Improving teacher's implementation of educational technology at scale by "nudging" teachers with usage information

Ethan Scherer

Dan Thal

Thomas J. Kane

Harvard Graduate School of Education

Douglas O. Staiger

Dartmouth College

April 2019

Abstract

We conduct a field experiment with 373 teachers across two local education agencies, to examine the impact of a low-cost, teacher-focused behavioral intervention on student-level usage of educational software in K-8. Teachers in the treatment group received emails containing information on their students' weekly usage compared to teachers in the same school; the controls received no information. Overall, these "nudges" did not increase average student usage. However, teachers below the school average increased their average student usage by 5 minutes per week (a 12% increase), while teachers above the school average decreased their usage by half as much. Our experiment provides preliminary evidence that targeted data-driven interventions can change classroom practice both initially and throughout the year assuming continual contact.

1. Introduction

Teaching is a complex and dynamic job that forces individuals to juggle multiple tasks and accurately prioritize their actions. With the creation of statewide longitudinal data systems and additional local data collection, one of these important tasks is utilizing student data to inform a teacher's practice (Campbell and Levin, 2009; McNaughton, Lai, and Hsiao, 2012). In 2018, 18 states passed 27 new laws on training, support, and resources for districts to data drives local decision-making (The Data Quality Campaign, 2018) Yet, while systems collect more types of data, such as interim test scores, social-emotional skills, absences, homework completion, educational software data, knowing which data point or points to focus remains difficult. Furthermore, while some districts develop teacher dashboard, often the information needed to assess student progress requires different systems and user credentials increasing the burden on teachers. Given the limited attention teachers can dedicate to any one task, it is a challenge to acquire all of the relevant and important information for practice decisions.

This paper presents an evaluation of a randomized controlled trial of an informational nudge with social comparisons within four local education agencies (LEA). Nudging interventions leverage research and tools identified from psychology and behavioral economics to improve policy outcomes (see Damgaard and Nielsen, 2018 for a recent review). The goal of the intervention was both to provide salient information in an easily digestible format and apply mild pressure to adhere to social norms in order to improve implementation of a common mathematics software, and ultimately, increase test scores. All of these LEAs observe significant variation in student usage classroom-to-classroom. In an effort to improve usage across the LEAs, each site provided teachers with personalized feedback on student usage of the educational software, a social comparison to usage within the same school and grade span, and generic tips on how to

help students in the software. The notifications were sent to the teachers' LEA email account about every month over the course of the 2016-17 school year.

The study finds two clear intervention results. In our first model, we assess the specific signal contained in the email message. When we examine the overall effect of the email, we find no significant difference between the treatment and control groups. However, the email contains a line of text comparing a teacher's use to other teachers using the same software in the same school. A bar chart allows teachers to easily make this comparison themselves. When we examine the subset of teachers in the treatment and control who were below the school average, as noted in the email, we see positive and significant effects initially, that decay over the subsequent weeks before the Thanksgiving holiday break. On average, the teachers below their school average significantly increased their average classroom use by 5 minutes per week and students made 0.2%[1] more progress per week through the curriculum compared to the control group. These gains from the teachers below the school average were partially offset by the teachers above the school average reducing the average minutes of use in their classrooms by 1.5 minutes per week, although this reduction was not significant. These effects are not driven by regression to the mean, because the high- and low-usage teachers in the treatment group are being compared to similar teachers in the control group.

Second, within the subsample of sites that continued to send the email throughout the year, we observe similar impacts over the course of the whole year. In the treatment group,

---

[1] Students tend to be in class 40 weeks during the school year, though this varies by LEA. Students are expected to complete 80-100% of the curriculum by the end of the year. As such, they should be making about 2% progress per week, on average.

teachers' status as being above or below the school average could switch from email to email based upon the classroom usage in the weeks leading up to the email. As such, individual signals of performance were endogenous after the first email. Thus, for the set of sites that sent multiple emails over the course of the school year, we instead identify below- or above-average teachers by their average classroom minutes of usage in the five weeks prior to the first email. Using this designation, we continue to see that, on average, students in below-school-average classrooms increase their software usage by about 5 minutes per week, and make 0.2% more progress per week compared to the control group. Similar to the initial email, these gains from the teachers below the school average were offset by the teachers above the school average reducing the average minutes of the classroom by 3 minutes per week, although this reduction was not statistically different from zero. Thus, consistently messaging teachers using the software less in the beginning of the year shows improvement in the average class usage.

Despite these improvement in the minutes, we fail to find a statistically significant improvement in test scores for the students of the low-usage teachers. However, given the relatively small improvement in the average number of minutes per week and the cross-sectional relationship between minutes and end of year achievement, the final sample was not powered to detect the magnitude expected.

These results contribute to a growing literature on teacher nudge interventions. With the emergence of longitudinal data systems that can integrate with electronic messaging software, these types of interventions have gained momentum among policy-makers due to the potential high benefit-cost ratios as well as early success. For example, Jackson and Makarin (2018) show that providing regular reminders for teachers to engage with online off-the-shelf mathematics lessons showed a positive, but statistically insignificant, effect on student test scores. Our

findings add to their work demonstrating that providing personalized information to teachers on student interim outcomes could de-bias their beliefs about their behavior, resulting in increased effort.

Yet, our work indicates a need for caution when using social comparisons. Prior work by Bergman and Hill (2015) examined the highly publicized effect of publicly releasing teacher value-added measures in Los Angeles. They found, similar to our findings, that teachers below the average tended to improve their performance while there was a negative effect for the above average performers. This mean reversion aligned with adjusting to the social norm (the average)[2]. In our study, the teachers operated in high-stakes accountability systems and thus, as the emails came from the district, performing above the average could have been interpreted as a signal that too much time was spent on the software. These findings underline the importance of understanding the context of the norms prior to initiating the experiment.

The organization of the paper is the following. Section 2 describes the context and the program. Section 3 discusses the data, sample, and empirical strategy. Section 4 reports the empirical results. In concluding, Section 5 summarizes and discusses the main results.

2. Background and Program Description

[2] Note that Pope (2015) did not find evidence of the negative effects on high performers using a different method and sample.

2.1 Background

Four LEAs participated in the experiment. All of them use a common mathematics educational software that has a focus on elementary grades (K-5), but also provides content to middle schools as well. All of the sites had been using the software for several years as of the beginning of the experiment. Generally, the software aids teachers in providing additional content to the students, but how this is delivered (e.g., large school-based labs, mobile laptop carts for classrooms) varied by site and even by school within a site. Particularly in the younger grades, most of the sites had 60-80% of their classrooms using the educational software.

Based upon prior work with the LEAs, all the sites had historically struggled with fidelity of implementation, as measured by student usage, across schools and classrooms. The teachers who used the software heavily in one year tended to do the same in the subsequent years as well. The converse was also true. After we presented evidence on the variation in teacher usage, district leaders decided to experiment with the teacher emails, to draw teachers' attention to how their own usage patterns compared to their peers.

