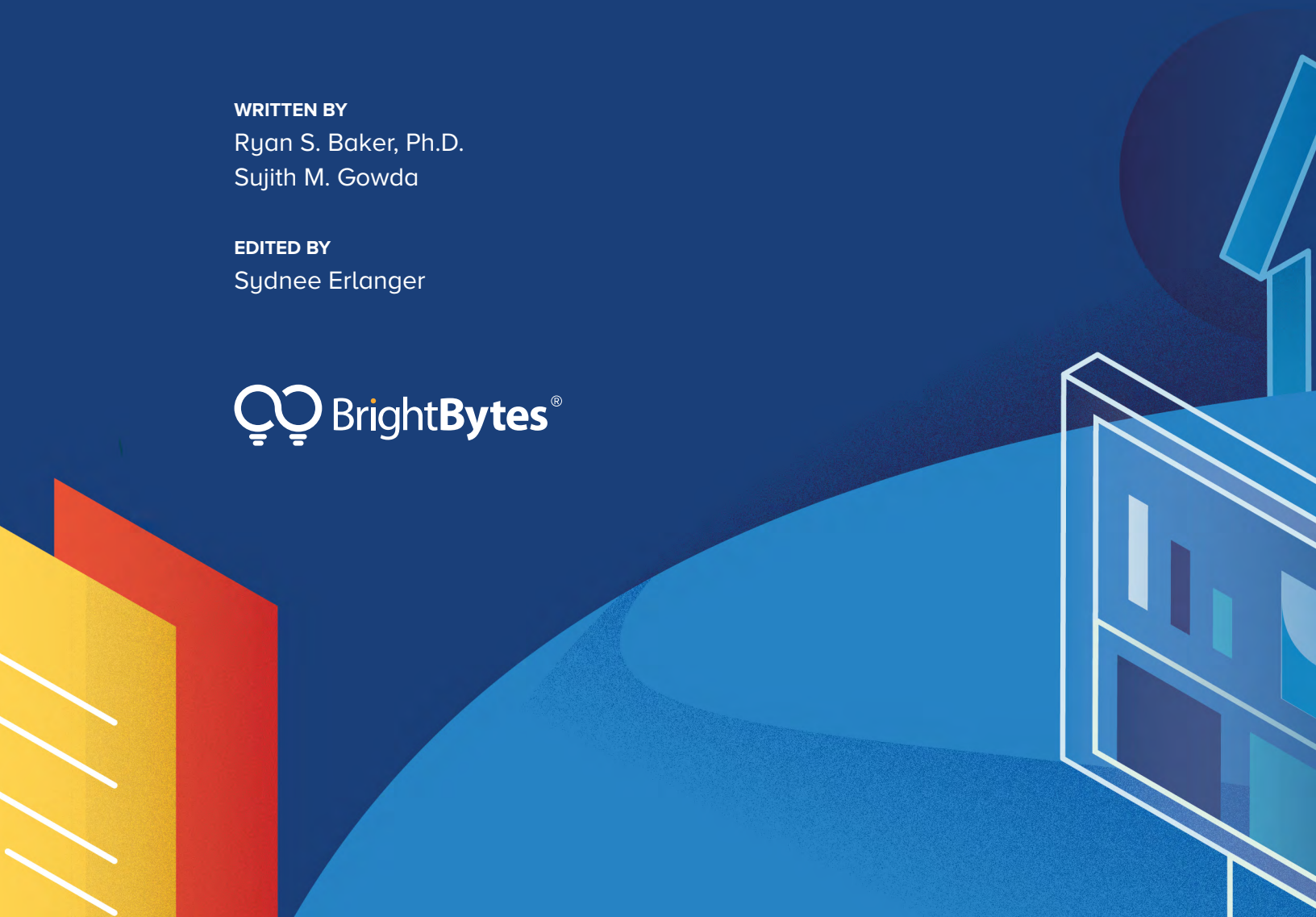
The 2018 Technology & Learning Insights Report:

Towards Understanding App Effectiveness and Cost

WRITTEN BY
Ryan S. Baker, Ph.D.
Sujith M. Gowda

EDITED BY
Sydnee Erlanger

BrightBytes®



EXECUTIVE SUMMARY

With access to an ever-expanding catalog of over 2,500 education apps, educators have the opportunity to adopt new approaches for learning and broaden horizons for all students. However, amid growing budget constraints and increasingly aggressive levels of accountability for student improvement, decision makers struggle to strategically invest time and resources into the programs that demonstrate the greatest impact on learning.

