The Effects of Educational Apps on Student Achievement and Engagement

Maya Lopuch*

April 2013

Abstract

The iPad and other tablets have democratized the development of educational content. While some of these development efforts have created substantive advances in teaching, others have fallen short. Identifying high quality apps is growing increasingly difficult. This paper uses a unique database of over 140,000 observations of 663 educational apps to determine whether app-based curricula measurably impact student achievement and engagement. Curated bundles of educational apps are associated with robust achievement gains. Students who supplement traditional classroom education with an app-based curriculum achieve 165% of their expected learning gains. Students enjoy the majority of the app-based content, even though it usually targets their academic weakness. There is a moderate positive relationship between achievement gains and student engagement, suggesting that these learning gains are sustainable. The cost of apps has no observable impact on either achievement or engagement.

*Maya Lopuch is a Data Scientist at eSpark Learning. Prior to joining eSpark, Maya was a researcher at the Spencer Foundation and the Harvard Graduate School of Education. Her research has investigated how public schools impact long-term student outcomes. Maya holds a Bachelor in Economics from Stanford University and a Master in Public Policy from the University of Chicago.

1 Introduction

It is now easier than ever to distribute educational content. Whereas previously only large-scale textbook publishers could distribute educational resources to wide audiences, now teachers, parents, and hobbyists can develop new content and distribute it digitally. The introduction of the iPad has led to a dramatic acceleration of this trend. The iPad platform allows anyone to develop content and distribute it through the App Store. Much of this content has focused on education. Within three years of its introduction, the number of distinct educational resources available for the iPad has grown to number over 100,000. Access to this content is growing rapidly. By early 2013, Apple has sold over 8 million iPads directly to educational institutions.

While some of these development efforts have created substantive advances in teaching, others have fallen short. With so many educational resources available for the iPad alone, a challenge for educators is to curate content smartly. Educators must often choose one app out of many that proclaim to teach similar skills. Unlike traditional textbooks, the content in the App Store is decentralized and dynamic.

Teachers, administrators, and parents have few resources to identify high-quality apps. The rate at which new content is introduced is also outpacing educators' abilities to identify which apps are currently most effective. This paper addresses both of these deficiencies. The goal of this work is to answer two questions: Are there high quality apps that can measurably affect student achievement? If there are, to what extent are observable app characteristics predictive of achievement growth and engaging student experiences?

New data show that educational apps are linked to compelling increases in student achievement. The average app-based curriculum analyzed in this paper increases achievement by 165% of expected student growth. This result is likely to represent a lower bound of the true effect because achievement growth estimates are measured conservatively. The data also show that apps that are more academically effective are correlated with higher engagement ratings. Students find effective educational apps fun, suggesting that these learning gains are sustainable. Interestingly, the cost of apps has no observable impact on either achievement or engagement.

These conclusions come from the analysis of a unique database of over 140,000 observations of 663 iOS apps. Data on several thousand students are observed as they progress through app-based learning curricula that supplement traditional classroom work. Achievement gains are measured with students' pre and post scores on a rigorous, nationally normed assessment. Student engagement is estimated with students' own approval ratings of each app. Students' own ratings of apps capture a measure of student engagement that is not available to teachers or administrators on a large scale. By aggregating thousands of students' own approval ratings for hundreds of apps and matching those patterns to achievement growth, this paper provides robust estimates of the academic effectiveness and engagement of educational apps.

2 Data

All data were collected from the internal database of eSpark Learning between October 2012 and January 2013. eSpark Learning creates personalized learning curricula using iPad-based instructional videos, third-party educational apps, and assessment tools. Each learning curriculum is tailored to a specific domain and grade level aligned with the Common Core State Standards (CCSS). Pedagogical experts select all components of these curricula, including the third-party educational apps. The apps analyzed in this sample do not represent a random sample of educational apps, but rather a highly curated set of apps that are judged by educational professionals as high quality. Public, private, and charter schools partner with eSpark to obtain access to these curricula, the associated apps, and technical and professional support.

Instead of working on a comprehensive curriculum, eSpark students focus on one or two goals within the set of Common Core domains. At the beginning of the academic year eSpark diagnoses each student's strengths and weaknesses using schools' existing assessment data. eSpark then recommends personalized learning goals for each student. The goal recommendations usually target students' existing weaknesses, but in some cases recommended goals target content ahead of grade level. Teachers and students then jointly review the eSpark recommendations and choose each student's goals. In the majority of cases, teachers and students use the eSpark recommended goals.