Finally, the comparison with other teachers in the same school stirred significant concern among some teachers in Site A. While only a small number of teacher raised these concerns, this LEA opted to drop the intervention after the first email. Furthermore, in Site D, due to teacher attrition and a small sample to start with, we lack treatment and control teachers in the same school and grade-band. As such, our annual analyses are based upon a subset of the sites.

2.2 Intervention

Teachers were randomized equally into three arms. One arm operated under business as usual and functioned as a control, while the two treatment arms received different versions of the

comparative email. Randomization was conducted within cells of school and grade span (K-1, 2-5, and 6-8). During the course of the intervention we sent five emails, on November 4, December 9, February 3, April 14, and May 19[3]. Emails were sent out after school on Friday, or early during the day on Monday of the following week.

For both arms of the treatment, emails reported the average use (in minutes) of the teacher's students over the last week, and told the teacher how that compared to the average use of all students in the same school and grade span. This was also translated graphically into a bar chart comparing those numbers.

For teachers in the second treatment arm, the email also reported what the software vendor's recommended weekly use was for students in the grade span (60 minutes per week for grades K-1, and 90 minutes for grades 2-8). This recommendation was incorporated into both the comparative text and the graphic. Both messages also contained a table with the names and usage of the three students in the class with the lowest use that week. Example emails for both arms are in section 7.4 of the appendix.

All emails were sent directly to the teachers by LEA personnel, rather than our research team. Some of these personnel were from data and research offices, while some were from education technology offices. We hoped that if the email came from an LEA email account, teachers would be more likely to open them. We also wanted teachers to be able to reply to the email directly if they had any questions.

---

[3] Site B did not send the 4th email due to an IT complication. Sites C and D ended the week following the 5th email, so we do not anticipate a large effect of that message, as software use drops off precipitously during the final week of school.

2.2.1 Mid-year changes to the intervention

During March, between the third and fourth emails, we convened the agencies to review preliminary results from the intervention, and propose changes to streamline the message and hopefully improve its effectiveness. From this convening, the agencies came up with four changes, which took effect starting with the fourth email.

First, we added the line "keep up the good work!" if a teacher's was above that of their peers, in an attempt to mitigate the decline in use among teachers above average. Second, we removed the table listing the three students in a class with the lowest use, as our results showed there was no difference in the effect of the email on these students[4]. Third, we changed to reporting use over the entire period since the last email, rather than just over the prior week, to mitigate concerns about fluky weeks. Lastly, we combined the two treatment arms into one, dropping the vendor comparison from all messages, because we saw no significant difference in impact between the two treatment arms, and the results for the vendor comparison arm were consistently lower[5].

Because our sample was too small re-randomize, we are not able to evaluate whether the revisions improved the efficacy of the messages. An example of the revised emails can be seen in the figures of the appendix.

---

[4] See appendix section 7.2 for details

[5] See appendix section 7.3 for details

3. Data, Sample, and Empirical strategy

3.1 Sample

Four sites initially participated in the experiment, scattered across the US in the southwest, east and west coast. The sample contained both charter management organizations as well as traditional school districts. Table 1 reports the number of students, teachers, and schools in our sample. In total, 1,227 teachers were randomized in 65 schools.

As noted in section 2.1, sites A and D are excluded from the annual analysis due to their decision not to continue or our inability to find treatment and control teachers within the same school and grade-band due to attrition, so the sample for the annual analyses is 373 teachers and 8765 students in 27 schools.

3.2 Data

We collected administrative data from the LEAs containing enrollment, student demographics, formative test scores (i.e., tests given two or three times within the same school year to help diagnose areas of concerns for students before the state test) and links between students and teachers to identify classrooms for each site during the 2016-17 school year. We then merged these data with weekly usage data from the educational software provider. The software data reported the number of minutes each student used the software, as well as the percent of the on-grade-level curriculum students completed. In order to understand if there were changes in quality of usage, we created a proxy defined as percent progress per hour of software use, to understand if students tended to cover more content in less time.

Columns (1) and (2) in table 2 report the baseline mean characteristics of the students in our sample, and their average software outcomes during the five weeks leading up to the first

email. Eighty-two percent of treatment students were Hispanic, 65 percent were designated as having limited English proficiency, and 89 percent received free or reduced price lunch, which often serves as a proxy for low socioeconomic status. To test the differences in these samples, we regressed the student characteristic on a treatment indicator and a set of school-grade span fixed effects, with the standard errors clustered at the school-grade span. This allows us to identify statistical differences between the students within randomization block. We note no significant differences for these characteristics.

We obtained scores on the fall math and English Language Arts (ELA) formative tests from each site, but the test companies varied by site. As such, we standardized each site's scores using national norming studies from their respective test providers. We then create concordance tables using prior year state test scores to put formative test scores across all of the sites onto a common scale. The treatment students perform 1.2 and 1.5 standard deviations below the national norms in ELA and math, respectively. Given that these sites tend to serve high poverty urban students, these means appear reasonable. Compared to the control group, students in the treatment classrooms do not differ significantly on baseline test scores either.

3.3 Empirical Strategy

3.3.1 Response to first email.

Our first model evaluates the impact of the first email we sent. We are interested in three measures of student software activity, as described in section 3.2. These outcomes are each student's average weekly software usage (measured in minutes of use per week), curriculum progress (measures as percent of the total curriculum completed each week), and efficiency of use (measured as progress per hour). For each student, we calculate the average outcome over

the five weeks[6] between when the first and second emails were sent, the weeks ending November 11 through December 9. We include students from all four agencies, as all participated in the first email.

$$y_{i,j} = \beta_0 + \beta_1 * \tau_j + y_{i,j,pre} + \gamma_{sg} + \varepsilon_{ij} \quad (1)$$

Equation (1) is an intent-to-treat (ITT) model where $y_{i,j}$ is the outcome (minutes, progress, or efficiency) for student *i* taught by teacher *j* averaged over the weeks after the first email was sent but before the second email was sent. We control for the treatment assignment of each student's teacher, $\tau_{j,}$ and include fixed effects for the school-grade span cells which were used for randomization, γ_{sg}. To improve precision, we additionally control for each student's average outcome measure during the five weeks prior to the first email, between the weeks ending September 30 and October 28, $y_{i,j,pre}$.

In addition to evaluating all teachers, we also run the model for two subsets of teachers: those who received an email informing them that their use was below that of their peers, and those who were informed that their use was above that of their peers. Because of the comparative language in the email, we had a strong *a priori* belief that the intervention would have a differential impact on teachers according to the message they received. Therefore, in addition to estimating an overall treatment effect, we also run the model separately for each of those two

[6] Site A and Site C were closed for Thanksgiving during the entire week of November 25th. For them, average use is calculated by dividing total use by four weeks, rather than five.

groups of teachers. The comparison group in both cases is the control teachers who were also below or above their peers, and would have received the same message.

3.3.2 Annual impact.

In our second model, we measure the impact of the emails over the course of the whole year, rather than just for the first email. As in the first model, we evaluate the impact in terms of the same three outcomes, minutes of use, curriculum progress, and efficiency of use. In terms of specification, this model is also a straight forward ITT model, nearly identical to the Equation (1), except the outcomes for this model are calculated over the entire period between the first email and the last week of school[7,8].

