Empirical Education
EMPOWERING EDUCATORS THROUGH EVIDENCE AND INSIGHT

For Comparison Group Studies on EdTech Products, What Counts as Being Treated?[1]

Valeriy Lazarev
Denis Newman
Empirical Education Inc.

Malvika Bhagwat
Newsela

Introduction

Edtech products are becoming increasingly prevalent in our schools as the cost of devices drops and the availability of digital connectivity increases. With the use of edtech growing in U.S. schools and with the need for evidence of impact to support school purchasing decisions, efficient methods for conducting rigorous studies are necessary.

There are over 4,000 edtech products for K-12 schools. While often attractive to teachers and engaging for students, few products have credible evidence that they make a difference for the student outcomes that matter to schools. The slow pace of conventional academic and government-funded research can't keep up with the pace of new product releases and their continuous updates. This calls for a much more efficient research process. We will also need a funding mechanism that is sufficient to pay for thousands of studies and not burdened by the slow pace of federal grants and contracts.

This paper uses studies we conducted on edtech products in US schools as case studies illustrating how data generated by the use and operation of the products, as collected in cloud-based servers, can be used in efficiently conducting comparison studies that meet the evidence standards built into the Every Student Succeeds Act (ESSA). The paper builds on the approach we are taking with *Evidence as a Service*™, which is aimed at the edtech providers looking for evidence of efficacy for marketing, and areas of strength and weakness to drive continuous improvement. More information on the service can be found here. The initial work on this service was supported under a contract with Reach Capital. The case studies reported here feature Reach portfolio companies and illustrate the considerations built into the initial prototypes.

[1] Paper presented at the annual meeting of the American Educational Research Association in New York April 13, 2018. This version incorporates analyses presented at the Bay Area Learning Analytics Network February 24, 2018.

Opportunities and Challenges in Studying Edtech

This paper explores opportunities as well as challenges in the use of cloud-based product usage data to identify the treatment group for matched comparison or quasi-experimental (QE) impact studies.

The nature of any given product's wide-ranging intensity of implementation can create challenges in identifying units that use the product adequately, such that it could reasonably be expected to make a difference in students outcomes and, more importantly, be recognized by decision-makers in other schools as having been treated.

We define edtech as software products that are internet-based such that usage metrics are collected routinely at the level of individual students or teachers. These data are needed for internal functioning, e.g., tracking what tasks a particular student has completed, and for reporting to teachers, e.g., which students were having difficulty with particular topics. With appropriate permissions, de-identification, and security, these data can be used by researchers to identify which schools, classes, or students were the users, the extent of their use, and even the quality or patterns of their implementation. These data can be linked at the student level with district administrative data, or at the school level with publicly available state data. When the product's usage is linked to characteristics of the schools, teachers, and students, it provides a very efficient means for identifying a treatment group for comparison group studies..

Distribution and Dispersion of Edtech

We have found that the implementation and use of edtech products are widely dispersed. The business practices that result in the dispersion are important here. Edtech can be widely distributed through the internet, often with a free version that can just be signed up for online by teachers or students. Payments are required for the full functionality or site-license features when purchased for a whole school or district. Given this mode of distribution, there can be a very wide range of implementation from a teacher or student who tries it out in a school where the product is adopted as a required part of the curriculum.

The implementation of edtech products tends to vary greatly by design, because they allow for substantial freedom to choose the intensity of engagement with the product. The result is a wide variability in the level of product usage across students, teachers, and schools. Edtech companies generally collect a wealth of usage data by recording user behavior within the system and/or by using web analytics tools.

Our analyses of datasets obtained from a number of edtech companies have shown that usage metrics are over-dispersed. Usage can vary by orders of magnitude even within the top quintile or quartile of most active schools, whereas the average per-student usage in the bottom half of the distribution is trivial and unlikely to make a measurable difference.

Identifying the Treatment Condition

While the amount of usage can be widely dispersed, it is documented in great detail and collected centrally via the servers that run the software, handle logins, and interact with students and teachers. Usage does not consist of a single measure, but rather multiple metrics per user. Products may generate dozens of metrics per user. The company may select metrics associated with what they consider to be critical aspects of the product implementation, such as number of tutoring sessions, or they may be more routine, such as average length of session. These may be reported individually, or aggregated at the grade- or school-levels. The two examples we discuss in this paper involved aggregated data.

Teachers are on the front line in the implementation of the program where finite resources of time, coaching, bandwidth, devices, and other factors may affect engagement with the program. Projects such as ISTE's Edtech Advisor are documenting these resources. At the same time, edtech service companies have data that largely defines the infrastructure on which the implementation is built. We show in two case studies how the usage data collected from teachers and students provided metrics for the implementation, which served as the basis for identifying the quasi-experimental treatment group.