To effectively execute this work, educators must have authentic quantitative and qualitative data on the conditions under which ed tech apps impact student learning. Until recently, there has been limited high-quality research on ed tech products, and the prohibitive cost and constraints for control groups make this research difficult to conduct in education environments.

Fortunately, recent advances in technology and education data mining may provide a solution to this. Although there are early tools that can offer summary data on areas such as app cost and usage, educators know that this information alone cannot determine effectiveness. In fact, in 2018, EdWeek MarketBrief published findings from a survey of over 500 district leaders in which administrators expressed most interest in ed tech data about students academic progress, "While data on students' academic achievement and deficits are in high demand, administrators said that information about the usage of ed-tech products–including how many times a teacher or student has logged in to the tool–is less important." Simply put, usage and cost data cannot answer the most critical question: **Is this app having a positive impact on student learning?**

This study used data about ed tech usage and cost captured by the BrightBytes Learning Outcomes module, and integrated that data with standardized test scores for math, science, and ELA, improving between one test (fall/winter) to another test (winter/spring), to measure impact.

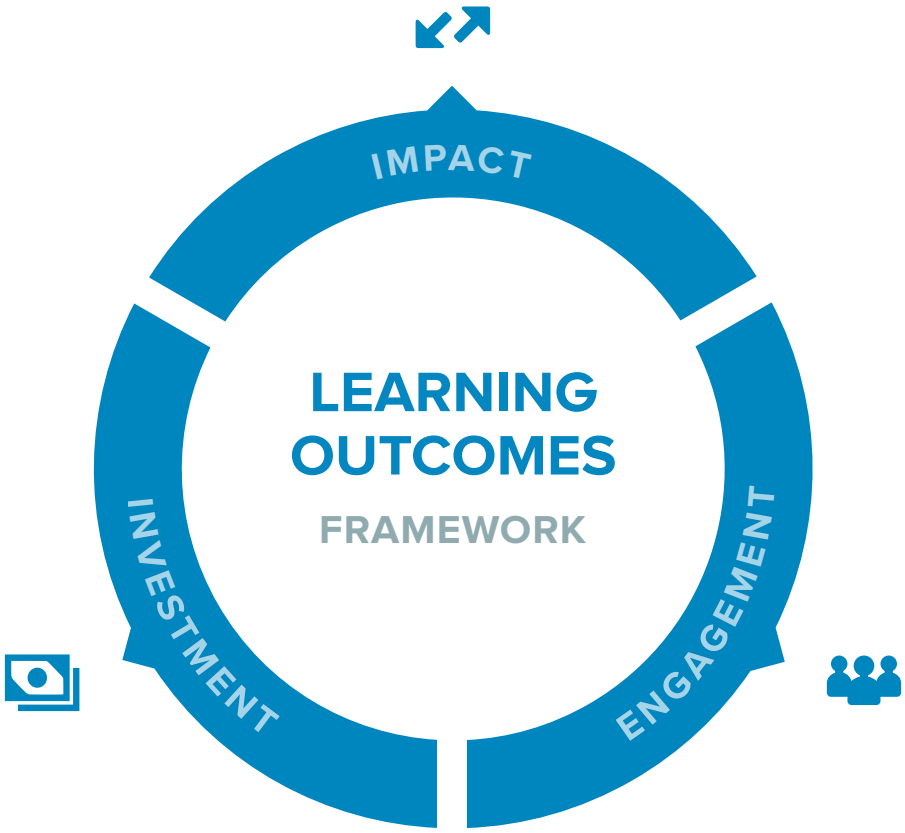
The dataset included:

392,603
STUDENTS

48
DISTRICTS

1.4M
HOURS OF USAGE

150
STUDENT-FOCUSED ED TECH PROGRAMS



The Learning Outcomes module, one of three modules in the BrightBytes 21st Century Learning suite, captures ed tech data via each district's existing web proxy and employs advanced analytics to correlate usage with student achievement data. The module's research-based framework measures data across three domains:

Investment- tracks the subscription, implementation, and maintenance costs of each app.

Engagement- measures student usage (how often students are using a particular app and to what degree) and student perceptions (how much do students like/dislike a particular app).

Impact- investigates the link between student usage in a particular app and how that may or may not influence their performance on assessments.

