# A Cross-State Evaluation of MIND Research Institute's ST Math Program and Math Performance

**Submitted to:**

Andrew Coulson
MIND Research Institute
111 Academy, Suite 100
Irvine, CA 92617

**Submitted by:**

Staci Wendt
John Rice
Jonathan Nakamoto

WestEd
4665 Lampson Avenue
Los Alamitos, CA 90720

**March 2018**

WestEd

# Contents

# Executive Summary

The MIND Research Institute (MIND) contracted with WestEd to conduct an independent assessment of mathematics outcomes in elementary schools across multiple states that were provided with the Spatial-Temporal Math (ST Math) program. The outcomes examined included: (1) mathematics scale scores on standardized state assessments; and (2) the percentage of students scoring at or above proficient in mathematics on those assessments. These outcomes were examined after at least one year of ST Math program implementation at the treatment schools. Data from 474 treatment schools in 16 states[1] that included grade levels 3 through 5 were in the evaluation. Of these 474 schools, 392 provided data for grade level 3, 366 provided data for grade level 4, and 374 provided data for grade level 5 (Exhibit ES1).

Outcomes were examined for all schools that were provided with ST Math as well as for the subset of schools where at least one grade level implemented the program with fidelity. For the purposes of this evaluation, the unit of analysis for the evaluation was a "grade-level cluster" within each school, as opposed to a classroom or the whole school. A grade-level cluster included all the classes in a school that taught content for a specific grade level. For example, the data from an elementary school with four grade-level 4 classes were included in the evaluation as a single "grade-level" cluster. Also, implementation with fidelity was considered to have occurred when at least 85 percent of the students who were enrolled in a grade level at a particular school had been logged into ST Math during the academic year and when an average of at least 50 percent of the grade-level material in ST Math had been covered by students in that grade level. Of the 474 schools that were provided with ST Math, 239 schools in 14 states[2] had at least one grade level that implemented the program with fidelity. Of these 239 schools, 168 implemented with fidelity in grade level 3, 156 implemented with fidelity in grade level 4, and 173 implemented with fidelity in grade level 5 (Exhibit ES1).

**Exhibit ES1. Number of Schools Provided with ST Math at Each Grade Level and that Implemented ST Math with Fidelity at Each Grade Level**

| Grade level | Number of schools that were provided with ST Math at a particular grade level | Number of schools that implemented ST Math with fidelity at a particular grade level | Percentage of schools that implemented ST Math with fidelity at a particular grade level |
|---|---|---|---|
| 3 | 392 | 168 | 43.0 |
| 4 | 366 | 156 | 42.6 |
| 5 | 374 | 173 | 46.3 |

[1] The schools were in the following states: California, Colorado, Connecticut, Florida, Georgia, Iowa, Massachusetts, Michigan, Minnesota, Missouri, Nevada, New Jersey, New York, Texas, Virginia, and Wisconsin.

[2] The schools were in the following states: California, Colorado, Connecticut, Florida, Georgia, Iowa, Michigan, Missouri, Nevada, New Jersey, New York, Texas, Virginia, and Wisconsin.

The evaluation utilized a quasi-experimental design that compared outcomes for: (1) grade-level clusters provided with ST Math matched to grade-level clusters that were not provided with ST Math; and (2) grade-level clusters that implemented ST Math with fidelity matched to grade-level clusters that were not provided with ST Math. For both the entire sample and the subsample that implemented with fidelity, the magnitude of the differences between baseline characteristics for treatment and comparison groups was less than a fifth of a standard deviation. Differences in mathematics outcomes between the treatment and comparison groups were examined across all grade levels, using hierarchical linear modeling, while accounting for the nesting of grade levels within schools. In addition, mathematics outcomes were compared for each of the three grade levels separately, using multiple linear regression to account for differences in several school characteristics as well as mathematics performance prior to the provision of ST Math.

## RESULTS FOR ALL SCHOOLS PROVIDED WITH ST MATH

An analysis of schools that were provided with ST Math revealed statistically significant differences between the treatment and comparison groups for mathematics outcomes, after adjusting for several school-level characteristics, math performance from the year before ST Math was provided, and the nesting of grade levels within schools. The strength of the effect was equivalent to 0.17 of a standard deviation for the percentage of students scoring at or above proficient on the state standardized mathematics assessment, and 0.13 of a standard deviation for scale scores on the state standardized mathematics assessment. These findings remained significant after applying a correction for multiple comparisons.