After teachers and students finalize their goals, students receive iPads loaded with curricular content specific to the Common Core domain and grade level that they have chosen. Students access their personalized learning curriculum through the eSpark iPad app. The majority of eSpark usage occurs during school hours under the supervision of a teacher. Usage patterns vary by school and teacher preferences. Some schools use eSpark for 20 minutes per day, five times per week, whereas others may use eSpark for one hour per day, two times per week.

When students log into the eSpark app, they progress through dozens of videos, third-party app-based challenges, and assessments to master their Common Core domain goal. For example, a third grade student might access curricular content on first grade Reading Foundational Skills, his weakest area. In an eSpark session, he would watch videos about word sounds and do activities in third-party apps like Phonics Awareness and Twinkl Phonics. The student would continue to work on phonics skills and complete written and video assessments before progressing to a subsequent skill like spelling.

The achievement analyses in this paper are estimated using a subset of students for whom eSpark received Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessment data in the fall and winter of the 2012-2013 academic year. The NWEA MAP is a computer adaptive assessment. The MAP is not a high-stakes test, but rather interim assessment that is designed to provide information about students' academic progress and guide classroom instruction. School administrators voluntarily provided eSpark with this data. All fall NWEA assessments were completed before students

Figure 1: A view of the eSpark app

Figure 2: Students rate third-party apps

began using the eSpark curricula.

The NWEA MAP data includes students' overall Rasch Unit (RIT) scores for math and reading as well as percentile rankings based on the national sample of all MAP test takers. The overall RIT score reflects a weighted average of students' understanding of Common Code domains within that discipline. RIT scores for a given student are expected to increase over time as students learn more material. Based on each student's fall RIT score and grade level, NWEA determines the number of RIT points the student is expected to achieve over the course of the academic year. Students with below average baseline scores are expected to achieve more RIT points than are students with higher baseline scores. Since NWEA's growth expectations are provided for the full academic year and the post-test in this analysis occurs in the winter, the growth expectations are divided by two to reflect expectations for one semester.

The engagement analyses in this paper are estimated using the larger sample of all eSpark users. Students' own ratings of third-party educational apps are used as a proxy for engagement. After students complete an app-based challenge, students are asked to rate the activity using a thumbs up or a thumbs down icon (see Figure 2). Students must rate each activity in order to continue.

3 Summary Statistics

Table 1 summarizes the composition of the dataset. The achievement sample contains data on 1,630 students who completed both pre- and post-tests in mathematics and 1,797 students who completed pre- and post-tests in reading. Most students in the sample are in early elementary school. The app sample includes 233 distinct math educational apps and 304 reading and language educational apps. The sample excludes data on 126 apps that are part of eSpark curricula but were used by fewer than 30 students before January 2013.

The remaining apps were accessed by an average of 263 students for a total of 141,338 app observations. The average student completed activities using 49.7 distinct educational apps during the data collection period.

Table 1: Distribution of students and apps by subject and grade level

Group	Apps	App completions	Students in test sample
Subject			
Mathematics	233	55,020	1,630
ELA	304	86,318	1,797
Grade			
PK	17	9,177	-
K	109	70,516	2,502
1	84	16,795	185
2	76	8,903	90
3	74	9,964	108
4	93	10,577	131
5	38	6,005	116
6	19	2,953	69
7	27	6,448	226

Table 2: Summary statistics of test sample

Metric	50th Percentile	N Students
Fall NWEA Score	148	3427
Winter NWEA Score	159	3427
Expected Growth	7.5	3427
Fall Percentile	51	3427
Winter Percentile	60	3427

Summary statistics on NWEA achievement data are shown in Table 2. In the fall of 2012, eSpark students scored slightly better than the national sample: a student in the 50th percentile of the eSpark distribution placed in the 51st percentile of the national distribution. In the winter of 2012-2013, after the students in the sample had started using eSpark, the distribution of test scores strongly outperformed national estimates. The median eSpark user placed in the 60th percentile of the national winter distribution.