Results are delivered across intuitive reports and dashboards that allow educators to truly understand program impact in a highly-engaging way that is both actionable and accessible. With this knowledge, educators can achieve on-demand access to knowledge about which programs are having a positive impact, and under what circumstances. This can inform strategic decisions about which programs to scale and adopt, and how to adjust resource allocations.

By analyzing the aggregate data of districts using the Learning Outcomes module, we were able to take this data one step further to examine national trends. In this report, we will discuss the results of a study performed using data from the BrightBytes platform. The study analyzed which apps are receiving usage in line with license purchases, how much usage is occurring, and which apps are more and less effective for learning outcomes for math, science, and ELA. We will also discuss findings related to a deep dive on the apps with the highest effectiveness, and analyze the contexts where they are producing positive impact.

TABLE OF CONTENTS

ABSTRACT

Schools use an increasing number of learning apps, but there is limited evidence on which apps work, and even less evidence on which apps work for which settings. In this report, we will discuss the results of a study performed using data from the **BrightBytes platform**. The study analyzed which apps are receiving usage in line with license purchases, how much usage is occurring, and which apps are more and less effective for learning outcomes for math, science, and ELA. We will also discuss findings related to a deep dive on the apps with the highest effectiveness, and analyze the contexts in which they are producing positive impact.

INTRODUCTION

In recent years, the number of learning apps and webpages used at schools in the United States has skyrocketed, with EdSurge's community-driven database of ed tech products listing over 2,500 apps (https://www.edsurge.com/product-reviews). However, the evidence for which apps are effective has not expanded at nearly the same rate. As of this writing, the Evidence for ESSA website (https://www.evidenceforessa.org) only has results listed for 110 programs (many of which do not have an online component) (Evidence for ESSA, 2018). Furthermore, the standards required for inclusion in Evidence for ESSA and the competing What Works Clearinghouse (US Department of Education, 2014) have more to do with the quality of a study's methodologies than its relevance to a specific school (US Department of Education, 2014; Evidence for ESSA, 2018). Evidence for ESSA and the What Works Clearinghouse prefer evidence from randomized controlled trials (studies involving pre-selected, randomly-assigned control and experimental groups), with quasi-experimental studies a distant second, but have no requirements for size, save those imposed by the need to obtain statistical significance, and no requirements for representativeness or breadth in the students studied. Furthermore, most published randomized controlled trials ignore issues of implementation (O'Donnell, 2008), looking at overall effects between conditions, but not diving deeper into issues of dosage (how much did students use the intervention?) (cf. Koedinger, Booth, & Klahr, 2013) or cost (cf. Levin & McEwan, 2000).

As such, the evidence base that exists is high quality, but remains limited in its usefulness. Consider a school district attempting to evaluate several apps for use. Most of the apps and webpages currently being used by schools are not represented in the existing evidence base. Even when there is a published report in Evidence for ESSA or the What Works Clearinghouse, it is relatively unlikely to involve students in settings similar to the school district making a decision today. And finally, some apps may look ineffective in the evidence base, but would have been effective if they had been implemented appropriately (see discussion in Feng et al. 2014).

This report attempts to take a first step towards addressing these limitations and improving the knowledge base available to schools. We take 150 popular apps and webpages targeted at students —more educational programs than are found in the entire Evidence for ESSA website— and collect data on these apps from hundreds of thousands of students nationwide. We analyze which apps are receiving usage in line with license purchases by school districts, and how much usage is occurring for each student and app in our sample. We then correlate this usage to standardized examination growth data to derive preliminary evidence on which apps have a pattern where more usage is associated with more learning. By investigating the amount of usage, the dosage of the intervention, we avoid concluding that a program is ineffective solely because it is scarcely used.

In analyzing these relationships, it is important to note that our data is correlational, observational in nature, rather than as experimental, and as such must be considered preliminary rather than conclusive. In addition, results may vary for specific districts, where due to implementational factors or differences in students, some apps may be more or less effective than the overall national results we report. At the same time, this report's breadth enables us to target apps which appear particularly promising, at a national level. These apps can become the subject of further investigation.

We also explore an analysis of the apps with the highest effectiveness, analyzing the contexts where they are producing positive impact. The ultimate impact of this type of research will not be in identifying the best apps overall, but in allowing individual school districts to look at evidence regarding effectiveness in their own district and in districts similar to them.