Case Study 1: Newsela

We draw on our experience conducting a large scale statewide comparison group study in California schools of a reading product from Newsela (Empirical Education, 2017). This paper takes the perspective of a researcher following the requirements of ESSA to generate what the legislation would define as Level 2 evidence. We studied the impact, over one school year, of the program implemented in hundreds of schools throughout California in grades four through eight.

Establishing the Footprint and Patterns of Usage

The first step in our *Evidence as a Service* process was to analyze the total national footprint of the product. Usage data from all customer schools were matched to our database of school characteristics, as explained in more detail below. The resulting footprint analysis served to inform internal marketing efforts and to provide demographic distributions of interest to impact investors. For the purposes of the subsequent impact study, the analysis gave us a broad picture of the patterns of usage. It also provided us with information on the distribution of schools that were considered "active users", which allowed us to identify states where an impact study would be feasible—those that had a sufficient number of active schools as well as enough similar schools to constitute a valid comparison group. For this initial study, we decided to conduct a school-level analysis using publicly-available state data on school outcomes in California. (Outcomes in California and many other states are reported at the grade-level within each school.) This process is described in detail in the following sections.

Identifying the Most Active Schools

Usage metrics were provided by Newsela for all users for whom school information was available. We created an aggregate usage score to identify the "Most Active Schools." This process involved the following steps:

- Usage metrics were aggregated at the school level
- Each usage metric was transformed into a log scale and adjusted to be proportional to the size of each school (i.e. the number of students)
- Usage metrics were combined statistically using principal component analysis. Since more than 90% of the total variance was accounted for by one component, this provided a general usage score for all user schools
- The top 20% of schools in terms of this combined usage metric were designated as the "Most Active Schools"

Dividing the schools into quintiles of overall usage resulted in a distribution shown in Figure 1, which displays the distribution of each metric across the quintiles.