Table 3: Frequencies of students beginning each eSpark curricular unit

Curricular Unit	K	1	2	3	4	5	6	7	Total
Counting and Cardinality	994	-	-	-	-	-	-	-	994
Geometry	159	32	11	-	-	-	-	-	202
Measurement and Data	145	27	19	10	13	-	-	-	214
Number and Operations in Base Ten	-	44	25	27	22	12	-	-	130
Number and Operations Fractions	-	-	-	11	10	-	-	-	21
Operations and Algebraic Thinking	127	42	16	14	19	-	-	-	218
Reading Foundational Skills	1,009	65	30	28	-	-	-	-	1,132
Reading Informational Text	96	21	13	22	38	48	27	88	353
Reading Literature	121	34	21	34	64	67	48	175	564
Total	2,651	265	135	146	166	127	75	263	3,828

Table 4: Quantitative app attributes

Attribute	Mean	Std. Dev.	Min	Max
Rating (%)	0.80	0.11	0.37	0.97
Duration (min)	9.9	2.5	6.2	21.8
Price ($)	2.17	1.82	0.00	9.99

Table 5: App price detail

Price Detail ($)	N apps	Percent
0	81	15.1
0.99	152	28.3
1.99	117	21.8
2.99	83	15.5
3.99	33	6.2
4.99 +	71	13.2
Total	537	100

Table 3 summarizes the content and grade level of eSpark curricular units. Each unit is closely aligned with the standards of the Common Core domain by the same name. The counts within each cell reflect the number of students linked to NWEA data that began the curricular unit in the fall of 2012. Within each curricular unit, students are exposed to the same content but progress through the content at their own pace.

Tables 4 and 5 display summary statistics of the quantitative characteristics of the 537 apps in the analysis sample. Students overwhelming liked the apps in the sample. The average app received an 80% approval rating. Students spent an average of 10 minutes on activities within the third-party apps. The average price of apps in the sample was just over two dollars, but apps ranged from free to $9.99.

4 Results on student achievement

Figure 3 visually shows how the distribution of NWEA results differed before and after eSpark usage. The black line is the kernel density of fall NWEA percentile scores in both math and reading. If the sample of eSpark students were perfectly representative of the national sample, one would expect to observe the same mass of students at each percentile ranking. The dashed line represents this uniform distribution. The black line shows that in the fall of 2012, eSpark students were disproportionately likely to score in the middle of the national distribution. Although eSpark partners with schools throughout the achievement distribution, the students in this NWEA sample tend to have average baseline test scores.

If eSpark usage had no effect on student achievement, one would expect the winter distribution of percentile scores to map closely to the fall distribution: when all students experience expected learning growth, their percentile rankings remain the same. This is represented by the hypothetical distribution shown in blue in the left panel of Figure 3. The right panel of Figure 3 shows the true observed distribution of winter scores in orange. The observed distribution shows that eSpark students experienced a marked difference in their achievement trajectories. Winter percentile scores among eSpark students dramatically shifted to the right of the initial distribution. The average eSpark student increased her national ranking by nine percentile points within one semester.[1]

The magnitude of this effect is quite large. These results are especially dramatic because students who use eSpark are exposed to curricular content that focuses on a subset of the items tested in the NWEA assessments. The large effects on discipline-level achievement suggest that eSpark curricular content creates positive spillovers to other academic domains.

Careful readers might note that the results shown in Figure 3 could be attributed to a school effect instead of a curriculum effect. Proactive administrators or an innovative learning culture within the schools that have partnered with eSpark might instead drive these

[1]Future research will investigate heterogeneity in achievement impacts. Preliminary work shows that achievement gains are significantly higher among students from the lowest tercile of baseline distribution.