DATASET SELECTION

CAPTURED ACROSS:

392,603
STUDENTS

48
DISTRICTS

1.48M
HOURS OF USAGE

SUFFICIENT TEST DATA WITH:

177
APPS TOTAL
2 Apps were eliminated due to technical issues in data

25
EDUCATOR FOCUSED APPS

150
STUDENT FOCUSED APPS
For this report, we analyze only student-focused apps

DATASET & DEFINITIONS

We obtained a sample of data from the **Learning Outcomes** module, on the BrightBytes Clarity platform, from 392,603 users in 48 school districts across the United States. BrightBytes was aware of at least one learning product license in each organization. These school districts were distributed across 26 states, from the Pacific to the Atlantic, from the Canadian border to the Mexican border, and in between. These users used 258 apps. Sufficient test data was available to analyze student improvement for 177 of these apps. We focus only on apps used by students, not on apps used by teachers or school administrators. In this example, 25 apps were focused on teachers or school administrators rather than students, and two apps appeared not to have been logged correctly, leaving a remaining 150 student-focused apps with test data. These apps were used for a total of nearly 1.5 million hours.

USER

A learner who has ever logged into an app

INTENSIVE USER

A learner who used the app at least 10 hours between assessments

A user of an app is defined as a learner who ever logged into an app. Generally, for apps where licenses are purchased there are more licenses purchased than users. However, in some cases there are single licenses purchased for an entire class or school. An intensive user of an app is defined as a learner who used the app at least ten hours in between two assessments, which is, in most cases still under an hour a week.

TIME SPENT

The total number of minutes the student used the system *(Deleting long pauses in which software was clearly not in use)*

DAYS USED

The total number of distinct days when the student used the software *(Ignoring how long the student used the software on each of these days)*

We analyze a student's amount of usage in two ways: the total number of minutes the student used the system (deleting extremely long pauses where the student was clearly not using the software), and the total number of distinct days when the student used the software (ignoring how long the student used the software on each of these days).

STUDENT IMPROVEMENT

Standardized test scores improving between one test (fall/winter) to another test (winter/spring)
(if a student completes three tests, we only consider the first two)

STANDARDIZED TESTS

Scores were collected from standardized tests in math, science, and ELA
(Tests normalized by district since different tests are on different scales)

We measure student improvement in terms of whether a student's scores improved from one standardized test (fall/winter) to a later standardized test (winter/spring). We look at standardized tests in math, science, and English language arts (ELA). Since different standardized tests have different scales, we turn each test's scores into "Z scores", a standard statistical technique (Mosteller & Bush, 1954). In Z scores, 0 is the average score, -1 is one standard deviation below average, +1 is one standard deviation above average, +2.3 is 2.3 standard deviations above average, and so on. We subtract the later test Z score from the earlier test Z score. For example, a student who was exactly 0.1 standard deviations better than average on their first test and was then exactly 0.6 standard deviations better than average on their second test has an improvement of 0.5. If a student completes three tests, we only consider the first two.

COST PER LICENSE

The reported cost of a license from a district

COST PER USER

The reported total cost of all liccenses divided by the number of users of the app

COST PER INTENSIVE USER

The reported total cost of all liccenses divided by the number of intensive users on an app

We obtained data on app costs and license purchases from 41 school districts. These districts provided data on the cost of 177 apps; a total of 393 reports were received on the cost of an app. However, 3.3% of these reports were default values of $1.00. We removed apps from consideration of analysis of costs (seven apps) if 50% or more of cost reports were default values. For the small number of cases (six apps) where default values were given but less than 50%, we calculated cost based on the districts that reported values. A total of 1.8 million licenses were purchased across students, representing an average of 5.02 licenses purchased per student.