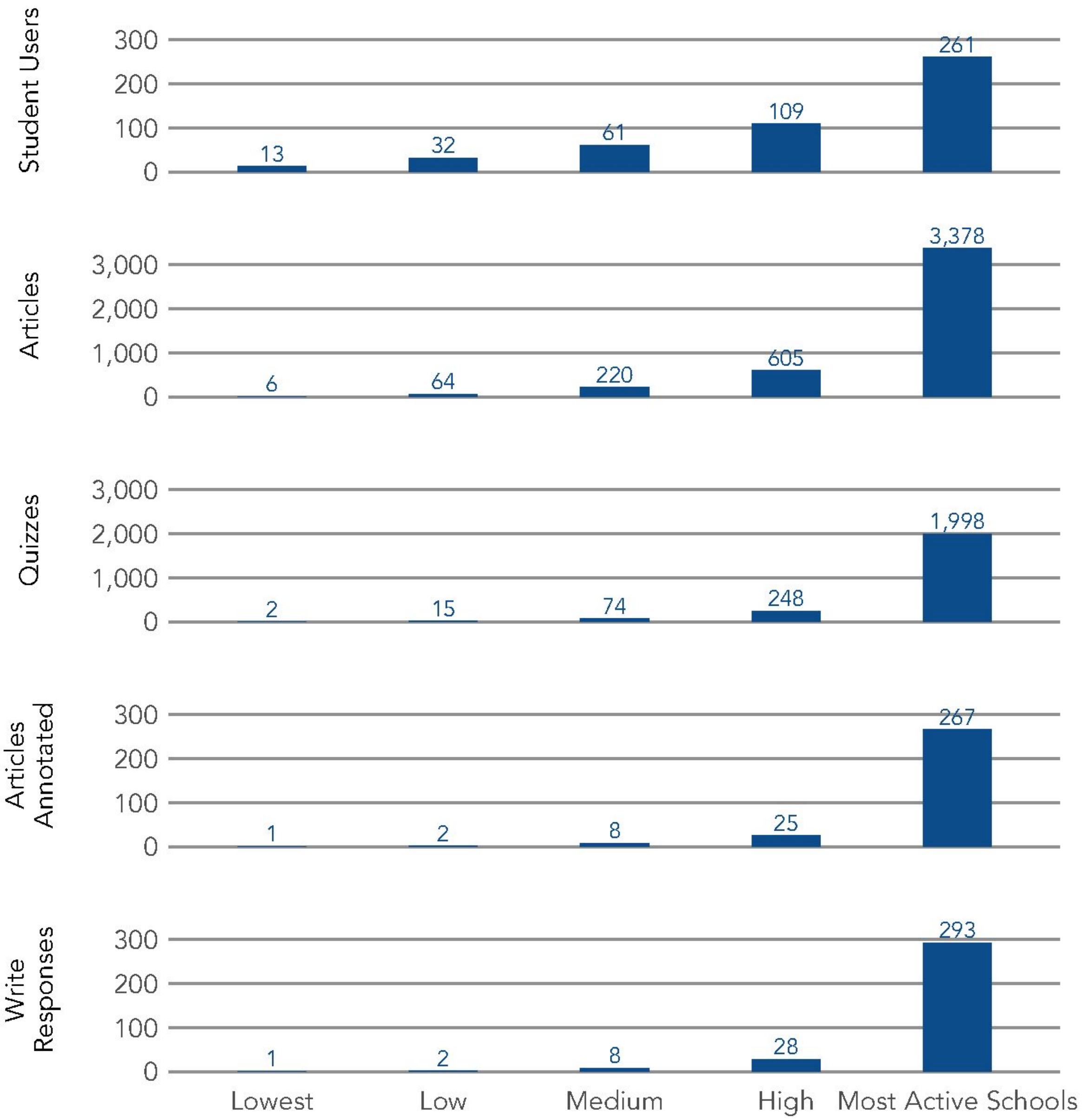

FIGURE 1. USAGE METRICS BY QUINTILE

Note. Values reported represent the mean of all schools within the corresponding quintile of usage.

As the figure illustrates, the vast majority of usage was in the top quintile. In our experience, this pattern is common in edtech usage data.

Footprint as Indicator of Feasibility

Identifying the most active users allowed us to move to the next step of identifying states in which there were sufficient numbers of such schools to conduct a study. Figure 2 shows all districts with the Most Active Schools. The size of each bubble reflects the number of Most Active Schools in that district.

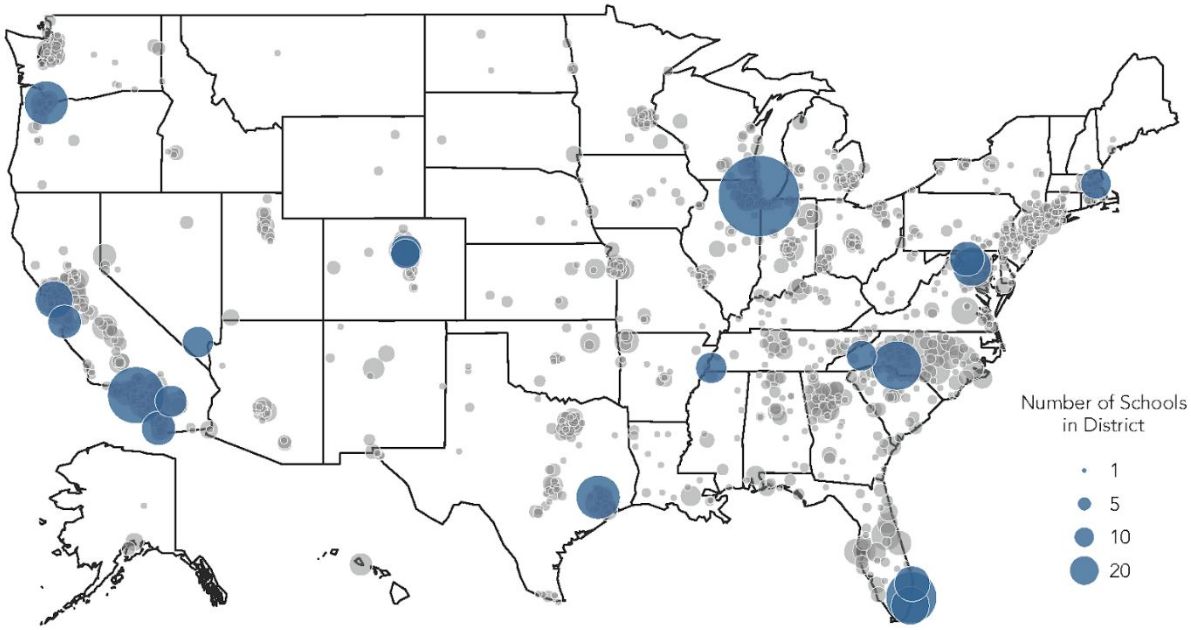

FIGURE 2. MAP OF MOST ACTIVE SCHOOLS

This illustration represents public school districts located in the continental US. Districts with fewer than 5,000 students are not included. The top 20 districts by number of Most Active Schools are indicated in blue. The size of each bubble reflects the number of active schools in that district.

California was selected as a good candidate, as the data showed more than one thousand active customer schools with relevant grades, a large number of non-customer schools with similar demographics, and a state test that addressed the reading outcome.