Statistically significant differences were also found when conducting separate analyses for grade levels 3, 4, and 5. Specifically, for schools that were provided with ST Math, there were statistically significant differences for the percentage of students at each grade level scoring "proficient or above" on state standardized math assessments compared to the percentage of students scoring "proficient or above" from matched grade levels in other schools. Similarly, for schools that were provided with ST Math, there were statistically significant differences in student scale scores at each grade level on state standardized math assessments compared to the scale scores of students in matched grade levels in other schools. These differences occurred after adjusting for several school-level characteristics as well as grade-level math performance from the year before ST Math was provided. These findings remained significant after applying a correction for multiple comparisons.

## RESULTS FOR SCHOOLS WITH AT LEAST ONE GRADE LEVEL THAT IMPLEMENTED ST MATH WITH FIDELITY

An analysis of only schools where the ST Math program was implemented with fidelity in at least one grade level revealed statistically significant differences between the treatment and comparison groups for mathematics outcomes, after adjusting for several school-level characteristics, math performance from the year before ST Math was provided, and the nesting of grade levels within

schools. The strength of the effect was equivalent to 0.36 of a standard deviation for the percentage of students scoring at or above proficient on the state standardized mathematics assessment, and 0.31 of a standard deviation for scale scores on the state standardized mathematics assessment.

Statistically significant differences were also found when conducting separate analyses for grade levels 3, 4, and 5. Specifically, for schools where ST Math was implemented with fidelity in at least one grade level, there were statistically significant differences for the percentage of students at each grade level scoring "proficient or above" on state standardized math assessments compared to the percentage of students scoring "proficient or above" from matched grade levels in comparison schools. Similarly, for schools where ST Math was implemented with fidelity in at least one grade level, there were statistically significant differences at each grade level on the student scale scores on state standardized math assessments compared to the scale scores of students in matched grade levels in comparison schools. These differences occurred after adjusting for several school-level characteristics as well as grade-level math performance from the year before ST Math was provided. These findings remained significant after applying a correction for multiple comparisons.

## LIMITATIONS

A limitation of the current study is that schools or grade-level clusters that were provided with ST Math might have been more focused on improving students' math skills or had more math-related supports at their disposal, compared to schools or grade-level clusters that were not provided with ST Math. Likewise, districts (or in some cases schools) elected to participate in the ST Math program; therefore, there might have been an underlying factor contributing to improvements in math scores, such as an emphasis on improving math, or a focus on integrating technology into the classroom. Further, schools that were included in the analysis of "implemented with fidelity" might differ from the full sample of schools in the study with regard to characteristics in areas that were not measured in our study, such as teacher quality, support from principals, or other schoolwide math reforms that occurred concurrently with ST Math.

# A Cross-State Evaluation of MIND Research Institute's ST Math Program and Math Performance

## Background

The stability of the U.S. economy and the productivity of its workforce depend on having a K–12 education system that produces students who possess strong mathematics skills (Hanushek, 2012). As students with more advanced skills move into the workforce, there are increases in productivity and earnings. Unfortunately, the low mathematics achievement of many students in the U.S. poses a threat to their future academic and employment prospects as well as the future competitiveness of the U.S. economy.

The rate of improvement between 1990 and 2013 in the National Assessment of Educational Progress (NAEP) mathematics scores for grade 4 students has slowed in recent years. From 1990 to 2013, the average scale score increased by 29 points. However, the average scale score of 240 in 2015 represented a 2-point decrease since 2013. In addition, only 40 percent of grade 4 students in 2015 scored at or above proficient on the NAEP mathematics assessment (National Center for Education Statistics, 2015).

The Trends in International Mathematics and Science Study (TIMSS) allows for the comparison of the mathematics performance of U.S. students in grade 4 with their peers from countries across Africa, Asia, Europe, and Latin America. The TIMSS data from 2015 show that U.S. grade 4 students scored higher than the overall average on the mathematics assessment. However, grade 4 students in seven educational systems, including those of the Russian federation, Hong Kong, Japan, and Singapore, outperformed U.S. students on the mathematics assessment by margins that reached statistical significance (Provasnik et al., 2016). Overall, the results from TIMSS and NAEP indicate that effective mathematics interventions are needed in U.S. schools. Enhancing program quality and engaging students in learning are two methods to increase achievement.

