PROVING GROUND

MAY 2020

Exploring ST Math's Implementation and Impact

A REPORT FROM THE PROVING GROUND NETWORK

Ethan Scherer, Sidrah Baloch, Lauren Dahlin, Thomas J. Kane, Douglas O. Staiger, Dan Thal, Jon B. Fullerton, Ashley Snowdon MacDonough

Proving Ground worked with 13 school systems across the country to assess the impact and implementation of ST Math, a mathematics educational software program. We found that students who increased their average weekly usage of ST Math from fall to spring experienced **higher math test score gains** than students whose usage remained the same.These gains were **highest for students with lower baseline usage**. Across sites, students used ST Math **less than the vendor's recommended amount**, though there was **considerable variation** within each site. Teachers and schools were responsible for **most of this variation**. *[Click on each finding to jump to its section.]*

Proving Ground's[1] first network of school districts and charter management organizations launched in 2015 to examine the use of educational software[2] in schools and classrooms. We worked with 13 school systems across the country to assess both the impact and implementation of ST Math, a mathematics educational software program designed for grades K–8 with a particular focus on elementary grades.[3] Between the 2015–16 and 2016–17 school years, we also worked with these 13 agencies to develop, pilot, and test strategies to improve the use of ST Math with the ultimate goal of boosting student learning and achievement.

The 13 partner agencies, which included three traditional school districts and 10 CMOs, were located in urban areas across five states. The network was comprised of approximately 230,000 students, with partners varying in size from a few thousand to 56,000 students. Eighty percent of those students were eligible for free or reduced-price lunch and 80% were African American or Hispanic. By the end of the 2015–16 school year, all but one of the 13 partner sites had been using ST Math for at least three years, and most of them had rolled it out to 60–80% of eligible classes.

HISTORICAL IMPACT RESULTS

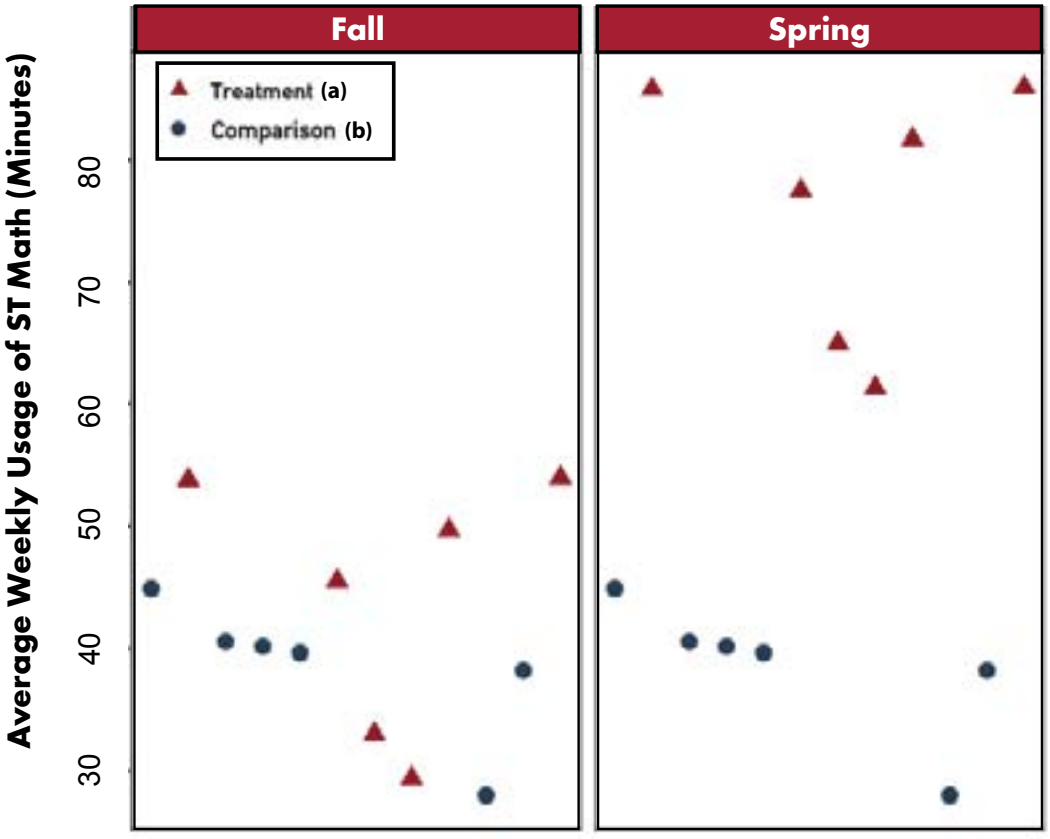
Our partner agencies wanted to know whether students were experiencing any meaningful growth in math test scores as a result of using ST Math. Because our partners had already rolled out ST Math widely by the 2015–16 school year, we were unable to do a randomized controlled trial to assess the impact of the software. Instead, we compared two categories of ST Math users: (a) students whose average weekly usage increased by over 10 minutes from the fall semester to the spring

How Does ST Math Work?