Conducting the Comparison Study

During the 2015-2016 school year, Newsela was implemented in more than 8,000 schools in California. For the purposes of this study, the pool of Newsela schools was limited to those where some students took at least one quiz during the year.

In the full sample, there were 2,525 such schools, and Newsela users comprised 35% of the student population of these schools. Due to the availability of achievement data, only grades four through eight were included.

The school customers in California displayed the same level of distribution as we found in the national population, so we began by identifying the treatment classes (grades within schools with sufficient usage).

Identifying the Treatment Cases

Newsela had a theory of action that pointed to a measure of "sufficient" use. Newsela recommended that students should take at least 24 quizzes in a year. The number of completed quizzes was a meaningful metric because quizzes were the culmination of a student being assigned a reading, reading the text, and taking a quiz. (Quizzes are shown in Figure 1 among the seven metrics provided.)

Developers often have some idea of the effective level of individual usage, but it can be problematic to translate these estimates into metrics that can be measured in large-scale implementations and that use school-wide or classroom data. Moreover, the wide distribution of usage levels may result in a treatment group that is too small for the purposes of statistical analysis.

An effective level of usage can be derived from the data. The analysis at this point was intended to understand the relationship between our simple "top quintile" usage, and the more meaningful criterion based on quizzes taken. We ultimately chose our top quintile criterion after determining the equivalence of the two through a sensitivity analysis.

This analysis is performed using linear regression of the outcome on the usage indicator, student demographics, school characteristics, and pretest scores. Appropriateness of the assumption about the linear functional relationship between the usage and the outcome was tested using a nonparametric (generalized additive) model.

Our correlational analysis yielded a statistically significant ($p < .05$) and positive estimate of association between the usage metric (quizzes per student) and the outcome metrics. A test for the shape of the relationship showed that the association was linear for almost all values of the usage metric. The relationship flattened out with very intensive usage (more than 70 quizzes per student) but the number of observations in this range was too small for reliable inference.

Matched Comparison Analysis

The second step involved QE evaluation of the product impact.

Method

This evaluation examined the effect of Newsela usage on reading performance in California schools during the 2015/16 school year, as measured by the Reading area of the Smarter Balanced ELA assessment. Reading area test results for each student are reported as one of three achievement levels: Below Standard, Near Standard, or Above Standard. California high-stakes test results are published at the class level, i.e. percentages of students in each of the three achievement levels among all students at a given grade level in a given school. For the purposes of evaluation, Newsela usage was aggregated at the class level in participating schools to match the available achievement data.

The analytical sample was established as follows:

- Student-level Newsela usage records were aggregated at the class level, and classes with at least one quiz taken were selected
- Schools with most active use were selected
- Newsela school data were linked to NCES demographic data
- Results from the 2016 Smarter Balanced test were collected for all schools in California and linked to Newsela school data.

The impact of Newsela was assessed using a matched comparison group design whereby each Newsela class was matched to up to four classes of the same grade level in non-user schools, but with similar (per nearest neighbor criterion) demographic characteristics and prior year test performance. A correlational analysis was performed using data from all Newsela classes with any greater-than-zero number of quizzes taken.

Two alternative criteria were used to select most active classes:

1. 24 or more quizzes per tested student

2. Top quintile of classes by the number of quizzes taken (more than 10 quizzes per student).

The first criterion was suggested by Newsela's own formative research, but resulted in too small a sample (144 Newsela classes). Use of the top quintile was suggested by prior *Evidence as a Service* studies and resulted in a larger sample size (394 Newsela classes). The appropriateness of the top-quintile threshold was tested by sequentially removing less active users by percentiles of the usage metrics and re-estimating the results on the smaller sample.

The impact analysis produced the same effect size estimates of 0.08 using either criterion (24 quizzes per student and top quintile of usage distribution). However, the first approach resulted in a smaller sample (including only 144 Newsela classes) and a low confidence level (p level .13), while the second approach resulted in a larger sample size (394 Newsela classes) and a much higher confidence level (p level .01). In addition, the first approach could not discern differential treatment effect by grade level, whereas the second identified a substantially higher impact in grade 7 (effect size of .16 with the p level of .06).

Sensitivity Analysis

Our test of the sensitivity of the estimates to the choice of usage threshold confirmed that only the estimates that included the top quintile of most active users (anywhere from 17 top percentiles to 21 top percentiles) resulted in statistically significant effect estimates. Higher usage thresholds resulted in treatment groups that were too small and provided insufficient statistical power. Lower thresholds resulted in larger treatment groups that were diluted by observations below the effective level (see Figure 3).