As before, we had a strong *a priori* belief that the impact of the messages would differ between teachers with relatively high and relatively low use. For the evaluation of the first email, we could directly stratify by the message each teacher received, as it was the first communication they received from us, and was therefore exogenous. When estimating an impact over the course of the year, however, this becomes less feasible. For one, teacher usage fluctuates over the year, so teachers received different messages at different points. Additionally, we expected the treatment group to be responding to the emails, so messages received after the first are endogenous, as they are a function of both use and response to the preceding messages.

---

[7] LEAs were open for different numbers of weeks, due to differing vacation schedules and school end dates. Site A was open for a total of 27 weeks after the first email was sent, Site B for 29, and Sites C and D were open for 23 weeks.

[8] These annual models exclude site A and D for the reasons mentioned in section 2.1.

We therefore stratify by teacher's average use during the baseline period, and whether that was above or below the average within their school and grade span during the baseline time period, the five weeks ending September 30 through October 28. Because this distinction is made using activity before the intervention began, it is exogenous.

3.3.3 Value added model.

Although software activity is our proximal outcome, the reason to focus on implementation of educational software and increasing use is that we ultimately believe that will raise student achievement, as measured by test scores. To this end, we run a value added model to estimate the effect of being in the treatment group on student achievement. A value added model is designed to show growth in student achievement relative to average growth within the sample. It does this by controlling for prior student achievement, as measured by prior test scores, each student's demographic characteristics, and the demographics of each student's peers.

$$y_{i,j,k,spring} = \beta_0 + \beta_1 * \tau_j + Y_{i,fall} + S_i + P_i + \gamma_{sg} + \varepsilon \quad (2)$$

Where: $y_{i,j,k,spring}$ is the spring formative test score for student *i* taught by teacher *j* in subject *k;* τ_j is an indicator for whether teacher *j* is in the treatment group and $Y_{i,fall}$ is a set of controls for student *i*'s fall formative test scores in both math and ELA. We model baseline achievement using a flexible cubic form of math and ELA. S_i is set of controls for student demographics, including gender, race, free or reduced price lunch, English language learner status, whether the student has an individualized education program and the student's grade. P_i represents a set of peer characteristics (i.e. the average of S_i among student *i*'s classmates), and γ_{sg} is a set of school-grade span indicators.

Unlike a traditional value added model, which measures spring-to-spring growth on state summative test scores, here we use fall-to-spring growth on formative assessments. We do this because such a large portion of our sample is outside of the traditional tested grades; typically only grades 3-8 are given the state assessments, and the need for a prior test score would further limit us to those in grades 4-8. There are concerns about the amount of time students spend on these formative tests in the fall and how seriously students take them (Pane et al., 2015), but they are the best measure we have available.

As in our model of annual usage impacts, we are interested in differences between teachers with above or below average use. Due to the same endogeneity concerns around using the message each teacher received for a year-long evaluation, we split our sample according to the same measure of baseline software use as explained in section 3.3.2.

4. Results

4.1 Response to the First Email

Figure 1 shows mean usage, in minutes per week, among students in both of the treatment arms compared to the control arm. For all teachers pooled together, we see no apparent effect. When the sample is split according to which message teachers received (or would have received, in the case of control teachers) we do see an apparent response, whereby treatment teachers who were told they were below average increased use relative to their control group, while teachers who were told they were above average decreased their use relative to their control group.

Table 3 shows the result of our student-level model over this time period, as described in equation 1. Each column shows a different model: column (1) includes all treatment and control teachers, columns (2) restricts the sample to only teachers who were below average use and column (3) restricts the sample to only teachers who were above average. The three main rows show results across our three software activity outcomes of interest, average weekly usage of the software, in minutes, average weekly progress in the software, in percentage of the curriculum, and average efficiency, in percent curriculum progress per hour of use.

Column (1) using all teachers shows no effect on the three outcomes. Among teachers told they were below the school average in column (2), we see a positive and significant effect on usage of the software, our primary short-term outcome and an accompanying positive and significant effect on curriculum progress. Importantly, the change in progress relative to the change in usage aligns with typical efficiency rates (approximately 2% progress per hour), and there is no change in overall efficiency. From this result, we are comfortable concluding that the intervention is not adding low-value time (e.g., placing students on the computer, but not helping

them progress). Lastly, among teachers who were told they were above average, shown in column (3), we see no significant effect on any outcome.

4.2 Annual Impact

We now move to our annual model, and restrict the sample to the two sites, B and C, who participated throughout the year and where we can identify above and below average teachers. As mentioned in the Methods section, we define high and low use teachers based on their relative performance during the five weeks prior to our first email, rather than based on which messages the teachers actually received, due to endogeneity concerns.

Figure 2 appears to show some overall differences between treatment and control. Among teachers with below average use during the baseline period, we see a stronger response. The picture for teachers with above average baseline use appears to show a drop in use among the treatment group, but the magnitude is smaller.

The results in table 4 broadly mirror what we see graphically and what we saw for the first email, in table 3, despite our standard errors almost doubling due to the loss in sample. In column (1), we do not see a significant effect on either usage, progress, or efficiency for the full teacher sample. Among teachers with below average use during the baseline period, in column (2), we see a positive and marginally significant effect on usage over the year after the emails began, similar in magnitude for teachers who were told they were below their peers in table 3. We also see a marginally significant effect on progress, which again is of a similar magnitude to the results from table 3, and in line with the estimated change in minutes, indicating that the

added time was not low-value, and assuaging concerns that teachers are merely leaving students logged in for longer to game the system.

Although these effects are relatively small, over the course of the year they would represent roughly an additional two hours of use and five percentage points of progress through the curriculum for low using teachers, roughly corresponding to adding two weeks of typical use to the year.

The possible negative effect on usage among high users, as seen in figure 2, is apparent in column (3), although roughly half the magnitude of the corresponding estimate in column (2), but it is not statistically significant. We see similar patterns for progress.

4.3 Value Added Model

Using our student test score model, we show estimated impacts on both math and ELA test scores. Because the emails focused on use of math software, and the software contains little or no language, we do not expect this intervention to raise ELA test scores[9]. Accordingly, we use ELA scores as a placebo test. As with the annual model, the sample is restricted to Sites B and C. Similar to the annual model, teacher's above- and below-average designations are made on the basis of average use during the baseline period, rather than by message received.

---

[9] This does not hold in the other direction: there are plausible explanations for how this intervention could decrease ELA scores, e.g. by increasing time spent on math at the expense of time spent on ELA. Also, non-cognitive factors, like persistence, could be enhanced by struggling through math problems that would also have an effect on ELA scores, but we would expect these to be small in magnitude.