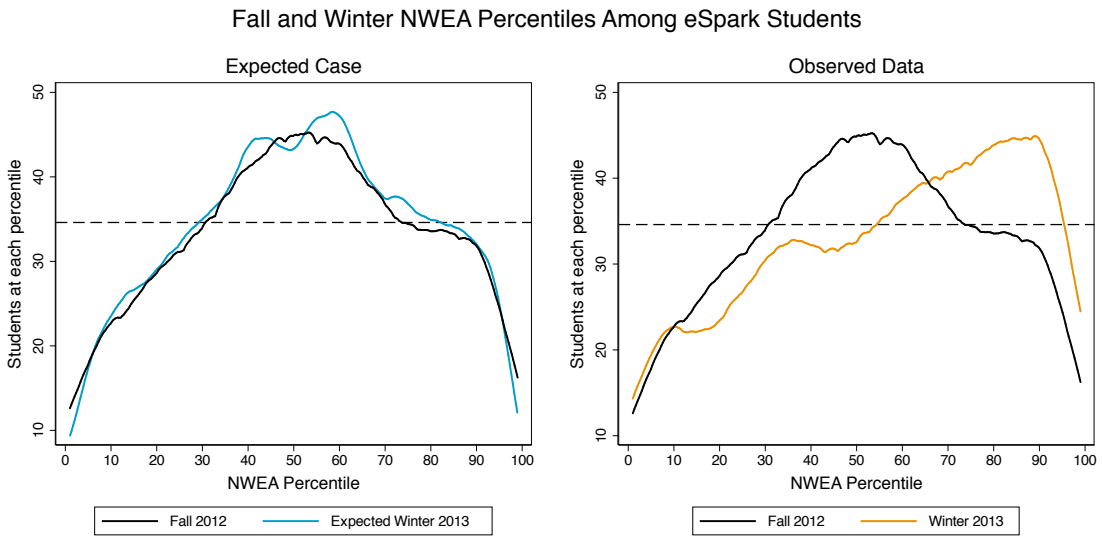

Figure 3: Observed post test scores are markedly higher than baseline scores.

positive achievement gains. One way to test this alternative hypothesis is to compare the pre and post NWEA percentile distributions among students within eSpark schools who do not use eSpark. eSpark does not collect this data from all of its partner districts, but one district did provide data on 1,656 students who did not participate in the eSpark program. The average nonuser student in an eSpark partner school increased her national ranking by 1.9 percentile points over the same time period. Using this group as a benchmark, these results suggest that about 20% of the overall increase in percentile rankings can be attributed to a school effect and 80% of the increase is associated with the app-based curriculum.

Another way to estimate achievement growth is to divide each student's difference between her winter and fall RIT scores by the semester-adjusted expected growth estimate provided by NWEA. For example, if a student had a RIT score of 180 in the fall, 182 in the winter, and her expected semester growth was 2 RIT points, her growth score of 1 would indicate that she is on target to meet her growth goals by the end of the school year. Growth scores that are greater than 1 indicate that students are learning more material than expected, and growth scores less than 1 indicate that students are not on track to meet annual academic goals. A growth score of less than 0 indicates that students have shown understanding of fewer concepts in the winter than they did in the fall.

A growth estimate for each curricular unit listed in Table 3 can be estimated by computing the average growth score among all students who began that unit. Figure 4 plots these growth estimates. The horizontal axis shows the cost of the third-party apps that

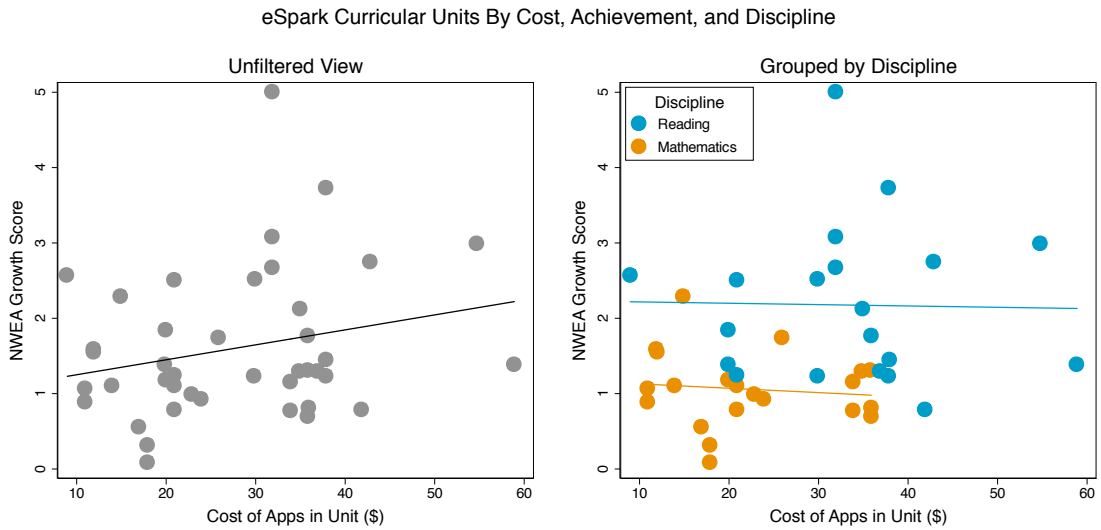

Figure 4: Once discipline is accounted for, there is no relationship between cost and achievement growth.

comprise the curricular units. Each unit shows positive average growth, and most produce growth results that outpace nationally normed expectations. Out of 41 curricular units, 30 have average growth scores of more than 1, and no units have growth scores less than zero. Despite these overall positive results, there is substantial variation. Some units yield growth estimates that are five times larger than others in the sample.

