Exploring ST Math's Implementation and Impact

PROVING GROUND

TECHNICAL APPENDIX

MAY 2020

Methods for Achievement Analysis

After assessing the Proving Ground network-wide context, we determined a matched comparison group design was the most rigorous analytic method to evaluate ST Math's impact on student achievement. Because most Proving Ground agencies had been implementing ST Math for several years before the network launched, we were unable to compare randomly-assigned treatment and control groups. Furthermore, though we did obtain longitudinal data from partner sites going back several years, we could not take advantage of longitudinal data because many sites did not consistently administer state tests during the transition to Common Core State Standards. Considering these limitations, we determined that a matched comparison group design was the best available quasi-experimental method for estimating the effect of ST Math. The goal was to match students who increased their usage of ST Math from fall to spring with students with very similar baseline characteristics whose usage remained relatively constant.

We created matched comparison groups using propensity score blocking. Sections B and C provide more detailed information about how we implemented the blocking method and the statistical models we used.

A. LIMITATIONS

Although this analytic approach was the most appropriate option for the situation, it has some limitations that must be mentioned. While we did confirm that there were few observable differences between the treatment and comparison groups, there is always a chance that the two groups differed in ways not captured by the data. If treatment students and comparison students were systematically different in a way that influenced growth on test scores, we may have falsely attributed these differences to the effect of increased ST Math usage. For example, we were unable to accurately link teachers to students in some of our partner agencies. If it was the case that students of highly effective (or ineffective) teachers were disproportionately represented in either the treatment or comparison group, our results could be biased. A general change in student motivation or instructional quality should have affected both math and ELA scores. Thus, we tested whether the students with the increase in ST Math usage also saw an increase in ELA scores (i.e. a "placebo" outcome). There was no improvement in ELA scores. Nevertheless, because we cannot rule out that there were other factors driving both the increase in software usage and math achievement, our impact results should be interpreted with some caution.

B. MATCHING METHODOLOGY

We examined the average weekly number of minutes students spent using ST Math in the fall semester and the spring semester of each school year. We defined the treatment group as students whose spring average weekly usage was more than 10 minutes greater than their fall average weekly usage. These were students who increased their usage of ST Math over the course of the year. We defined the comparison group as students whose average weekly usage remained stable (that is, their average weekly usage changed by less than 10 minutes between fall and spring).

We used a propensity score blocking technique advocated by Imbens (2015).[1] We first controlled for a cubic function of prior math and ELA test scores, grade-fixed effects, and site-fixed effects. We then added covariates (gender, race, free- and reduced-price lunch status, English language learner status, special education status, and peer aggregate prior test scores and demographics) in a stepwise fashion. We determined the best fitting model using a likelihood ratio test. We then trimmed outliers from the sample and re-estimated the propensity score within the trimmed sample. The final model included race, free- and reduced-price lunch status, and peer aggregate prior test scores and demographics.

Next, we exact-matched students based on grade level and agency type: students in traditional school districts were matched only with other students from the same district; students in CMOs were matched with other students from any CMO. We then created propensity score blocks of students. The size of the blocks was determined by the method proposed by Imbens (2015) and were no smaller than two plus the number of covariates used in the propensity score estimation.

C. STATISTICAL ESTIMATION STRATEGY

Using the trimmed sample, we estimated the residuals for all observations using the following equation:

$$y_{i,t} = \alpha + \beta * Y_{i,pre} + \gamma * S_i + \delta * P_i + \ \eta_t + d_d \ \varepsilon \qquad (1)$$

Where $y_{i,t}$ is the test score outcome for student i during time t. $Y_{i,pre}$ is a cubic function of prior math and ELA test scores. S_i is a set of demographic characteristics including gender, race, free- or reduced-price lunch status, English language learner status, and special education status. P_i is a set of peer averages of each of the demographic characteristics and prior test scores. η_t is a school-year-fixed effect. d_d represents agency-fixed effects.

Within each of the propensity score blocks, we estimated a linear regression of the residuals on the treatment indicator and a set of covariates. The overall treatment effect was calculated as the weighted average of the within-block estimates, where the weight for a given block was the proportion of total students within that block.

For impact analyses broken down by student characteristics (e.g. prior achievement), we subset the sample into the categories of interest (e.g. baseline minutes of usage), blocked the propensity scores within those categories, and estimated weighted treatment effects as described above.

In addition to the balance checks in Section D, we assessed whether there were any systematic differences between the groups by also examining students' end-of-year ELA scores, which would not be expected to change in response to using ST Math. This served as a placebo test, similar to those discussed in a medical context: if students who increased their usage of ST Math experienced greater math *and* ELA gains than other comparable students, there was likely

[1] Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373-419.

some other explanation for the difference in their academic outcomes than the effect of the software. The results indicated no effect of increasing ST Math usage on ELA test scores.

D. CHARACTERISTICS OF SAMPLE AND BALANCE CHECKS

Table A1 shows the descriptive statistics and number of students in our matched-comparison state assessment samples. No variable was significantly unbalanced.

Table A1. Balance Table

	2015-16			2016-17		
Baseline Characteristics	**Treatment Group**	**Comparison Group**	**Adjusted Difference (s.e.)**	**Treatment Group**	**Comparison Group**	**Adjusted Difference (s.e.)**
Male	0.502	0.495	0.009 (0.056)	0.510	0.497	0.012 (0.063)
African-American	0.107	0.228	0.005 (0.039)	0.276	0.293	-0.004 (0.045)
Hispanic	0.829	0.674	0.006 (0.042)	0.627	0.583	0.002 (0.049)
Free or Reduced Price Lunch	0.822	0.779	0.013 (0.031)	0.291	0.335	0.011 (0.032)
English Language Learner	0.674	0.476	-0.002 (0.041)	0.486	0.416	0.003 (0.05)
Special Education	0.131	0.129	-0.001 (0.038)	0.102	0.094	0.005 (0.037)
Average weekly usage of ST Math in fall semester (minutes)	45.528	42.083	-3.048 (2.535)	44.227	37.792	1.116 (2.832)
Average prior state math test score (standardized)	-0.266	-0.167	-0.049 (0.101)	-0.290	-0.191	-0.012 (0.118)
Average prior state ELA test score (standardized)	-0.395	-0.258	-0.031 (0.099)	-0.300	-0.208	-0.002 (0.113)
Number of students	5,049	4,354		4,294	4,626	
Notes: Each cell represents the result of a separate regression of the treatment indicator and a set of agency fixed effects, with standard errors clustered at the school-grade level. *p<0.10, **p<0.05, ***p<0.01.						