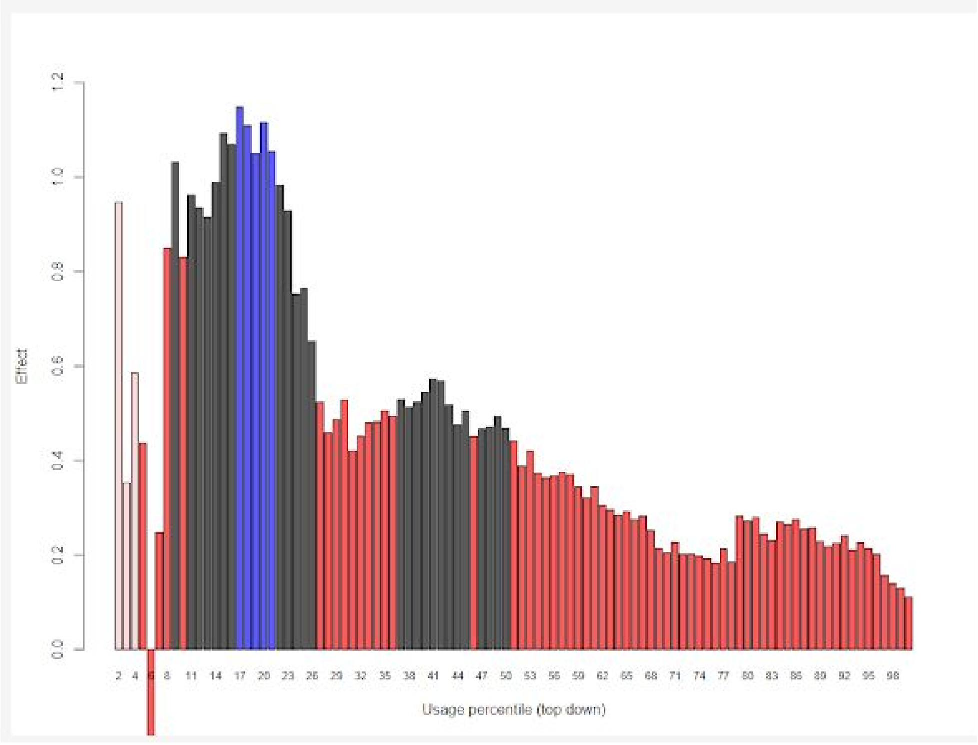

FIGURE 3. QE IMPACT ESTIMATES UNDER DIFFERENT USAGE THRESHOLDS

Note. Horizontal axis shows the threshold percentiles of usage included in the sample. Vertical axis: corresponding effect estimate (on the outcome metric scale). Bar color corresponds to the confidence (p level) of estimate: red - greater than .2, black – between .05 and .2, blue – less than .05.

Results

Active use of Newsela was shown to result in improved student outcomes on the reading portion of the California state test, as compared to a matched sample of non-users. The effect of Newsela on schools was found to be, on average across grades 4-8, a 1.34 percentage point increase in the proportion of students performing "near standard" or "above standard" in the Reading area of the California Smarter Balanced ELA Assessment. We have high confidence of this result (*p* < .05). Figure 4 illustrates the overall result.