## OVERVIEW OF THE ST MATH PROGRAM AND THE EVALUATION

Spatial-Temporal Math (ST Math) is game-based, instructional software for K–12 students, created by the MIND Research Institute (MIND). The purpose of the program is to boost math comprehension through visual learning. ST Math is integrated into classroom instruction but can also be used in a computer lab or at home. The ST Math software games begin without language or symbol abstractions by posing math problems as purely visual puzzles. Interactive, animated visual manipulatives provide informative feedback on student solutions. Puzzle scores of 100 percent are

required for progression through the levels. The ST Math software games follow JiJi, a cartoon penguin, who passes obstacles when students solve spatial math puzzles.

MIND contracted with WestEd to conduct an independent evaluation of ST Math as implemented in schools in several states. The evaluation compared the math performance of students in grade-level clusters that were provided with ST Math against the performance of students in grade-level clusters that were not provided with ST Math but were otherwise in comparable grade levels and schools in the same states. The outcomes examined were average student scale scores from the state standardized math assessments and the percentage of students who were proficient or above (based on their scale scores from the same assessments). Analyses estimated program effects for grade levels within schools that were provided with ST Math; additional analyses estimated program effects on schools only where at least one grade level implemented the program with fidelity.

## METHOD

For the purposes of this evaluation, the unit of analysis for the evaluation was a "grade-level cluster" within each school, as opposed to a classroom or whole school. A grade-level cluster included all the classes in a school that taught content for a specific grade level. For example, the data from an elementary school with four grade 4 classes were included in the evaluation as a single "grade-level cluster." The current study utilized a matched-comparison, quasi-experimental design that matched grade-level clusters that were provided with the ST Math program to grade-level clusters that were not provided with ST Math.[3] WestEd examined grade-average standardized math assessment scale scores and the percentage of students in each grade-level cluster who were proficient or above on the state standardized math assessment. Because outcomes of interest were examined across multiple states and these states have differing standardized tests, z-scores were created for each outcome of interest. To examine the effect of ST Math, analyses were conducted using multiple linear regression (Stata command "regress"), which allowed for inclusion of covariates (in this case, school-level demographic factors and baseline achievement scores) in the statistical model. Program effects were estimated for all schools that were provided with ST Math and separately for schools where the program was implemented with fidelity in at least one grade level. Outcomes were computed using hierarchical linear modeling to account for the nesting of grades within schools.

[3] Matching is a quasi-experimental alternative to a randomized controlled trial. When conducted with large samples, randomization makes the treatment and control groups equal on all characteristics other than the treatment condition, allowing for any differences between groups seen after the treatment or program to be causally determined as a result of exposure to the treatment or program. Without randomization, the possibility that two groups differ on other characteristics besides exposure to the treatment or program is a threat to causal conclusions (Shadish, Cook, & Campbell, 2002).

## DATA PROVIDED BY MIND

MIND provided all applicable data to WestEd. MIND combined three sources of data prior to providing a single dataset to WestEd: ST Math implementation data collected by MIND, school-level demographic data from Market Data Retrieval (MDR), and assessment score data that MIND staff culled from each state's publicly accessible school accountability websites and research tables.

In preparing the data for WestEd, MIND examined its own dataset for all active users of ST Math in grade levels 3 through 5 during the 2015/16 school year. MIND then narrowed the treatment pool to schools that had grade levels with either one, two, or three years of consecutive program use by the end of the 2015/16 school year. Schools were eliminated in states with missing data in either the "baseline" year (i.e., the year before the first year of ST Math implementation) or the 2015/16 school year. In addition, schools in states with unavailable or incomplete state standardized test scale scores or proficiency rates were removed from the sample. Finally, schools for which MDR did not provide baseline year demographic data were removed from the treatment group dataset. The comparison pool was compiled in a similar way, using only schools that never used the ST Math program in any year, based on MIND's records.

In order to combine the three datasets (i.e., ST Math implementation data from MIND, school demographic data from MDR, and state assessment data from state websites) and to assign unique identifiers to all schools (and grade-level clusters within schools), MIND used the school names provided in the MDR data to match it to both the MIND data and state data. When data between three data sources could not be matched using this method, MIND researchers manually checked the ST Math schools in order to match data from all sources. However, this latter procedure was not used for schools in the comparison pool when a match was not initially successful using the school name. Thus, potential comparison schools may have been dropped. After the data were selected and combined using the aforementioned criteria, MIND provided a single dataset to WestEd prior to any matching or data analysis.