The sum cost of app bundles within each curricular unit also varies substantially, ranging from $9 to $59. At first glance, there appears to be a strong positive relationship between achievement growth and cost (left panel of Figure 4). The positive relationship disappears when fit lines are drawn within discipline (right panel of Figure 4). Although reading curricula are associated with higher achievement growth than are math curricula, reading units also have systematically higher costs. Within subject, the correlation between cost and growth is statistically indistinguishable from zero.

Using the growth metrics described above, the mean growth score across all curricular units is 1.65. This indicates that students assigned to the average performing curricular unit achieved 65% more than they were expected to achieve in the first semester of the school year. In other words, students had achieved 83% of their annual goal before January. Most eSpark students are on track to dramatically exceed academic expectations by the end of the school year.

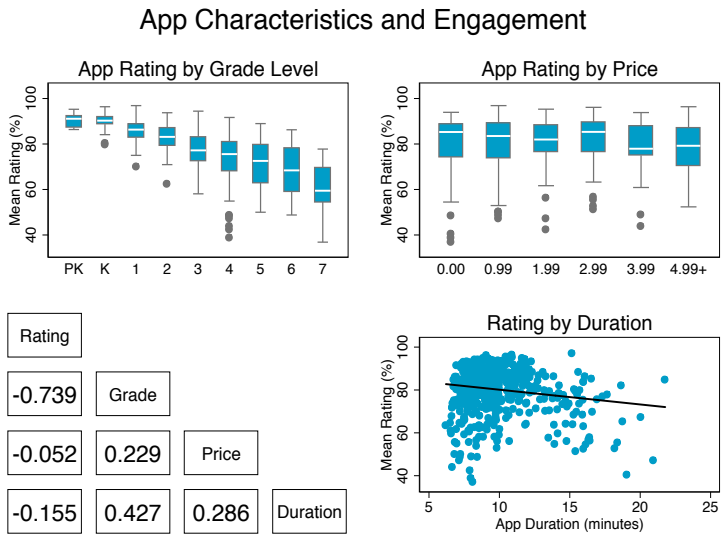

Figure 5: Grade level is a stronger predictor of app ratings than price and duration.

5 Results on student engagement

Figures 3 and 4 have established that iPad-based educational content has large effects on student achievement, and these effects are unrelated to cost. This next section investigates how educational apps impact student engagement. While achievement effects are estimated using bundles of apps sequenced together in an educational curriculum, engagement effects can be estimated using individual apps.

How do the quantitative characteristics of educational apps correlate with student engagement? Figure 5 summarizes the relationships between ratings, grade level, price, and duration of app activities. The grade level of app content has a strong negative effect on students' ratings ($\rho = 0.739$, p-value$<$0.01). The boxplot in the upper left quadrant visualizes this relationship. Apps in the 25th percentile of the second grade distribution receive higher approval ratings than apps in the 75th percentile of the seventh grade distribution. Older students tend to rate iPad-based activities lower than do younger students.

The upper right quadrant of Figure 5 shows that the price of apps does not have a discernible effect on student ratings. The correlation coefficient of -0.052 is not significantly different from zero. While there is little distinguishable difference in average quality, there is more variation in the ratings of less expensive apps, particularly on the low end of the ratings distribution. Less expensive apps are disproportionately likely to be less engaging. Among the ten apps with average approval ratings less than 50%, six are priced below $1, and eight are priced below $2.

The lower right quadrant of Figure 5 shows a weak negative relationship between app

ratings and mean amount of time students spend on apps. High-leverage outliers on the upper end of the duration distribution heavily influence this pattern, and these apps almost exclusively cover seventh grade material. Upon restricting older students from the sample, the correlation between ratings and duration becomes statistically indistinguishable from zero at -0.009.

Overall, Figure 5 shows that grade level is the strongest predictor of student ratings among the characteristics measured in this sample. Grade level alone explains more than half of the variance in students' ratings of apps. Perhaps surprisingly, average student ratings are insensitive to the cost of content.

6 The relationship between achievement and engagement

Educators seek curricula that are both effective and engaging in order to sustain students' interest and produce learning gains over long periods of time. This next section investigates whether more academically effective app content is associated with greater student engagement. Figure 6 plots curricular units by their growth and share of apps that were rated positively by students. Each curricular unit shows positive average achievement growth and the majority of the apps within the unit are rated positively. There is a moderately positive relationship between achievement growth and engagement. Curricular units that have higher growth results tend to have higher student ratings. The black line represents this bivariate relationship ($\rho = 0.367$, $p<0.05$). Even if the high leverage outlier is excluded, the correlation remains significant ($\rho = 0.300$, $p<0.10$).