We do not expect impacts on software outcomes either overall or among teachers with above average use given our results on interim outcomes and table 5 confirms these result showing no test score impacts in columns (1) or (3). However, column (2) also shows no significant impact on student test scores. Given our reduced sample size, the null result is not unexpected. Using non-experimental estimates of the relationship between usage and end of year achievement (Forthcoming white paper), a back of the envelope calculation indicates that an increase in usage of roughly five minutes per week, would result in expected test score gains of 0.01 standard deviations. We are not powered to detect this effect size.

5. Conclusion and Discussion

Education policy-makers are increasingly excited for nudge-based interventions due to their relatively low cost and preliminary evidence of their effectiveness. Our study adds to this emerging literature by assessing whether providing personalized information with a social comparison to teachers can enhance the fidelity of implementation of educational software across four local education agencies. We find moderate increases in software usage and syllabus progress initially, as well as throughout the year, among teachers with relatively low use. These improvements do not appear to be adding low-value time (e.g., placing the students in front of the computer, but not helping them learn concepts) because we do not see changes in the efficiency of the students, and the observed impacts on progress and usage are consistent with each other. Yet, with these moderate increases, we are unable to detect an effect on standardized test scores.

Our results provide important policy implications. Data-driven decision-making can change classroom practice, however practitioners often use data in simplistic ways (Marsh, Pane and

Hamilton, 2006; Oláh, Lawrence, and Riggan, 2010; Nelson et al., 2012) and it is hard to know which metrics to use when. Our results show that providing important data in an easily accessible nudge can be persistently effective in changing teachers' practice as long as the communication continues. It should be noted that for the site who only participated in the first email, but then decided to remove themselves from the study, we did not observe any difference between treatment and control after the initial five weeks following the first email. This implies that sending a single notification to teachers is not sufficient.

Second, these types of interventions can be done at a relatively low cost. On average, it took the staff at each LEA less than an hour to send each email round. After the initial upfront cost of writing code to combine the data and generate the emails, these interventions can be done relatively easily.

Although the effect of the nudge alone isn't enormous, we see potential for using nudges like this to reinforce a broader professional development initiative. Our findings suggest that nudges could be used as an informal, low touch, way to remind teachers about the professional development, and to report their level of implementation as a way of increasing the effort teacher's put into enacting the changes. The comparison to other teachers could also be used to facilitate partnering between teachers who are struggling to implement with teachers who are doing well. This pairing of high- and low-performing teachers has been shown to work for coaching (Papay et al., 2016), and could be extended within a framework such as this.

While encouraging, our work has several implications for future work in the field. First, while the intervention itself was relatively low-cost, based upon anecdotal feedback from our partners, these emails generated a desire among teachers to understand best practices in the field to improve their classroom use of software. Providing this type of additional support is not cost-

free. To accommodate these concerns, it appears as this field continues to grow, nudges either need to be accompanied by supplemental professional development or concise actionable steps to improve teacher practice. Providing comparison information with generic "tips" on its own is insufficient, and adding useful follow-up will increase the cost of the interventions.

Second, the below-average teachers increased their usage of the software, yet we do not know the opportunity cost of that additional time. For example, teachers could have reduced ELA software or class instruction time in order to accomplish the gains we observe. Since these interventions do appear change teacher behavior, it will be important for local education agencies to think through their priorities and assure the nudges align.

Similarly, sites need to carefully think through the context, particularly when using social norms to influence motivation. We saw evidence of this in three areas. First, as noted, one of the sites did not participate after the first email because some teachers in the district felt that the comparison was punitive, though this was not the district administration's intent. Furthermore, we observed that high-use teachers may have reduced their usage if they interpreted the message as saying that their time on the software was excessive. Finally, we saw some suggestive evidence (noted in the Appendix) that providing unrealistic goals (i.e., the vendor targets, which were significantly above average school usage) reduced the effectiveness of the intervention for teachers who were both below and above their school average. Past work with students (Karlson and Varhaug, 2016) provided evidence that pushing goals too high could have negative effects. Therefore, agencies should be careful to clearly lay out expectations and purpose before sending these nudges to ensure they achieve the goal of boosting the desired outcome.

Finally, due to better systems that can provide data more frequently and the importance of relatively small changes in language and context for nudge interventions, experimental designs

must be adaptable. As we noted, the intervention changed mid-year but we were unable to separately assess these changes and their value. As such, researchers must increasingly utilize methods like factorial design, sequential multiple assignment randomized trial (SMART), and Bayesian adaptive methods as well as be appropriately powered to learn more quickly and under what conditions a nudge can be most effective. The use of these methods in conjunction with nudges could become a powerful tool in the local policy-maker's tool box.

6. References

Bergman, P., & Hill, M. (2015). *The effects of making performance information public: evidence from Los Angeles teachers and a regression discontinuity design* (CESifo Working Paper Series No. 5383). Retrieved from http://www.columbia.edu/~psb2101/BergmanHillMakingPerformancePublic.pdf

Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability*, 21(1), 47-65.

Damgaard, M. T., & Nielsen, H. (2018). Nudging in education. *Economics of Education Review, 64*(C), 313-342.

Jackson, K., & Makarin, A. (2018). Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment. *American Economic Journal: Economic Policy, 10*(3), 226-54.

Karlsen, A. M. H., & Varhaug, M. (2016). *Kan nudging øke oppmøtet til lærerstudiet?: en studie initiert av Kunnskapsdepartementet* (Master's thesis).

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). Making sense of data-driven decision making in education. RAND.

McNaughton, S., Lai, M. K., & Hsiao, S. (2012). Testing the effectiveness of an intervention model based on data use: a replication series across clusters of schools. *School Effectiveness and School Improvement*, 23(2), 203-228.

Nelson, T. H., Slavit, D., & Deuel, A. (2012). Two dimensions of an inquiry stance toward student-learning data. *Teachers College Record*, 114(8), 1-42.

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, 85(2), 226-245.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data* (NBER Working Paper No. 21986). Retrieved from https://www.nber.org/papers/w21986.pdf

Pope, N. (2015). *The effect of teacher ratings on teacher performance* (Working paper). Retrieved from http://www.econweb.umd.edu/~pope/la_ny_paper.pdf

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2011). Classroom assessment for student learning: impact on elementary school mathematics in the central region. (NCEE 2011-4005). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/central/pdf/REL_20114005.pdf

The Data Quality Campaign. (2018). Education Data Legislation Review: 2018 State Activity. Retrieved from https://2pido73em67o3eytaq1cp8au-wpengine.netdna-ssl.com/wp-content/uploads/2018/09/2018-DQC-Legislative-Summary.pdf

Table 1. Overview of our sample

	Number of Participating Schools	Number of Participating Teachers	Number of Participating Students
Site A	36	846	21,877
Site B	14	138	3,464
Site C	13	235	5,301
Site D	2	8	458
Total	**65**	**1,227**	**31,100**