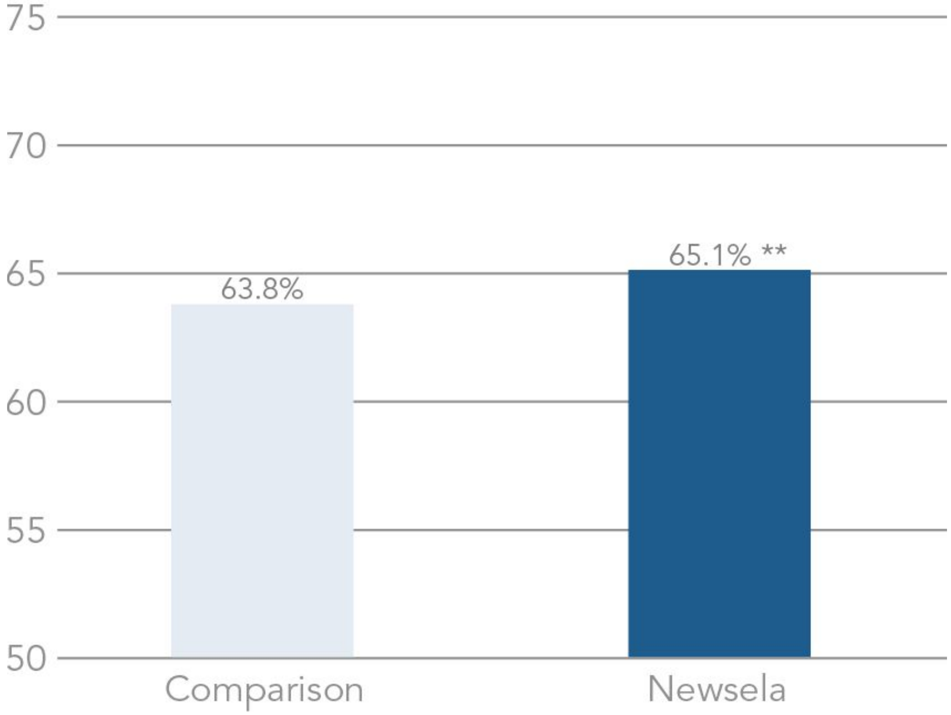

FIGURE 4. EFFECT OF NEWSELA. COMPARISON OF COMPARISON AND NEWSELA SCHOOLS IN PROPORTION OF STUDENTS PERFORMING NEAR OR ABOVE STANDARD, PERCENTAGE POINTS.

Note. Asterisks indicate significant differences between comparison and product. (p < .2, ** p < .05). Reported percentages are adjusted for group differences at baseline.*

We also tested the impact of Newsela on schools that differed in the percentage of the following subgroups of students: economically disadvantaged, English learners, major ethnic groups (Black, White, and Hispanic). We did not find a difference in the impact of Newsela on schools regardless of the percentage of students with these characteristics.

Case Study 2: Multiple Usage Patterns

We saw in the case of Newsela that the principal component analysis resulted in a single dominant component, which could be used as a general usage metric. We conducted the footprint analysis of another edtech company, Kaymbu, whose product is sold to pre-K schools and centers. Kaymbu also supplied usage data for all US schools that used the tool between October 2013 and November 2017. The data included four usage metrics shown in Table 1.

TABLE 1. USAGE METRICS PROVIDED BY KAYMBU

Metric	Description
Daily Notes	Routine notes created for specific students (e.g. related to diapering, naps, etc.)
Outgoing Messages	A message sent to a particular parent or group of parents
Moments	Photos or videos uploaded for a classroom
Storyboards	A collection of moments that are put together for a newsletter or some other type of communication

We followed a similar set of steps with these schools, although in this case we used a database of pre-schools provided by MDR and supplemented with our own analytics. This provided information on each customer including size, demographics, and location of each school.

Usage metrics were aggregated to the school level, then adjusted to account for factors such as the length of exposure to the product, the number of classrooms using the tool, and the number of students at the school. The principal component analysis in this case indicated that there was no single dominant component. The first component explained 63% of total variance and the second component explained 28%.

In this case, a factor analysis showed two kinds of use with characteristics sufficiently distinct such that the Kaymbu developers were able to name them: General Usage and Documentation. From there, each school could be given a score indicating the extent to which it exemplified one or the other type.

With these scores we were able to identify two clusters of schools with high scores on one or the other usage type. A third category of schools did not score sufficiently high on either type to be identified as part of the cluster.

Figure 5 shows how the clusters were defined by our statistical algorithms. Each school is represented by a bubble. The size of each bubble is proportional to the number of students enrolled. Each school has a score on each of the two dimensions of usage (General Use and Documentation). Schools toward the right side of the graph and shaded turquoise were characterized by high levels of general usage (and only a moderate amount of documentation), and are hence labeled “Generalists”. Schools toward the top of the graph and shaded purple were characterized by high levels of documentation (and only moderate amounts of general use), and are labeled “Documenters”.