## SELECTION OF TREATMENT AND COMPARISON GROUPS BY WESTED

WestEd conducted two sets of analyses. This first set of analyses estimated effects for all schools that were provided with the ST Math program, regardless of the extent to which the program was implemented. The second set of analyses included only a subset of these schools where the ST Math was implemented with fidelity in at least one grade level. The following paragraphs discuss the selection of the analytic samples for the treatment and comparison groups for both sets of analyses.

### IDENTIFICATION OF THE SAMPLE OF SCHOOLS PROVIDED WITH ST MATH

The treatment group consisted of grade-level clusters within schools where ST Math was provided. Each cluster consisted of all the classes within a school that were from the same grade level, either grade level 3, 4, or 5. The clusters in the treatment group were from schools where ST Math was provided for either one year (beginning in the 2015/16 school year), two years (beginning in the

2014/15 school year), or three years (beginning in the 2013/14 school year). All follow-up data were collected from the 2015/16 school year. The treatment pool began with all 1,794 grade-level clusters that had received a maximum of three years of ST Math. Exhibit 1 shows the number of grade-level clusters that received one, two, or three years of ST Math.

Schools were eliminated from the treatment pool if applicable grade-level data were missing (this includes data from MDR and from state data sources) for the year prior to ST Math implementation (i.e., the pre-intervention school year, or "baseline") or for the most recent year of ST Math implementation (i.e., the intervention school year(s), or "follow-up"). In addition, schools were eliminated if demographic variables from the baseline school year were missing; these variables were to be used in the matching process. After these exclusions, the treatment group included 474 schools. Of these, 392 schools had ST Math data for grade level 3, 366 schools had ST Math data for grade level 4, and 374 schools had ST Math data for grade level 5 (Exhibit 2). The schools were in the following states: California, Colorado, Connecticut, Florida, Georgia, Iowa, Massachusetts, Michigan, Minnesota, Missouri, Nevada, New Jersey, New York, Texas, Virginia, and Wisconsin.

## IDENTIFICATION OF SCHOOLS WHERE ST MATH WAS IMPLEMENTED WITH FIDELITY IN ONE OR MORE GRADE LEVELS

The second set of analyses included only schools where ST Math was implemented with fidelity in at least one grade level. For these analyses, implementation with fidelity was considered to have occurred for a grade level when at least 85 percent of the students in that grade level at a school had been enrolled in ST Math during the academic year and an average of at least 50 percent of the grade-level material in ST Math was covered by those students.[4] These cutoff criteria are considered the minimum to obtain a status of implementing with fidelity. Ideally, schools would have a higher percentage of student enrollment and completion. Of the 474 schools that were provided with ST Math, 239 schools in 14 states[5] had at least one grade level that implemented the program with fidelity. Of the 239 schools, 168 implemented with fidelity in grade level 3, 156 implemented with fidelity in grade level 4, and 173 implemented with fidelity in grade level 5 (Exhibit 2).

**Exhibit 1. Number of Years of ST Math Implementation**

| Number of grade-level clusters | 1 year of ST Math implementation | 2 years of ST Math implementation | 3 years of ST Math implementation |
|---|---|---|---|
| Provided with ST Math | 258 | 163 | 711 |
| Implemented with fidelity | 91 | 79 | 327 |

[4] To calculate enrollment percentage, the denominator was the number of students who took the mathematics test in the implementation year, and the numerator was the number of students who were enrolled in ST Math, which was obtained from MIND. The average percentage of grade-level ST Math material covered was also obtained from MIND.

[5] The schools were in the following states: California, Colorado, Connecticut, Florida, Georgia, Iowa, Michigan, Missouri, Nevada, New Jersey, New York, Texas, Virginia, and Wisconsin.

**Exhibit 2. Number of Schools Provided with ST Math at Each Grade Level and that Implemented ST Math with Fidelity at Each Grade Level**

| Grade level | Number of schools provided with ST Math at a particular grade level | Number of schools that implemented ST Math with fidelity at a particular grade level | Percentage of schools that implemented ST Math with fidelity at a particular grade level |
|---|---|---|---|
| 3 | 392 | 168 | 43.0 |
| 4 | 366 | 156 | 42.6 |
| 5 | 374 | 173 | 46.3 |

Note: For the purposes of this study, implementation with fidelity was considered to have occurred for a grade level within a school when at least 85 percent of the students in that grade level at the school had been enrolled in ST Math during the academic year and at least 50 percent of the grade-level material in ST Math was covered by those students.