ST Math teaches math concepts visually, initially avoiding the unnecessary complexity of abstract symbols and language. The animated interactive puzzles must be mastered to advance, and provide personalized informative feedback. This visual approach is accessible to all student subgroups. The program is structured similarly to a textbook, with students progressing through a sequence of learning objectives taught through interactive games that are part of a grade-level syllabus chosen by their teacher. ST Math recommends a weekly usage level in minutes and an approximate amount of content that students should complete each week in order to complete a majority of their assigned content by the end of the school year.

semester, and (b) students whose average weekly usage did not significantly change from fall to spring.[4] Figure 1 provides an illustrative example of how the students were divided into the two categories. Students in the first category increased the time they spent using ST Math each week by an average of 32 minutes from fall to spring, compared to an average change of 0 minutes by students in the second category. If we assume that increased usage was the only major change in math instruction these students received, the difference in math test score gains between these two groups can be interpreted as the impact of increased ST Math usage on math achievement.

Figure 1. Illustrative Example of How the Treatment and Comparison Groups Changed



Key Finding: Increased ST Math usage was associated with higher math gains.

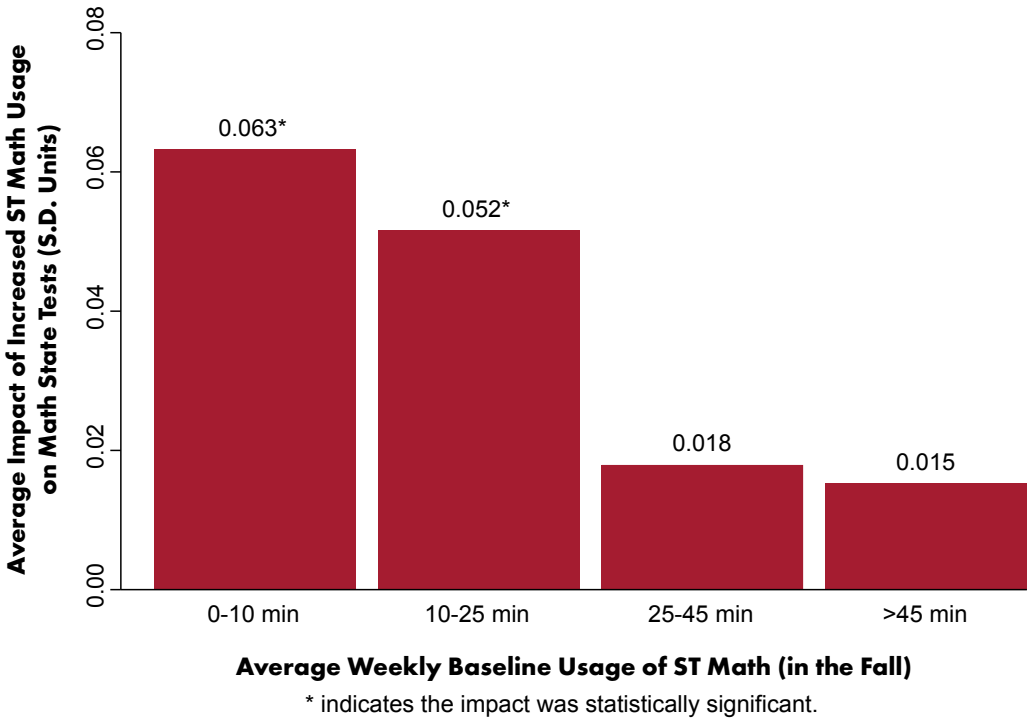
Using the above method, we estimated the impact of increased ST Math usage on mathematics end-of-year state tests [e.g., Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC)] for 4th and 5th grade students.[5] The pooled results for 2015–16 and 2016–17 indicated that increasing ST Math usage from fall to spring by more than 10 minutes was associated with higher math achievement gains. The size of this effect was close to 0.04 standard deviation units, which is comparable to approximately half the difference in effectiveness between a first- and fifth-year teacher. In order to test whether there were other factors driving the change in software usage and achievement, such as a

Understanding Standard Deviation Units

These impact results are reported in standard deviation units rather than test score units. This allows the analysis to combine the results of students in different grade levels who took different tests and calculate one overall effect size. It also allows us to compare these impact results to the impact of other interventions. For example, reducing class sizes from 22 to 15 students was found to improve student test scores by 0.15 standard deviation units.[6] In addition, teacher effectiveness research has found that the difference in math test score gains that students experience from having a novice teacher compared to teachers with five or more years of experience is 0.08 standard deviation units.[7]

general improvement in student health, motivation or instructional quality, we measured the impact on students' ELA scores as well. We did not see differential gains on the ELA state assessment, which would be consistent with the assumption that the treatment and comparison groups differed only in their increased ST Math usage.