APP EFFECTIVENESS

MOST EFFECTIVE APPS: MATH TESTS

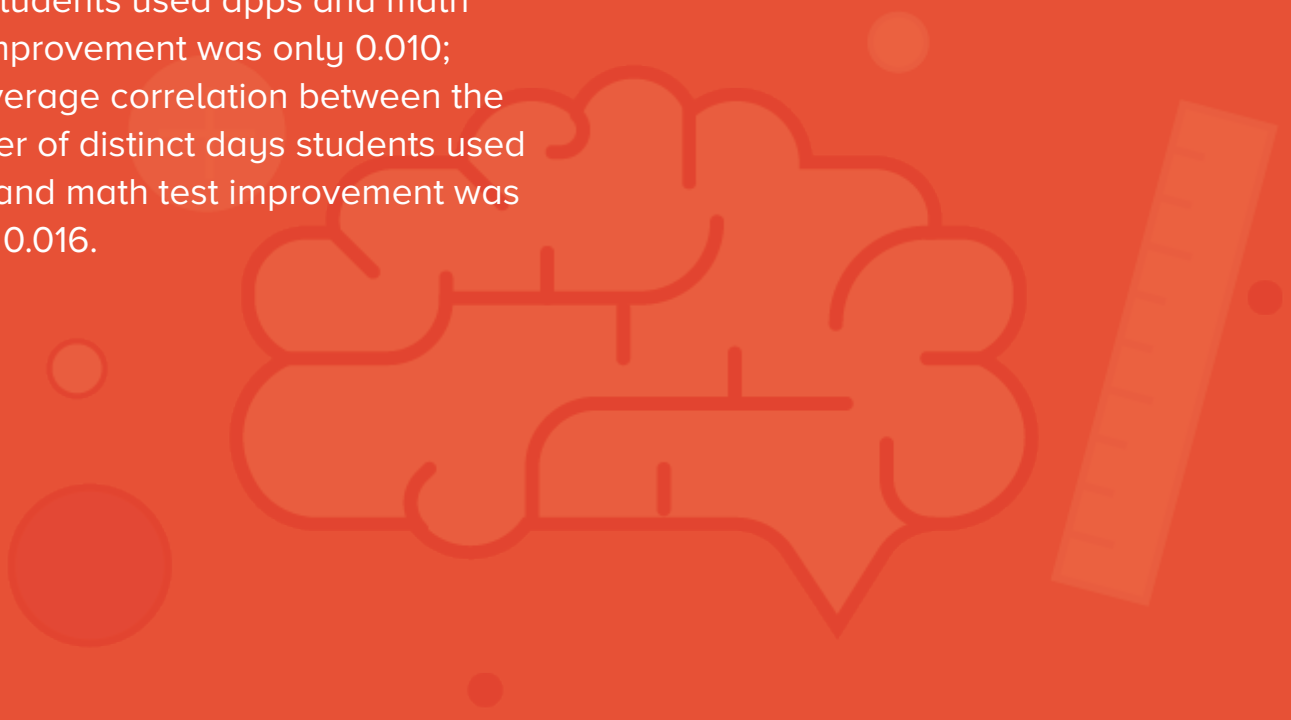
We can look at the degree of effectiveness of each app in terms of the question:

If a student uses the app more, does their performance improve?

A total of 149 apps had both usage data and two math test scores for at least 30 students. Across all of the apps, there was not a general trend towards use of the apps being effective for learning. The average correlation between the amount of time students used apps and math test improvement was only 0.010; the average correlation between the number of distinct days students used apps and math test improvement was a tiny 0.016.

1.8M
Licenses Purchased Across All Students

5
Average Number of Licenses per Student



We analyzed whether a student's use of an app was associated with significantly higher performance, using hierarchical linear models (HLM) (Raudenbush & Bryk, 2002) to control for variance at the school level. Post-hoc tests, though typically desirable for correlation mining analysis such as this (e.g. Baker, 2018), were not used here, since the goal of this analysis is to identify cases for further investigation rather than drawing causal or definitive conclusions. For 21 apps, the more time the student used the app, the more their math test performance improved between assessments (statistically significant or marginally significant effect, $p < 0.1$, with no reversal of direction in HLM due to collinearity). For 24 apps, the more days the students used the app, the more their math test performance improved between assessments (statistically significant or marginally significant effect, $p < 0.1$, with no reversal of direction in HLM due to collinearity). Again, the reader should note that these findings are correlational rather than causal; we cannot conclusively infer that a correlation between more usage of an app and better outcomes indicates that more usage causes the better outcomes. For instance, teachers who use a specific app more may also engage in other positive pedagogical practices to a greater degree.

Figure THETA shows some of the most effective apps in terms of both days used and time spent. The reader will note that most of the top ten apps for impact per days used were also in the top ten for impact per time spent. One key exception was Canvas, which was #9 for days used, but only #33 for time spent, suggesting that consistency of use of Canvas was more important than how long it was used. This is plausible, especially when many teachers use Canvas to post assignments or resources rather than for learning interaction within Canvas. Another interesting exception was Kids Discover Online; students who spent a lot of time using this system did very well, but students who used the system on many days did very poorly. This pattern of results may suggest that Kids Discover Online is best used in an intensive fashion for specific content.