Table 2. Demographics of our sample

Measure	**Control Mean**	**Treatment Mean**	**Adjusted Difference (SE)**	**Control SD**	**Treatment SD**	**Number of Students**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Baseline average weekly minutes on the software	71.23	71.86	1.191 (1.5999)	42.17	43.33	31,100
Baseline average weekly progress in the software (% of curriculum)	2.24	2.18	-0.024 (0.0563)	1.94	1.88	31,100
Baseline average efficiency (progress per hour)	2.00	1.88	-0.089 (0.0819)	1.38	1.29	30,708
Proportion Male	0.51	0.51	-0.007 (0.0049)	0.50	0.50	31,100
Proportion Black	0.13	0.14	<0.001 (0.0022)	0.33	0.35	31,100
Proportion Hispanic	0.83	0.82	-0.002 (0.0030)	0.38	0.39	31,100
Proportion designated Limited English Proficiency	0.67	0.65	-0.005 (0.0048)	0.47	0.48	31,100
Proportion on an Individualized Education Program	0.09	0.10	0.009 (0.0054)	0.29	0.30	31,100
Proportion receiving free or reduced price lunch	0.88	0.89	0.004 (0.0047)	0.32	0.31	31,100
Average fall formative test score – English Language Arts	-1.22	-1.22	-0.010 (0.0286)	0.99	1.03	23,827
Average fall formative test score – Math	-1.47	-1.49	-0.032 (0.0254)	0.99	1.02	29,217

Note. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Standard errors are clustered at the school-grade span level. The sample includes all students who were present during the week before the first email was sent. Formative test scores are standardized using national means and standard deviations by grade and subject. We create a concordance of different formative tests using common state assessments. Students with zero baseline use are necessarily missing an efficiency measure. The number of students with fall formative math test scores is low due to missing test scores, as some students may have been absent the day of testing, or not yet enrolled in the sites. The number of students with fall formative ELA students is low for the same reasons, and also because site A does not test ELA uniformly in grades K-2.

Table 3. Response to the first email

	All teachers	Teachers receiving a low message	Teachers receiving a high message
	(1)	(2)	(3)
Average weekly minutes	0.52 (1.28)	4.53** (1.75)	-1.33 (2.08)
Average weekly progress	0.046 (0.051)	0.191*** (0.064)	-0.040 (0.090)
Average efficiency	0.018 (0.076)	0.210 (0.207)	-0.127 (0.080)
Number of teachers	1,227	546	451
Number of students	31,100	13,319	11,507

Note: *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Standard errors in parenthesis are clustered at the school-grade range level. The "all teachers" sample includes all teachers, and all students who were present the week before the first email was sent. The "low message" and "high message" samples exclude school-grade range cells where there was not at least 1 treatment and 1 control teacher who received/would have received that message. Of 124 school-grade range cells, 29 are excluded for the "low message" sample and 32 are excluded for the "high message" sample. Students with 0 minutes of use are necessarily excluded from efficiency models.

Table 4. Annual impact

	All teachers	Teachers with low baseline use	Teachers with high baseline use
	(1)	(2)	(3)
Average weekly minutes	2.71 (1.83)	5.38* (2.92)	-2.83 (2.25)
Average weekly progress	0.090 (0.056)	0.180* (0.097)	-0.083 (0.067)
Average efficiency	-0.066 (0.070)	-0.050 (0.066)	-0.036 (0.063)
Number of teachers	373	175	139
Number of students	8,765	3,954	3,219

Note: *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Standard errors in parenthesis are clustered at the school-grade range level. The "all teachers" sample includes all teachers in Site B and Site C, and all students who were present the week before the first email was sent. The "low use" and "high use" samples exclude school-grade range cells where there was not at least 1 treatment and 1 control teacher who had below or above average use, respectively. Of 52 school-grade range cells in the "all teachers" sample, 10 are excluded for the "low use" sample and 16 are excluded for the "high use" sample. Students with 0 minutes of use are necessarily excluded from efficiency models.

Table 5. Student test score results

	All teachers	Teachers with low baseline use	Teachers with high baseline use
	(1)	(2)	(3)
Formative spring test score – math	-0.024 (0.029)	0.006 (0.041)	-0.036 (0.041)
Formative spring test score – ELA	-0.004 (0.029)	0.019 (0.036)	0.019 (0.041)
Number of teachers	347	161	128
Number of students	7,845	3,304	2,857

Note: *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Standard errors in parenthesis are clustered at the school-grade range level. The "all teachers" sample includes all teachers in Site B and Site C, and all students who were present the week before the first email was sent and had a valid spring formative test score in math. The "low use" and "high use" samples exclude school-grade range cells where there was not at least 1 treatment and 1 control teacher who had below or above average use, respectively. Of 50 school-grade range cells in the "all teachers" sample, 10 are excluded for the "low use" sample and 18 are excluded for the "high use" sample. The different school-grade range cell samples is the reason the sign of the "all teachers" result does not necessarily align with the sign of the two subsamples.

Table A1. Demographics of annual impact sample

Measure	Control Mean	Treatment Mean	Adjusted Difference (s.e.)	Control SD	Treatment SD	Number of Students
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline average weekly minutes on the software	50.098	51.799	0.8173 (2.2249)	35.224	35.822	8765
Baseline average weekly progress in the software (% of curriculum)	1.480	1.528	0.0412 (0.0733)	1.491	1.470	8765
Baseline average efficiency (progress per hour)	1.843	1.788	-0.0016 (0.0493)	1.602	1.342	8538
Proportion Male	0.504	0.504	0.0002 (0.0094)	0.500	0.500	8765
Proportion Black	0.461	0.462	-0.0000 (0.0081)	0.499	0.499	8765
Proportion Hispanic	0.464	0.471	-0.0022 (0.0084)	0.499	0.499	8765
Proportion designated LEP	0.289	0.291	-0.0071 (0.0083)	0.454	0.454	8765
Proportion with IEP	0.103	0.105	-0.0012 (0.0113)	0.304	0.307	8765
Proportion FRPL	0.870	0.890	0.0053 (0.0070)	0.337	0.313	8765
Average fall formative test score - ELA	-1.157	-1.107	0.0233 (0.0429)	0.935	0.968	7281
Average fall formative test score – Math	-1.433	-1.381	0.0528 (0.0430)	0.994	1.007	7617

Note. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Standard errors are clustered at the school-grade range level. The sample includes all students in Sites B and C who were present during the week before the first email was sent. The number of students with fall formative math test scores is low due to missing test scores, as some students may have been absent the day of testing, or not yet enrolled in the sites.

Table A2. Response to first email, differential impact by arm

		All teachers	Teachers receiving a low message	Teachers receiving a high message
		(1)	(2)	(3)
Average weekly minutes	Main treatment effect (β_1)	1.62 (1.27)	5.18** (2.17)	1.63 (2.32)
	Vendor arm difference (β_2)	-2.20 (1.62)	-3.46 (2.89)	-4.30 (2.87)
Average weekly progress	Main treatment effect (β_1)	0.093* (0.054)	0.199** (0.078)	0.061 (0.118)
	Vendor arm difference (β_2)	-0.095 (0.069)	-0.110 (0.102)	-0.126 (0.138)
Average efficiency	Main treatment effect (β_1)	-0.034 (0.039)	0.072 (0.079)	-0.168 (0.111)
	Vendor arm difference (β_2)	0.106 (0.134)	0.398 (0.394)	0.047 (0.066)
Number of teachers		1,227	459	364
Number of students		31,100	11,257	9,406

Note: *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Standard errors in parenthesis are clustered at the school-grade range level. The "all teachers" sample includes all teachers, and all students who were present the week before the first email was sent. The "low message" and "high message" samples exclude school-grade range cells where there was not at least 1 treatment and 1 control teacher who received/would have received that message. Of 124 school-grade range cells, 56 are excluded for the "low message" sample and 63 are excluded for the "high message" sample. Students with 0 minutes of use are necessarily excluded from efficiency models.

