.123)	−0.180 (0.123)
N					378
			Panel B: Second Quarter		
Math	−0.024	0.985^{***} (0.011)	0.115 (0.094)	0.102 (0.095)	0.103 (0.094)
N					381
English	−0.016	0.983^{***} (0.012)	0.075 (0.125)	0.059 (0.126)	0.060 (0.126)
N					366
			Panel C: Third Quarter		
Math	−0.062	0.927^{***} (0.014)	0.158 (0.116)	0.136 (0.112)	0.147 (0.118)
N					365
English	−0.059	0.917^{***} (0.015)	0.220 (0.136)	0.204 (0.130)	0.223 (0.139)
N					356

Table A.2. Continued.

	Control Mean	First Stage	OLS	Reduced Form	2SLS
	(1)	(2)	(3)	(4)	(5)
		Panel D: Fourth Quarter			
Math	−0.055	0.896*** (0.017)	0.171 (0.115)	0.098 (0.109)	0.109 (0.119)
N					363
English	−0.074	0.884*** (0.026)	0.306 (0.208)	0.300 (0.198)	0.340 (0.217)
N					199

Notes: This table reports the two-stage least squares estimates of the effects of the eSpark intervention on student quarterly test scores. Random assignment to eSpark instruments for the proportion of time in the program for the quarter. Test scores are centered to the school's mean for the grade and year. All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-test.

***Significant at 1%.

Table A.3. Effects on State Centered Test Scores

	Control Mean	First Stage	OLS	Reduced Form	2SLS
	(1)	(2)	(3)	(4)	(5)
Math	0.145	0.897*** (0.017)	0.136* (0.081)	0.138* (0.077)	0.154* (0.085)
N					394
English	−0.251	0.900*** (0.016)	−0.027 (0.091)	−0.014 (0.087)	−0.016 (0.095)
N					397

Notes: This table reports the effects of the eSpark intervention on students' test scores. Column 1 displays the mean test score for untreated students. All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-test. All test scores are centered to the state's average score in the grade and year. Data are at the student year level and standard errors are not clustered. Random assignment to eSpark instruments for proportion of time in the length of the program.

*Significant at 10%; ***significant at 1%.

Table A.4. Test Score Effects by Initially Assigned Subject

	Control Mean	First Stage	OLS	Reduced Form	2SLS
Initial Assignment – Subject of Exam	(1)	(2)	(3)	(4)	(5)
		Panel A: End-of-Year Exam (MCAS)			
Math Group – Math Score	−0.377	0.874*** (0.025)	0.056 (0.145)	0.077 (0.135)	0.089 (0.150)
N					187
Math Group – ELA Score	−0.042	0.874*** (0.025)	0.002 (0.148)	0.057 (0.138)	0.065 (0.153)
N					187
ELA Group – ELA Score	0.049	0.923*** (0.022)	−0.096 (0.143)	−0.147 (0.138)	−0.160 (0.146)
N					210

Table A.4. Continued.

	Control Mean	First Stage	OLS	Reduced Form	2SLS
Initial Assignment – Subject of Exam	(1)	(2)	(3)	(4)	(5)
		Panel A: End-of-Year Exam (MCAS)			
ELA Group – Math Score	0.311	0.921^{***}	0.197	0.188	0.204
		(0.023)	(0.129)	(0.126)	(0.132)
N					207
		Panel B: Quarterly Exams			
Math Group – Math Score	−0.379	0.916^{***}	−0.003	0.000	0.000
		(0.031)	(0.111)	(0.113)	(0.122)
N					514
Math Group – ELA Score	−0.077	0.915^{***}	0.002	0.067	0.074
		(0.032)	(0.137)	(0.127)	(0.137)
N					419
ELA Group – ELA Score	−0.017	0.959^{***}	0.251^{*}	0.204	0.213
		(0.016)	(0.133)	(0.134)	(0.137)
N					502
ELA Group – Math Score	0.244	0.956^{***}	0.190^{**}	0.135	0.141
		(0.016)	(0.093)	(0.094)	(0.096)
N					595

Notes: This table reports the effects of the eSpark intervention on students' test scores by which subject they were initially assigned to focus on. Column 1 displays the mean test score for untreated students. All models control for gender, ethnicity, grade, free or reduced-price lunch status, English Language Learner status, special education status, and baseline test scores. Models also control for whether the student performed relatively lower on the English pre-test compared to the Math pre-test. All test scores are centered to the school's average score in the grade and year. Panel A displays estimates for the state-standardized end-of-year exam (Massachusetts Comprehensive Assessment System; MCAS). Data is at the student year level and standard errors are not clustered. Panel B shows estimates for the quarterly exam (ANet [Achievement Network]). Data are stacked at the student by quarter level, with data from the second through fourth quarters (the time of the intervention). Standard errors are clustered at the individual student level. Random assignment to eSpark instruments for the length of the program (panel A) or the proportion of time in the program for the quarter (panel B). ELA = English Language Arts; OLS = ordinary least squares; 2SLS = two-stage least squares.

*Significant at 10%; **significant at 5%; ***significant at 1%.