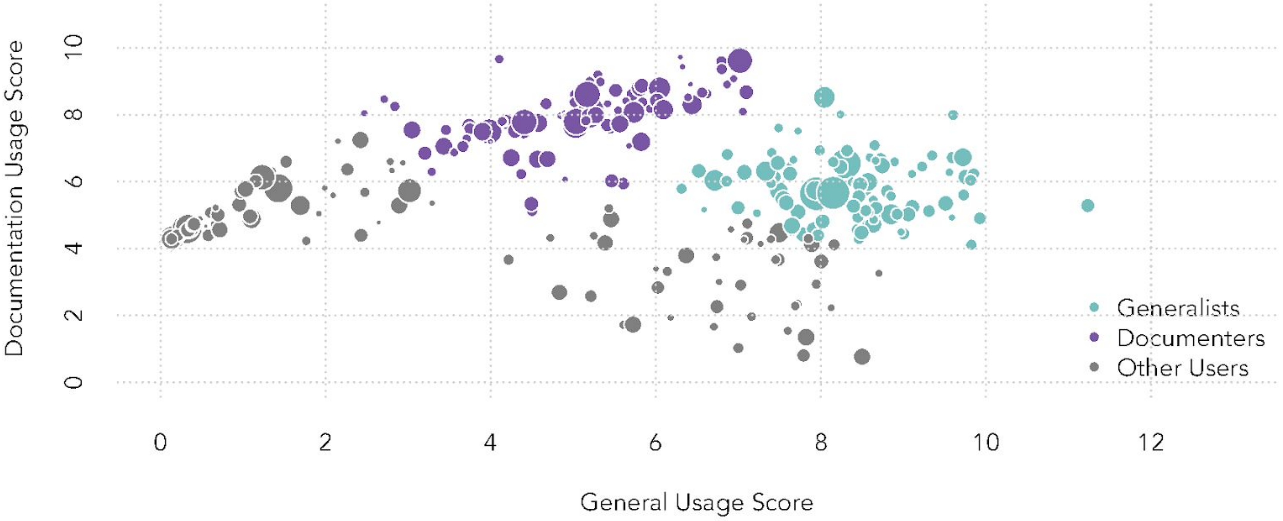

FIGURE 5. SCHOOL CLUSTERS BY USAGE PATTERN

Schools in gray may have had moderate scores as Generalists or Documenters but had low scores on the other dimension, and therefore did not exhibit a level of overall usage the algorithm needed in order to confidently identify with one cluster or the other.

Figure 6 shows differences in the four usage metrics and illustrates the patterns that define the two clusters of Generalists and Documenters. The last set of bars includes all other schools.

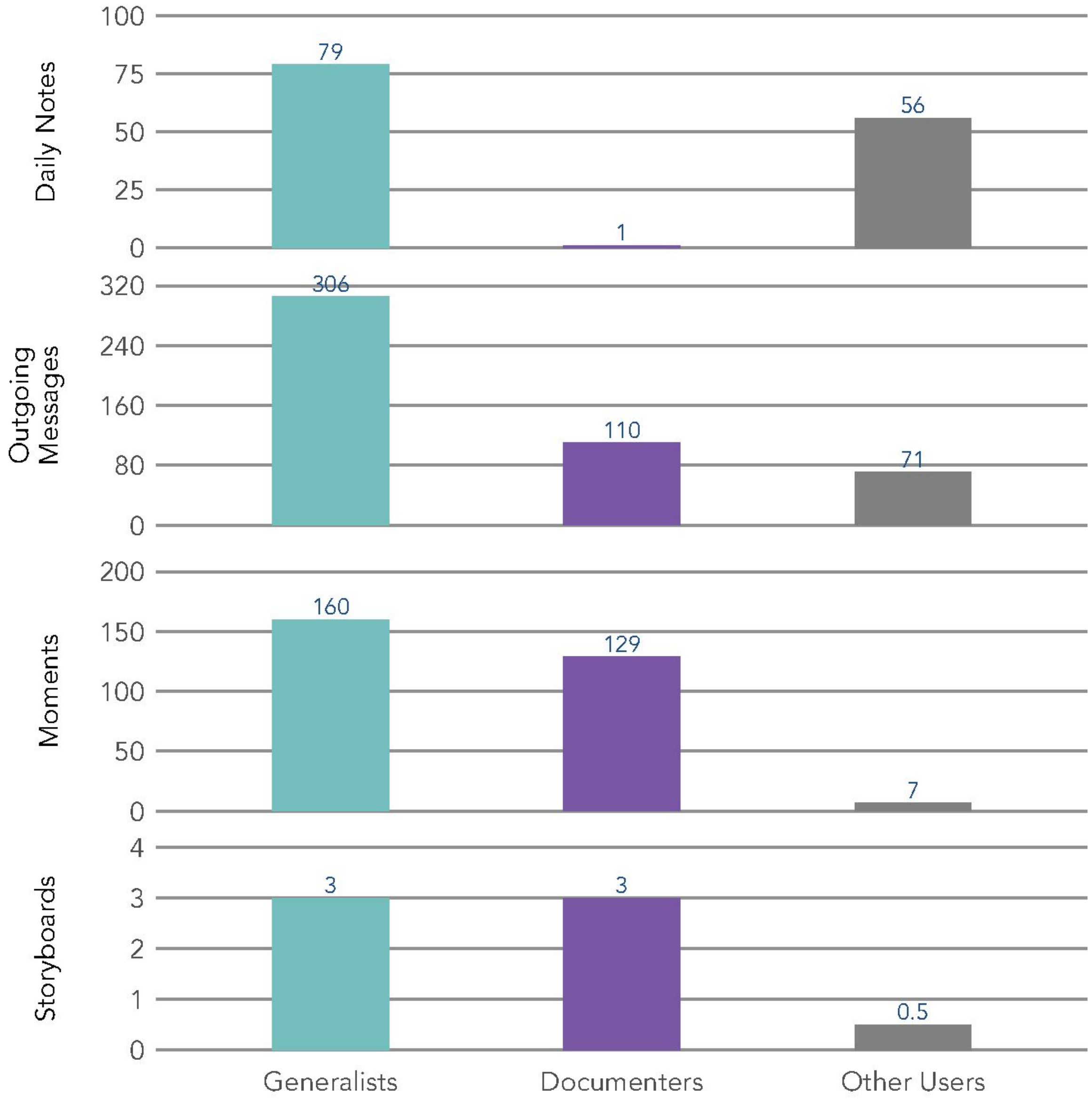

FIGURE 6. USAGE METRICS BY USER TYPE

Note. Values reported represent the mean of all schools within the corresponding usage category.