## IDENTIFICATION OF COMPARISON GROUP

The comparison group was from clusters (i.e., one or more classrooms) in grade levels 3 through 5 that had not been provided with the ST Math program prior to, or during, the baseline school year. There were 165,583 such grade-level clusters in the pool of potential comparison clusters. Next, grade-level clusters were excluded if they were missing applicable data for the baseline or follow-up years. Excluding clusters that were missing applicable grade-level data reduced the comparison pool to 108,332 (or 65.4 percent) of the potential comparison grade-level clusters. Appendix C outlines the selection of treatment and comparison groups.

## MATCHING

WestEd used a matching procedure to identify a comparison group. The purpose of matching is to create two groups that are essentially equal on the observable variables known to be related to the outcome of interest.[6] Several different types of matching strategies exist (Guo & Fraser, 2010), and propensity score matching is one such technique. Using the Stata command "psmatch2," WestEd identified a group of comparison schools to match each of the treatment schools. Matching for each grade-level cluster was done within the same state as the treatment schools. Grade-level clusters were matched one-to-many, and frequency weights were used in analyses to establish equal group sizes. To examine the reliability of the matching technique, treatment and comparison clusters were

[6] Matching is a quasi-experimental alternative to a randomized controlled trial. When conducted with large samples, randomization makes the treatment and control groups equal on all characteristics other than the treatment condition, allowing for any differences between groups seen after the treatment or program to be causally determined as a result of exposure to the treatment or program. Without randomization, the possibility that two groups differ on other characteristics besides exposure to the treatment or program is a threat to causal conclusions (Shadish, Cook, & Campbell, 2002).

compared on the matching variables. After identifying the treatment group for the analysis of implementation with fidelity, matching was redone, again restricted to within the same state as the treatment clusters in order to identify a set of comparison grade-level clusters that were most similar to the implementation-with-fidelity grade-level clusters. The comparison and treatment groups did not significantly differ on matching characteristics. Exhibits A1–A6 in Appendix A show the results of *t*-test comparisons for each grade-level cluster for both the entire sample and the fidelity sample. For both the entire sample and the fidelity sample, the magnitude (i.e., effect size) of the differences between treatment and comparison clusters was less than a quarter of a standard deviation.

## ANALYSES

Regression models were used to examine the effects of ST Math for treatment and comparison groups on two outcomes: mathematics scale scores, and the proportion of students who were proficient or above based on the scale scores.[7] Regression is an appropriate analysis technique because it models covariates, or variables known to be related to the outcome. The inclusion of covariates provides more precise estimates of the effect of ST Math program participation on outcomes than if the covariates were not included. The regression models included the following as covariates: baseline grade-level cluster percentage at or above proficient in math or average math scale score (depending on outcome); school-level percentages of White, Asian, Latino, Native American/American Indian, and African American students; the number of students with high socioeconomic need in the school; and the number of students enrolled in the school. To calculate the overall treatment effect across grade-level clusters, WestEd used hierarchical linear modeling (HLM) (Raudenbush & Bryk, 2002) to account for the nesting of grade-level clusters within schools.[8] In addition, as the number of outcome comparisons increased, the likelihood of Type-I errors also increased. To address this issue, WestEd used the Benjamini-Hochberg (BH) correction for comparisons of each group of two or more outcomes within a grade level (Benjamini & Hochberg, 1995). This was conducted for analyses by grade level and across grades. The results of each analysis are presented with and without the BH correction.

[7] For the regression analysis: Outcome = $\alpha$ + $\beta_1$(Treatment) + $\beta_2$(Demographic Covariate 1) +… +$\beta_3$(Demographic Covariate 8) + $\varepsilon_{ij}$

[8] For the HLM analyses: Level 1: Outcome$_{ij}$ = $\beta_{0j}$ + $\beta_{1j}$(Baseline Percent at or Above Proficient)$_{ij}$ + $\beta_{2j}$(Treatment)$_{ij}$ + $\beta_{3j}$(Demographic Covariate 1)$_{ij}$ +… + $\beta_{11j}$(Demographic Covariate 8)$_{ij}$ + $\beta_{12j}$(Grade dummy code)$_{ij}$ + $\beta_{13j}$(Grade dummy code)$_{ij}$ + $r_{ij}$. Level 2: $\beta_{0j}$ = $\gamma_{00}$ + $\gamma_{01}$(Treatment Status)$_j$ + $u_{0j}$