FIGURE THETA. *The most effective apps for math learning, in terms of impact per number of days used and time taken. Statistically significant relationships are shown in boldface. Marginally significant relationships are shown in italics.*

App	Correlation (days used)	Rank (days used)	Correlation (time spent)	Rank (time spent)
ALEKS	**0.896**	**1**	0.246	6
Wikipedia	**0.330**	**2**	0.167	9
LearnZillion	**0.305**	**3**	**0.257**	**5**
DreamBox	**0.278**	**4**	**0.264**	**3**
Seesaw	**0.262**	**5**	**0.378**	**2**
Starfall	**0.255**	**6**	**0.259**	**4**
Mission US	**0.244**	**7**	*0.159*	*13*
Science Companion Prime	0.219	8	0.195	7
Canvas	**0.215**	**9**	**0.097**	**33**
Culture Grams	*0.204*	*10*	0.140	19
Kids Discover Online	-0.174	144	**0.378**	**1**

APP EFFECTIVENESS

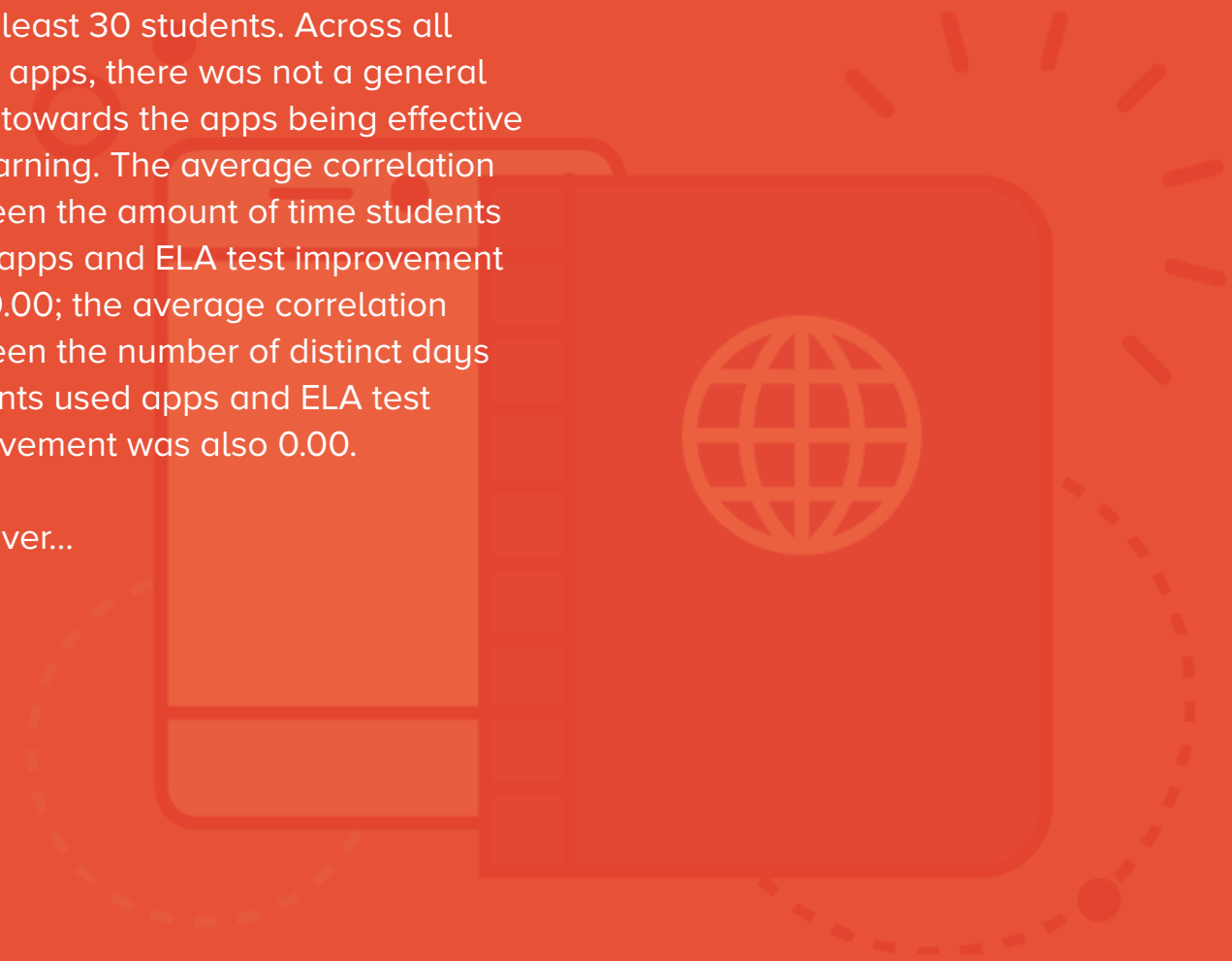
MOST EFFECTIVE APPS: ELA TESTS

We can look at the degree of effectiveness of each app in terms of the question:

If a student uses the app more, does their performance improve?