Table A3. Response to first email, differential impact by whether student was named

		All teachers	Teachers receiving a low message	Teachers receiving a high message
		(1)	(2)	(3)
Average weekly minutes	Main treatment effect (β_1)	-0.12 (1.33)	3.79** (1.85)	-1.21 (2.13)
	Named student difference (β_3)	1.62 (1.91)	1.11 (2.32)	-1.38 (2.68)
Average weekly progress	Main treatment effect (β_1)	0.024 (0.053)	0.173** (0.067)	-0.046 (0.092)
	Named student difference (β_3)	0.044 (0.068)	0.023 (0.087)	0.011 (0.116)
Average efficiency	Main treatment effect (β_1)	0.064 (0.083)	0.216 (0.208)	-0.041 (0.048)
	Named student difference (β_3)	-0.322* (0.187)	-0.122 (0.127)	-0.716 (0.514)
Number of teachers		1,227	546	451
Number of students		31,100	13,319	11,507

Note: *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. Standard errors in parenthesis are clustered at the school-grade range level. The "all teachers" sample includes all teachers, and all students who were present the week before the first email was sent. The "low message" and "high message" samples exclude school-grade range cells where there was not at least 1 teacher in each treatment arm and 1 control teacher who received/would have received that message. Of 124 school-grade range cells, 29 are excluded for the "low message"

sample and 32 are excluded for the “high message” sample. Students with 0 minutes of use are necessarily excluded from efficiency models.