That students highly enjoy more academically effective content is an encouraging result for the future of app-based curricula. This result may be especially surprising given that students in this sample most often work on content that targets their existing academic weaknesses. Skeptical readers may ask if this relationship is driven by selection bias: even if all curricula were equally effective, we could observe this positive relationship if students put in more time and mental effort to material they liked. This alternative explanation seems unlikely. Students tend to spend the same amount of time on app-based activities whether they rate the app with a thumbs up or a thumbs down.

Figure 6 also shows how cost influences the relationship between achievement growth and student engagement. Earlier results failed to show significant relationships between cost and achievement and between cost and engagement. Unsurprisingly, Figure 6 also fails to show that cost interacts with the relationship between achievement and engagement. This finding is encouraging to stakeholders who seek to maximize the effectiveness of educational apps. One does not necessarily need to spend more money to yield stronger achievement and engagement results.

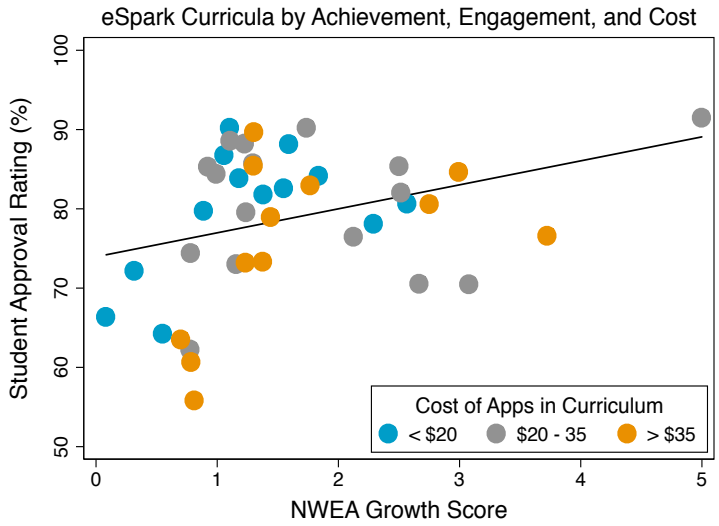

Figure 6: There is a positive relationship between achievement growth and student ratings, but cost is unrelated to either of these outcomes.

7 Discussion

This paper uses a large, unique dataset to identify whether app-based educational curricula produce measurable achievement and engagement results. Achievement results are robust. Students who accessed an app-based curriculum grew on average nine percentile points on a nationally normed assessment. Results on engagement are also compelling. Despite the fact that most students focused on their academically weakest area, students approved of 80% of the app-based content.

Although selection into these app-based curricula is not random, the effectiveness evidence presented in this report is likely to underestimate true effects. Students disproportionately completed educational activities on content with which they have previously struggled. This suggests that achievement gains and engagement may be higher if students worked on topics for which they have existing proclivities. Moreover, achievement growth estimates are measured conservatively. The growth estimates used in this paper are derived from discipline-specific test scores. eSpark curricula focus only on a subset of this material. For example, a given student may work only on fractions content in eSpark, but his NWEA math score will also assess his knowledge of algebraic thinking and measurement. Growth scores that are based only on content which students explored in eSpark would likely be higher. Additionally, these estimates count all students who began work on any of these curricula, regardless of how much content they complete or how teachers were trained on the curricula. Future research will investigate how usage and implementation patterns af-

fect student outcomes.

While this report has found that app-based curricula can have large impacts on student outcomes, this does not imply that all educational apps are equally successful. One important caveat is that the data analyzed in this report do not represent a random sample of educational apps. All apps included in this dataset were hand-selected by pedagogical experts who specialize in curriculum design. These experts chose apps that they deemed to be best of breed within the set of apps that cover the same educational standard. The results described in this paper are more likely to hold true for apps that have been previously screened for quality as opposed to a random selection of educational apps.

This paper concludes with an optimistic assessment of the future of app-based curricula. Educational apps have demonstrated large and positive impacts on student outcomes. There is, of course, more work to be done. While student achievement and engagement are important outcomes, but they do not represent a complete picture of educational quality. Parents, teachers, and administrators are interested in a multitude of other student outcomes such as persistence and self-control. App-based curricula also raise important questions of equity, particularly in the context of the Digital Divide. Future research using the eSpark database will explore these and other effects of digital educational content.