Figure 2. Differences in Impact by Baseline Level of ST Math Usage



* indicates the impact was statistically significant.

Key Finding: The impact of increased ST Math usage was greater for students with low baseline levels of usage.

Among the students who increased ST Math usage from fall to spring, there were greater benefits for students starting from a baseline of lower levels of usage. Students who used the software for an average of 10 minutes or less per week in the fall experienced greater test score gains after increasing their usage in the spring than students who began the school year with previously higher average weekly usage rates (shown in Figure 2).

HISTORICAL IMPLEMENTATION RESULTS

School systems allocate money and resources to deliver educational software products to students and naturally want to know if they are being used and, if so, by whom. To provide partners with a complete picture of both the impact of the software and how it was being used, we complemented the effectiveness analyses just discussed with a description of how ST Math implementation

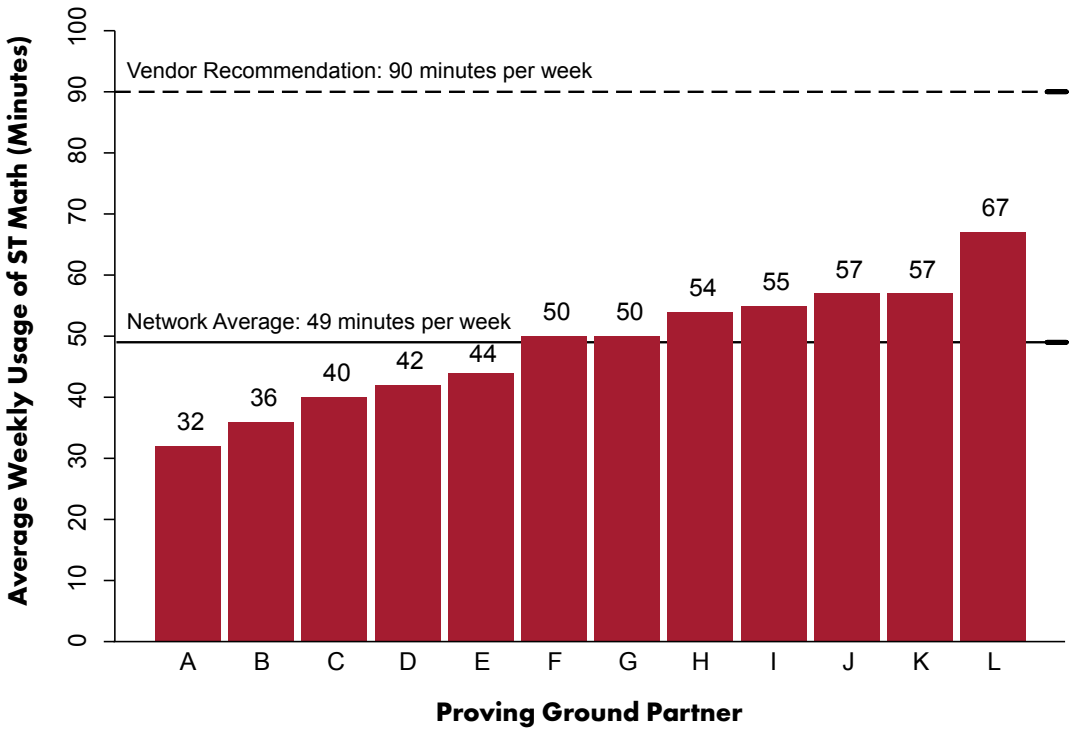
A Note about Proving Ground's Methodology

Because there are many factors that may influence both the amount of time students spend using ST Math and their performance on math assessments, several student characteristics were accounted for when calculating the impact of increased ST Math usage on achievement. Only students with similar baseline math and ELA test scores, demographic attributes, and fall ST Math usage levels were compared, so that the only apparent difference between students was whether or not they increased their usage of ST Math from fall to spring. We tested this assumption by examining students' end-of-year ELA scores, which would not be expected to change in response to using ST Math. If students who increased their usage of ST Math experienced greater math and ELA gains than other comparable students, there was likely some other explanation for the difference in their academic outcomes than the effect of the software. This served as a placebo test, similar to those discussed in a medical context. The results indicated no effect of increased ST Math usage on summative ELA test scores, suggesting that our assumptions were valid. (See technical appendix for a more detailed explanation of our methodology.)

varied across the Proving Ground network. Many of these analyses were based on questions that emerged through discussions with partners.