Figure 1. Response to the first email

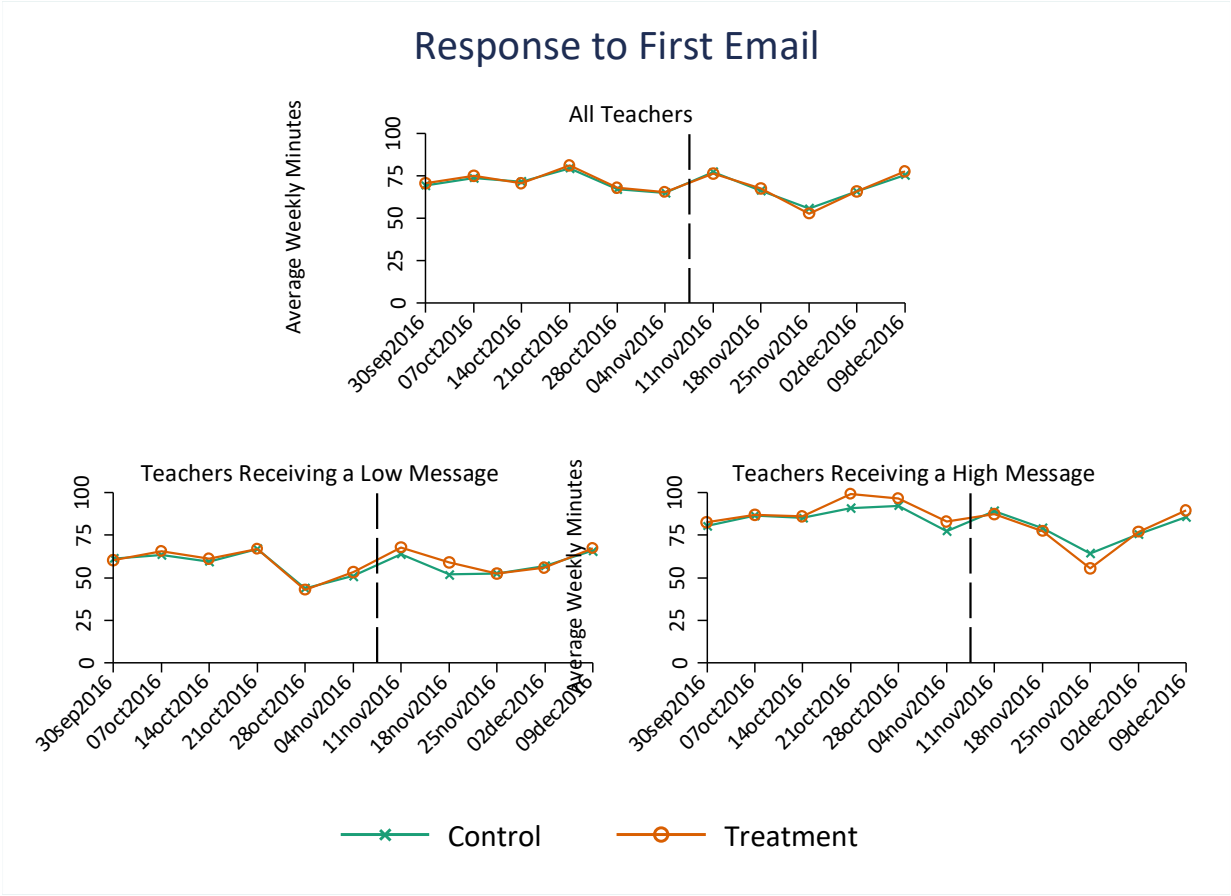


Figure 2. Annual impact

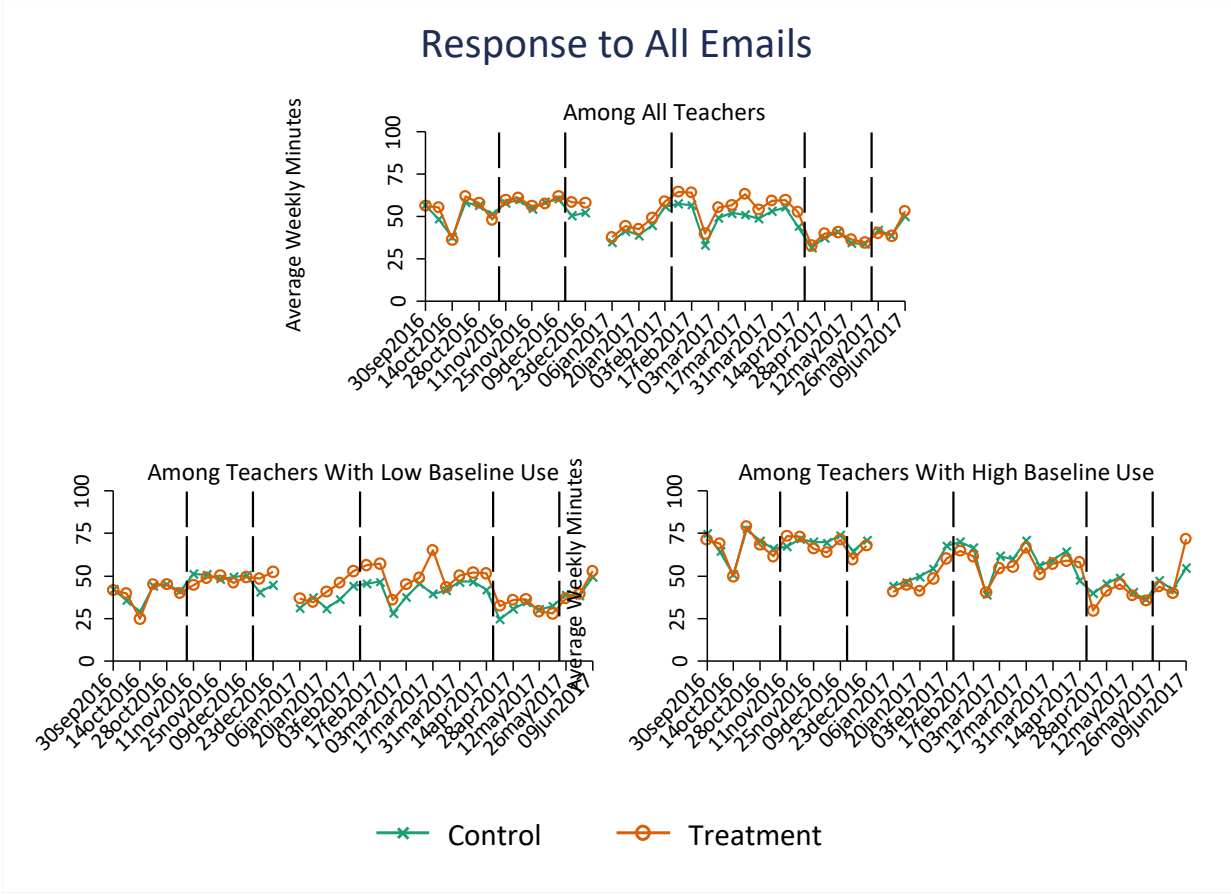


Figure A1. Example message, peer comparison only

Dear George Washington,

Your class used [redacted] an average of 48 minutes during the week of October 22nd - 28th, which is less than other classes in grades K-1 in Lincoln Elementary.

The following three students had the lowest [redacted] usage in your class:

Name	Average minutes per week
John Adams	5
Thomas Jefferson	8
James Madison	12

Students learn more and earn higher scores in mathematics when they use [redacted] in addition to receiving high quality classroom instruction.

You can have a big effect on your students' [redacted] usage going forward – and we appreciate your help. For additional [redacted] resources, please visit [redacted].com.

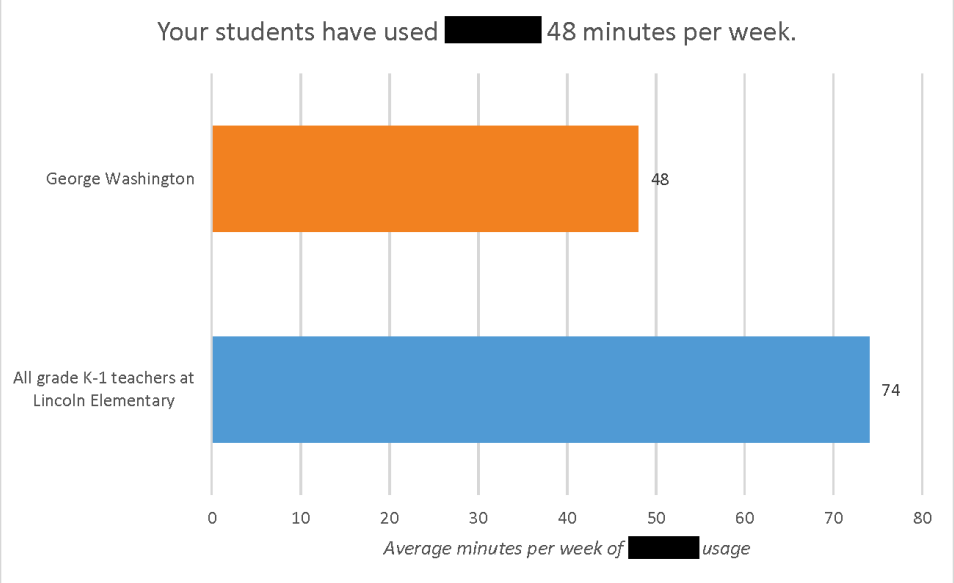


Figure A2. Example message, peer and vendor comparison

Dear George Washington,

Your class used [redacted] an average of 48 minutes during the week of October 22nd - 28th, which is less than other classes in grades K-1 in Lincoln Elementary, and less than the vendor recommends.

The following three students had the lowest [redacted] usage in your class:

Name	Average minutes per week
John Adams	5
Thomas Jefferson	8
James Madison	12

Students learn more and earn higher scores in mathematics when they use [redacted] in addition to receiving high quality classroom instruction.

You can have a big effect on your students' [redacted] usage going forward – and we appreciate your help. For additional [redacted] resources, please visit [redacted].com.