As Figure 6 shows, "Generalists" used all the modes and overall had a higher level of usage. "Documenters" were much less inclined to utilize daily notes and outgoing messages, and utilized moments and storyboards at a level similar to Generalists. Other users tended to have a lower level of usage across the board.

Discussion

Usage data linked to data on the schools can provide value to the edtech companies both for marketing and sales, as well as for product development and improvement. If studies can be conducted and reported efficiently, they can provide a sufficient return to warrant a modest investment by the provider in data collection and analysis.

Our case studies used alternative methods of identifying quasi-experimental treatment groups in large-scale edtech product adoptions. Edtech products, through their usage metrics, are opening up important avenues for improving quasi-experimental studies by providing systematic approaches for identification of the treatment group. We have also shown that using all customers as the treatment group will include a large number of units that had a trivial level of usage. A line must be drawn. Often the developer is unaware of the relationship between usage and impact when tested on a large scale. We have addressed a new problem for research methods in schools—using edtech usage data as a measure of implementation and a means for identifying the treatment group of a comparison group study.

Objections to Use of Usage Data to Identify Treatment Cases

With edtech products, the usage data allows for precise measures of exposure and whether critical elements of the product were implemented. Providers often specify an amount of exposure or kind of usage that is required to make a difference. And educators often want to know whether the program has an effect when implemented as intended. Researchers can readily use data generated by the product (usage metrics) to identify compliant users or to measure the kind and amount of implementation.

Since researchers generally track product implementation, and statistical methods allow for adjustments for implementation differences, it is possible to estimate the impact on successful implementers, or technically, on a subset of study participants who were compliant with treatment. It is, however, very important that the criteria researchers use in setting a threshold be grounded in a model of how the program works. This will, for example, point to critical components that can be referred to in specifying compliance. Without a clear rationale for the threshold set in advance, the researcher may appear to be "fishing" for the amount of usage that produces an effect.

Some researchers reject comparison studies whereby identification of the treatment group occurs after the product implementation has begun. This is based, in part, on the concern that the subset of users who comply with the suggested amount of usage will get more exposure to the program. More exposure will result in a larger effect. This assumes of course, that the product is effective, otherwise the students and teachers will have been wasting their time and likely to do worse than the comparison group.

There is also the concern that the "compliers" may differ from the non-compliers (and non-users) in some characteristic that isn't measured. Even after controlling for measurable variables like prior achievement, ethnicity, English proficiency, there could be a personal characteristic that results in an otherwise ineffective program becoming effective for them. A similar point is made by non-researchers when they point out that the impact of a product is controlled by the quality and resources available for implementation of the product--a greater amount of usage is thought to be associated with greater gains. This paper takes the position that a product's effectiveness can be strengthened or weakened by many factors. A researcher conducting any matched comparison study can never be certain that there isn't an unmeasured variable that is biasing it--that's why the WWC only accepts QEs "with reservations." But we believe that as long as the QE controls for the major factors that are known to affect outcomes, the ESSA requirement that the researcher "controls for selection bias" is met. With those caveats, we believe that a comparison group study, which identifies users by their compliance to a pre-specified level of usage, is a good design.

Value of Efficacy Research to the Edtech Industry

Studies that look at the measurable variables that modify the effectiveness of a product can not only be useful to schools in answering their question, “is the product likely to work in my school?” but points the developer and product marketer to ways the product can be improved.

Our footprint reports provide a first step in mapping out the location and demographic characteristics of the most active customers, and in some cases can provide quantitative indications of distinct patterns of usage. While the company’s customer relationship team may have an intuitive sense of the modes of usage, the footprint report can map those patterns to demographic characteristics and indicate geographic distributions of these patterns.

The impact studies, tying usage to outcomes, can provide evidence of which usage metrics and patterns are associated with improvements at the school. This provides what ESSA considers “evidence of promise” but helps the company’s trainers support more successful implementation. The quasi-experimental comparison group study results provide moderate evidence and, with connections to demographics, can show which subgroups benefit the most, or—as in one of our case studies—can show that the product is equally effective for all subgroups in the population under study.