Our exploration relied on the most basic measures of ST Math usage: the average number of minutes per week students were logged into and using the software. A key limitation of focusing on average weekly minutes is that it does not address deeper questions about ST Math implementation, such as the best mode of delivery (e.g., computer lab or laptops in the classroom); how to schedule usage during the school day or week; or the type of teacher instruction that should accompany use of the software. Thus, the following findings should be viewed as a starting point to thinking about mode and quality of implementation.

Figure 3. ST Math Usage by Partner (Grade 2–5)



Key Finding: Average ST Math usage was lower than the vendor's recommended levels.

ST Math recommends that students in grades K–1 use the software for at least 60 minutes each week; the recommendation for grades 2–8 is 90 minutes per week. In partner sites, on average, students' weekly ST Math usage did not approach these vendor recommendations. Figure 3 shows the average weekly usage levels of the 12 partner agencies that had ST Math users in grades 2–5 during the 2016–17 school year.[8] None of the sites reached the recommended level of 90 minutes per week, despite most of them having implemented ST Math for at least three years.[9] There was some variation in the average level of usage across sites, but most were within 20 minutes of each other. This was also consistent over multiple years, as the network-wide average in 2016–17 was almost identical to the average in 2015–16.

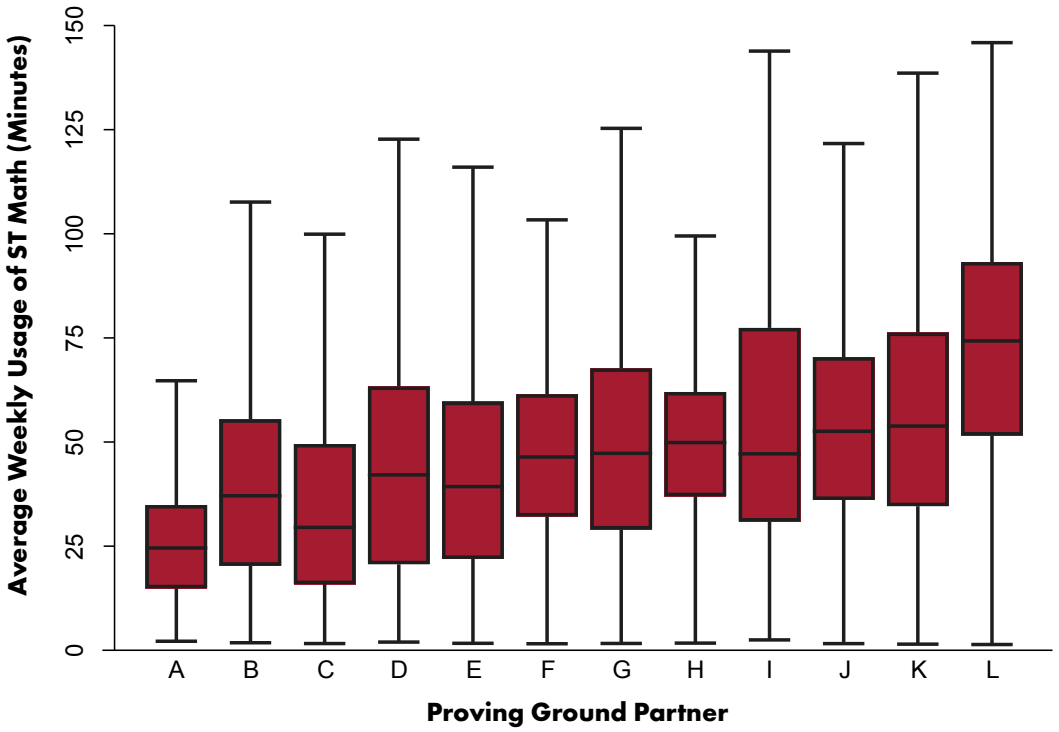
Key Finding: There was considerable variation in students' ST Math usage within sites.

While the usage of ST Math overall was below expectations, the box and whisker plots in Figure 4 show that there was considerable variation in students' average weekly usage within each site. For example, in site I, about 25% of the students spent less than 35 minutes logged into ST Math each week. However, another 25% of students used the software for more than 76 minutes each week, approaching the vendor's recommended usage.