There were 138 apps that had both usage data and two ELA test scores for at least 30 students. Across all of the apps, there was not a general trend towards the apps being effective for learning. The average correlation between the amount of time students used apps and ELA test improvement was 0.00; the average correlation between the number of distinct days students used apps and ELA test improvement was also 0.00.

However...



Some individual apps appeared to be effective. For 10 apps, the more time the student used the app, the more their ELA test performance improved between assessments (statistically significant or marginally significant effect, $p < 0.1$, with no reversal of direction in HLM due to collinearity). For 11 apps, the more days the students used the app, the more their ELA test performance improved between assessments (statistically significant or marginally significant effect, $p < 0.1$, with no reversal of direction in HLM due to collinearity). Even so, the correlations of apps to ELA tests were generally much lower than the correlations of apps to ELA tests, and, due to small samples, many of the apps with relatively higher correlation were not statistically significant. One relatively clear bright spot was LearnZillion, which had good correlations to ELA test improvement for both the number of days used and the total time spent. Brainingcamp was also correlated for both measures. More days of usage were associated with better outcomes for Varsity Tutors and Wikipedia. More time spent was associated with better outcomes for TED-Ed. Figure IOTA shows some of the apps where usage had the highest correlation to learning outcomes in terms of days used and time spent.

FIGURE IOTA
THE MOST EFFECTIVE APPS FOR ELA LEARNING, IN TERMS OF IMPACT PER NUMBER OF DAYS USED AND TIME TAKEN.
Only apps with significant or marginally significant correlations above 0.1 for at least one measure are shown. Statistically significant relationships are shown in boldface. Marginally significant relationships are shown in italics.

App	Correlation (days used)	Rank (days used)	Correlation (time spent)	Rank (time spent)
Varsity Tutors	**0.289**	**1**	0.124	14
LearnZillion	**0.250**	**2**	**0.255**	**2**
Wikipedia	**0.227**	**4**	0.068	26
Brainingcamp	**0.186**	**8**	*0.162*	*8*
Google Classroom	**0.111**	**13**	**0.055**	**31**
TED-Ed	0.044	40	**0.153**	**3**

APP EFFECTIVENESS

MOST EFFECTIVE APPS: SCIENCE TESTS

We can look at the degree of effectiveness of each app in terms of the question:

If a student uses the app more, does their performance improve?

Unfortunately, very few students in our sample took multiple science tests. Only eight apps had both usage data and two science test scores for at least 30 students. Among these eight apps, only two had statistically significant or marginally significant correlations between the number of days the app was used and science test improvement, and only three had statistically significant or marginally significant correlations between the amount of time the app was used and science test improvement.

Somewhat surprisingly, the top apps did not seem particularly strongly connected to science learning. The app most associated with science improvement was Desmos, a graphing calculator app. The app second most associated with science improvement was Typing.com, an app to learn how to type. Among all the statistically significant or marginally significant apps, only the app third whose time spent was most associated with science improvement, CREATOMbuilder, was obviously connected to science content.

FIGURE KAPPA
THE MOST EFFECTIVE APPS FOR SCIENCE LEARNING, IN TERMS OF IMPACT PER NUMBER OF DAYS USED AND TIME TAKEN.
Only apps with significant or marginally significant correlations above 0.1 for at least one measure are shown. Statistically significant relationships are shown in boldface. Marginally significant relationships are shown in italics.

App	Correlation (days used)	Rank (days used)	Correlation (time spent)	Rank (time spent)
Desmos	**0.311**	**1**	**0.347**	**1**
Typing.com	**0.306**	**2**	0.212	2
CREATOMbuilder	0.123	6	*0.173*	*3*

DIFFERENTIAL EFFECTIVENESS

Although these apps were the most effective overall, they had some variation in where they were most effective. Among the math apps, for example, ALEKS performed particularly well in one medium-sized city, but relatively less well in smaller communities. DreamBox had one particularly successful implementation in the suburbs of a different medium-sized city, but relatively less well in a small city fairly nearby, with in-between performance in other communities. Wikipedia was associated with relatively consistent outcomes in different areas.