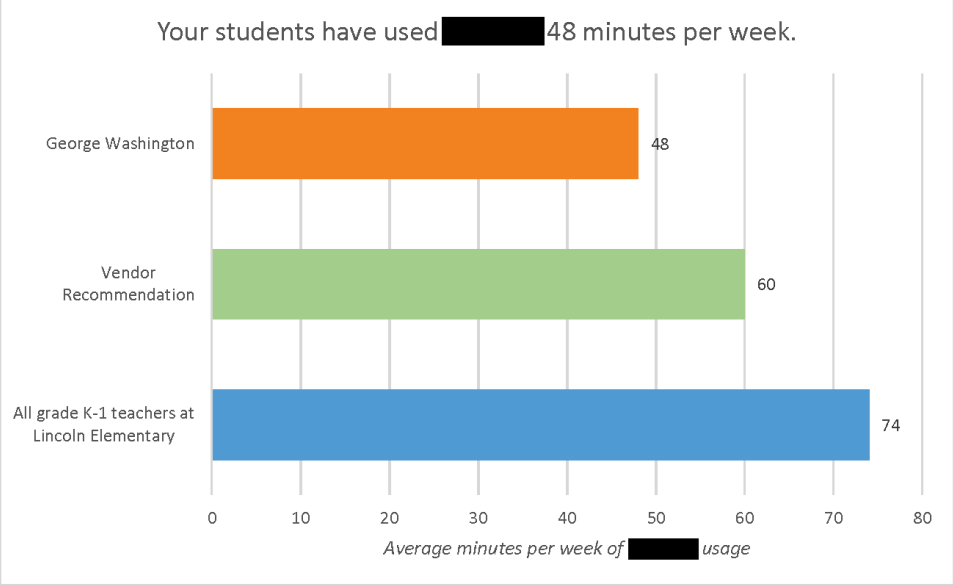


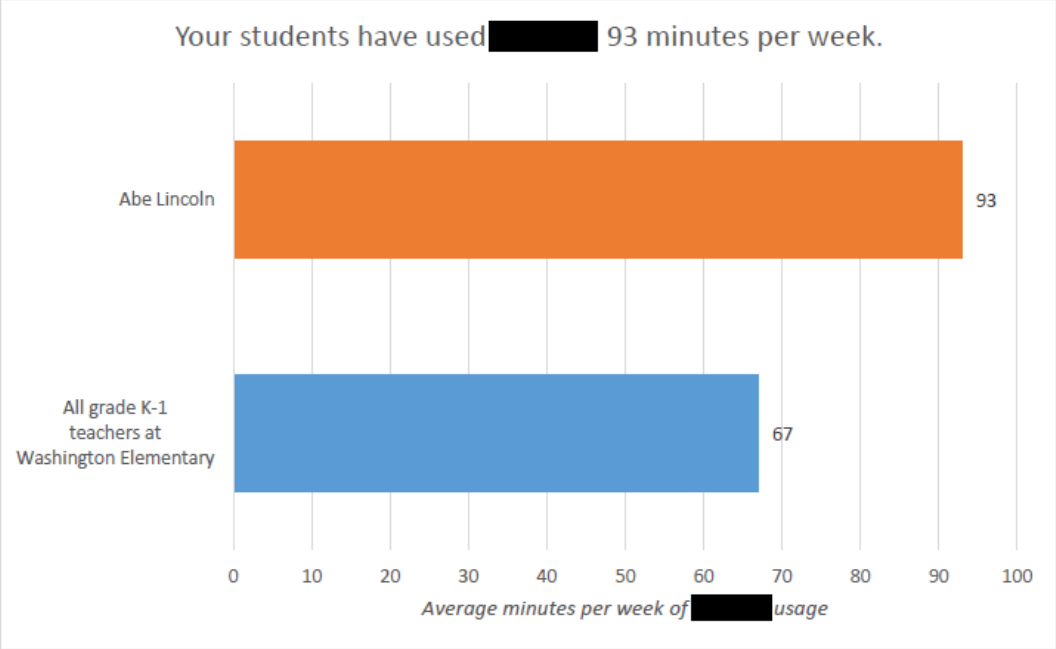
Figure A3. Example message, midyear revision

Dear Abe Lincoln,

Your class used [redacted] an average of 93 minutes during the week of April 1st - 7th, which is more than other classes in grades K-1 in Washington Elementary; keep up the good work!

Students learn more and earn higher scores in mathematics when they use [redacted] in addition to receiving high quality classroom instruction.

You can have a big effect on your students' [redacted] usage going forward – and we appreciate your help. For additional [redacted] resource, please visit [redacted].com.



7. Appendix

7.1 Demographic Balance Among Annual Impact Sample

Table A1 shows the demographic balance of the treatment and control groups for Sites B and C, the only sites included in the annual impact and value-added models. The groups are balanced on observable measures, with no statistically significant differences. Additionally, the standard deviation of all measures appear to be roughly equal across the two groups. Compared to the full sample demographics, seen in table 2, the restricted sample has lower software usage, a higher proportion of black students, and a lower proportion of Hispanic students. A smaller share of students in this sample are designated LEP, but the share with an IEP and receiving FRPL are approximately the same. Fall test scores are also moderately lower.

7.2 Response to First Email by Treatment Arm

This model evaluates the impact of the first email differentially by treatment arm. The first treatment arm received emails comparing their use to the use of other teachers in their school and grade span while teachers in the second treatment arm (the vendor comparison arm) received additional information comparing their use to the software vendor's recommended weekly use.

This model is nearly identical to the model in equation 1, used to evaluate the response to the first email. The one difference is that it includes an indicator for whether a student's teacher was in the vendor comparison arm. This term lets us estimate the difference in the impact of the email between the two treatment arms, and it is our coefficient of interest.

$$y_{i,j} = \beta_0 + \beta_1 * \tau_j + \beta_2 * \tau_{vendor\_arm,j} + y_{i,j,pre} + \gamma_{sg} + \varepsilon \quad \text{(A1)}$$

Where $y_{i,j}$ is the for student *i* taught by teacher *j* averaged over the weeks after the first email was sent but before the second email was sent; $y_{i,j,pre}$ is the outcome *before* the first email was sent. τ_j is an indicator for whether teacher *j* is in either arm of the treatment group and $\tau_{vendor\_arm,j}$ is an indicator for whether teacher *j* is in the vendor comparison arm of the treatment group. As before, γ_{sg} is a set of school-grade range indicators.

The estimates for the main treatment effect in table A2 in all columns for all outcomes are similar to the results presented in table 3, and again we see a positive and significant overall effect on teachers below their school average, in column (2). None of the estimates for the vendor arm impact difference are significant, so we do not have evidence that the effect differed by treatment arm. In fact, while not significant, we see a negative impact of the vendor comparison across all teachers. Because the coefficients on the vendor arm difference are all negative for minutes and progress, our main proximal outcomes, we decided to remove the vendor comparison for all teachers in order to simplify the messaging. These negative effects also could be related to providing unrealistic comparisons (Karlson and Varhaug, 2016).

7.3 Response to First Email Among Named Students

This model evaluates the impact of the first email by whether a student was one of the three students identified as having the lowest usage during the week preceding week the email. The model is nearly identical to the model in equation 1, used to evaluate the response to the first email. It differs by including an indicator for whether a student was named in the first email, or

would have been named in the case of the control group (i.e. was one of the three lowest use students). The model also includes the interaction of this indicator with our treatment indicator. In this way, we can identify both the main effect of the treatment and the differential impact of the treatment among named students.

$$y_{i,j} = \beta_0 + \beta_1 * \tau_j + \beta_2 * named_{i,j} + \beta_3 * \tau_j * named_{i,j} + y_{i,j,pre} + \gamma_{sg} + \varepsilon \quad \text{(A2)}$$

Where $y_{i,j}$ is the for student *i* taught by teacher *j* averaged over the weeks after the first email was sent but before the second email was sent; $y_{i,j,pre}$ is the outcome *before* the first email was sent. τ_j is an indicator for whether teacher *j* is in either arm of the treatment group and $named_{i,j}$ is an indicator for whether student *i* was (or would have been, for control teachers) listed as one of teacher *j*'s three lowest-use students. As before, γ_{sg} is a set of school-grade range indicators.

Once again, the estimates for the main treatment effect in table A3 in all columns for all outcomes are similar to those seen in table 3. None of the estimates for the impact difference among named students are significant, implying that teachers were not making any specific effort to help these students. The coefficient in column (1) for the efficiency outcome is marginally significant, but because there is no readily apparent explanation, and only one of 9 named student coefficients tested is marginally significant, we believe this is merely spurious.

Because of the lack of evidence of effectiveness of this portion of the message, combined with anecdotal concerns about the accuracy and actionability of the named students (specifically that many of them had either transferred from the class or been sick for the week), we decided to remove the named students portion of the message when we revised the intervention midyear.