Key Finding: Teachers and schools were responsible for most of the variation in students' ST Math usage.

We examined different factors that could influence students' usage of ST Math and found that usage varied much more by teacher and by school, than by student within the same classroom or school. Furthermore, we found that the average weekly usage of most teachers' classes was consistent from one year to the next. This meant that teachers whose classes had low average weekly usage of ST Math in one year were likely to have classes with low average weekly usage in the subsequent year. Similarly, teachers with high-using students in one year tended to have high-using students in the following year. Such findings suggest that much of the variation in student usage is due to differences in how teachers incorporate the software into their classroom practices.

Figure 4. Variation of ST Math Usage by Partner (Grade K–5)



How to Interpret a Box and Whisker Plot

Box and whisker plots help illustrate the distribution, or spread, of a data group. In Figure 4, the middle line of each box represents the median, or 50th percentile, of average weekly usage in a particular agency. The top and bottom of a box represent the 75th and 25th percentiles. The ends of the whiskers represent close to the maximum and minimum amounts of average weekly usage.

CONCLUSION

The findings discussed above were important for our partners as they thought about implementing ST Math. On the one hand, we found encouraging (though not conclusive) evidence that ST Math was effective with their students. On the other hand, our analyses found strong evidence that ST Math was not being implemented as designed and that implementation varied strongly across teachers. As a result, the benefits ST Math could provide were likely not being maximized.

In response to these findings, Proving Ground network members tried several strategies to improve the amount of usage of ST Math and the effectiveness of that usage, such as messages informing teachers about how their own students' usage compared to that of their colleagues' students. Details on the outcomes of these strategies can be found in a companion brief on the Proving Ground website.[10] However,

such strategic responses would not have been desirable or pursued without evidence that ST Math can have positive effects and the knowledge that it was being underutilized.

The information in this brief is not intended to provide specific advice or to endorse any particular educational software or program including ST Math. Our findings are based on data collected between 2015-17 in 13 districts.

Endnotes

1 Proving Ground, an initiative of the Center for Education Policy Research at Harvard University, seeks to make evidence-gathering and use a routine part of how education agencies conduct their daily work. Proving Ground brings together a network of education organizations to collaborate in solving shared challenges and supports the network with data analysis, strategic advice, hands-on assistance, and peer networking opportunities. Through our continuous improvement framework, Proving Ground helps partners understand the pressing challenges they face, rapidly identify potential solutions that align with those challenges, and test evidence-based solutions that work for their students, families, and schools. We empower partners to address challenges in their own context while benefiting from lessons learned across the network.

2 Educational software refers to online programs designed to deepen student learning by combining traditional classroom instruction with academic content tailored to individual children's needs and abilities.

3 Proving Ground worked with multiple software vendors between fall 2015 and spring 2017. ST Math was used by all 13 partner agencies, which facilitated collaboration across the entire network and allowed for the pooling of data.

4 Proving Ground used this comparison because comparing the test scores of ST Math users and non-users would not have yielded an accurate estimate of the software's impact. As partners had already rolled out ST Math widely, the minority of students who were not using it were likely different in some way from the majority of students who were. Proving Ground needed to find an alternate way to rigorously evaluate the impact of ST Math. To address this issue, Proving Ground focused only on students with basic levels of ST Math usage during a full school year (at least five logins and 50 minutes of usage over the course of a year) and examined whether those who increased their usage over time had achievement gains relative to those whose usage level remained the same.

5 State tests are administered in 3rd, 4th, and 5th grade. Because our methodology accounts for students' test scores from the previous year, our analysis includes only 4th and 5th grade students.

6 Schanzenbach, D. (2006). What Have Researchers Learned from Project STAR? Brookings Papers on Education Policy, (9), 205–228.

7 Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. Economics of Education Review, 27(6), 615–631.

8 The 13th partner did not have any students using ST Math in this grade band.

9 We observed similar patterns for grades K–1 and 6–8: no sites reached the vendor-recommended level of weekly ST Math usage.

10 provingground.cepr.harvard.edu/resources

Acknowledgments

We thank each of the members of the Proving Ground Network for both their partnership and providing the administrative data for this work. We additionally thank the leadership of MIND Research Institute for being collaborative research partners and providing helpful nuances of the software. Any errors of fact or interpretation are our own.

Suggested Citation

Scherer, E., Baloch, S., Dahlin, L., Kane, T. J., Staiger, D., Thal, D., Fullerton, J. B., & Snowdon MacDonough, A. (2020). *Exploring ST Math's Implementation and Impact: A Report from the Proving Ground Network*. Cambridge, MA: Center for Education Policy Research.