Among the ELA apps, LearnZillion performed particularly well in one small town, but relatively poorer in suburban areas and small cities. By contrast, Varsity Tutors performed best in suburban areas (including one of the same suburban areas that LearnZillion performed poorly in) but more poorly in rural areas. Unlike with math, Wikipedia was associated with differential results in different communities, achieving excellent results in one small town but poor results in another small town, and intermediate results in suburban areas, rural areas, and a small city.

CONCLUSION

CONCLUSION

The number of educational apps available to schools has increased rapidly in the last several years, but the evidence base available to schools to decide which apps are effective has not increased to anywhere near the same degree. In this report, we discuss evidence from the BrightBytes Clarity platform on which apps are associated with better student outcomes. We find evidence that in over thirty apps, more usage is associated with higher gains on standardized examinations for math, ELA, or science. Our evidence also suggests that some of the effective apps are more effective for some patterns of usage than others (does it matter how many days an app is used, or how many total minutes?) and that even highly effective apps are associated with different outcomes in different school districts. This result points to the potential value of giving schools access to data on effectiveness in addition to the more widely-used paradigm of relying on data from experimental studies conducted in other contexts.

One limitation that is important to notice is the risk of selection bias (Rothstein, 2009) in our findings. Since randomization was not used, we do not know if the relationships between usage and learning outcomes that we see are due to app usage causing the learning outcomes, or due to some third factor causing both app usage and learning outcomes. For example, it is possible that in some cases teachers who are pedagogically more skilled choose to use apps more, or that teachers of gifted students use different apps than teachers of struggling students. As such, any correlational findings of the type seen here should be examined more closely, and followed up with more conclusive research if possible.

Keeping that in mind, while correlation is not causation, causal results may not hold for new contexts (Sullivan, 2011). An ideal decision-making approach will leverage both types of data. The type of correlational data used here is more preliminary but more scalable.

One of the other primary findings of this report is that usage of apps is generally lower than might be expected. Most apps are used only for limited amounts of time, and most apps purchased by districts go unused. This has an impact on efficacy —an app cannot be effective if it is not used. It also has implications regarding current procurement practices. While some of the data we are seeing may simply reflect site licenses never intended for use across all students, it may also suggest that districts should review their app purchases and attempt to bring purchases more in line with use.

One way to integrate the type of evidence collected here with traditional RCT evidence is as follows: First, a district may want to use existing published evidence from other districts (randomized controlled trials and quasi-experiments) when deciding which apps to adopt. Once adoption has occurred, the district can review their data in a platform such as the BrightBytes Clarity Platform to see if the app is being used enough. If there are challenges limiting usage, the district can work with schools and teachers to address these challenges. In parallel, the district can investigate if the app seems to be associated with better outcomes in their schools. Correlational evidence suggesting efficacy can be followed up with a more formal randomized controlled trial or quasi-experiment, which can be added to the nation's public knowledge base.

By using district financial resources more efficiently, monitoring usage to increase implementation fidelity, and choosing apps that work better for their individual districts, school districts may be able to produce better results for their students with limited additional effort.

REFERENCES

REFERENCES

Baker, R.S. (2018) Big Data and Education. 4th Edition. Philadelphia, PA: University of Pennsylvania.

Evidence for ESSA (2018). Evidence for ESSA. Retrieved 10/14/2018 from https://www.evidenceforessa.org. Baltimore, MD: Johns Hopkins University.

Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R. (2014). Implementation of an intelligent tutoring system for online homework support in an efficacy trial. In *International Conference on Intelligent Tutoring Systems (pp. 561-566). Springer, Cham.*

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, 342(6161), 935-937.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: technology bringing learning science to the classroom. In K. Sawyer (Ed.), T*he Cambridge handbook of the learning sciences* (pp. 61-78). New York: Cambridge University Press.

Levin, H. M., & McEwan, P. J. (2000). *Cost-effectiveness analysis: Methods and applications* (Vol. 4). Sage.

Mosterller, F.M., & Bush, R.R. (1954) Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of Social Psychology: Vol. 1: Theory and Method*. Cambridge, MA: Addison-Wesley.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, 78(1), 33-84.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: *Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education finance and policy*, 4(4), 537-571.

Sullivan, G. M. (2011). Getting off the "gold standard": randomized controlled trials and education research. *Journal of Graduate Medical Education*, 3(3), 285-289.

US Department of Education (2014). What Works Clearinghouse: Procedures and standards handbook (Version 3.0). *Washington, DC: US Department of Education.*