## Results for All Schools Provided with ST Math

The analyses of schools that were provided with ST Math revealed statistically significant differences between these schools and comparison schools for all grade levels with regard to the percentage of students at or above proficient in math, and for math scale scores on the state standardized exam (Exhibit 3). These differences were found after adjusting for several school-level characteristics and math proficiency rates (or scale scores) from the year before ST Math was provided. Specifically, students in clusters of grade level 3 that were provided with the ST Math program had z-score proficiency rates in math that were, on average, 0.13 of a standard deviation higher compared to students in clusters of grade level 3 that were not provided with the ST Math program. Likewise, students in clusters of grade level 4 that were provided with the ST Math program had z-score proficiency rates in math that were, on average, nearly a quarter of a standard deviation higher compared to students in clusters of grade level 4 that were not provided with the ST Math program. Students in clusters of grade level 5 that were provided with the ST Math program had z-score proficiency rates that were 0.15 of a standard deviation higher compared to students in clusters of grade level 5 that were not provided with the ST Math program. All three findings remained statistically significant after applying the correction for multiple comparisons. When comparing the two groups' average student math scale scores on the state standardized assessment, the findings were statistically significant for grade levels 4 and 5 only, and the magnitudes (or effect sizes) were smaller compared to the proficiency outcomes. In addition, the findings for grade level 4, but not grade level 5, remained statistically significant after applying the correction for multiple comparisons. The unadjusted baseline and follow-up means and standard deviations are included in Appendix B.

**Exhibit 3. Differences in Mathematics Performance for Schools Provided with ST Math, by Grade Level**

| Grade level 3 | | | | | | |
|---|---|---|---|---|---|---|
| Outcome | Adjusted mean Treatment (N = 392) | Adjusted mean Comparison (N = 392) | Adjusted mean difference | *t*-test | Effect size | *p*-value |
| Scale score | 0.40 | 0.29 | 0.11 | 2.89 | 0.12 | .01*† |
| % proficient | 0.49 | 0.36 | 0.13 | 2.91 | 0.13 | .01*† |
| **Grade level 4** | | | | | | |
| Outcome | Adjusted mean Treatment (N = 366) | Adjusted mean Comparison (N = 366) | Adjusted mean difference | *t*-test | Effect size | *p*-value |
| Scale score | 0.56 | 0.41 | 0.15 | 3.90 | 0.17 | .01*† |
| % proficient | 0.69 | 0.47 | 0.22 | 5.37 | 0.23 | .01*† |
| **Grade level 5** | | | | | | |
| Outcome | Adjusted mean Treatment (N = 374) | Adjusted mean Comparison (N = 374) | Adjusted mean difference | *t*-test | Effect size | *p*-value |
| Scale score | 0.47 | 0.38 | 0.09 | 2.22 | 0.10 | .03* |
| % proficient | 0.54 | 0.38 | 0.16 | 3.58 | 0.15 | .01*† |

* Statistically significant at *p*-value < .05, two-tailed test.
† Statistically significant at < BH critical value correcting for the false discovery rate under multiple testing within each grade.
Note: All outcomes adjusted for baseline grade-level percentages of students at or above proficient in math (or the baseline average math scale scores); and for baseline school-level percentages of White, Asian, Latino, Native American/American Indian, and African American students, percentage of students with high socioeconomic need, and number of students enrolled.
Treatment N = 1,132 grade-level clusters; comparison N =1,132 grade-level clusters; across 16 states.

The analyses that combined grade levels 3 through 5 revealed statistically significant differences between the treatment and comparison groups for proficiency rates. This was the case after adjusting for the z-score baseline grade-level clusters' percentages of students at or above proficient in math (or the z-score baseline average math scale scores); and for the baseline school-level percentages of White, Asian, Latino, Native American/American Indian, and African American students, percentage of students with high socioeconomic need in each school, and number of students enrolled in each school. The analysis also accounted for the nesting of grade levels within schools, and for school characteristics (Exhibit 4). Specifically, grade-level clusters that were provided with the ST Math program had higher proportions of students who were at or above proficient in math compared to students in grade-level clusters that were not provided with the ST Math program. The pattern was similar for scale scores. The magnitude (or effect size) of the difference was 0.17 of a standard deviation for proficiency rates and 0.13 for scale scores.

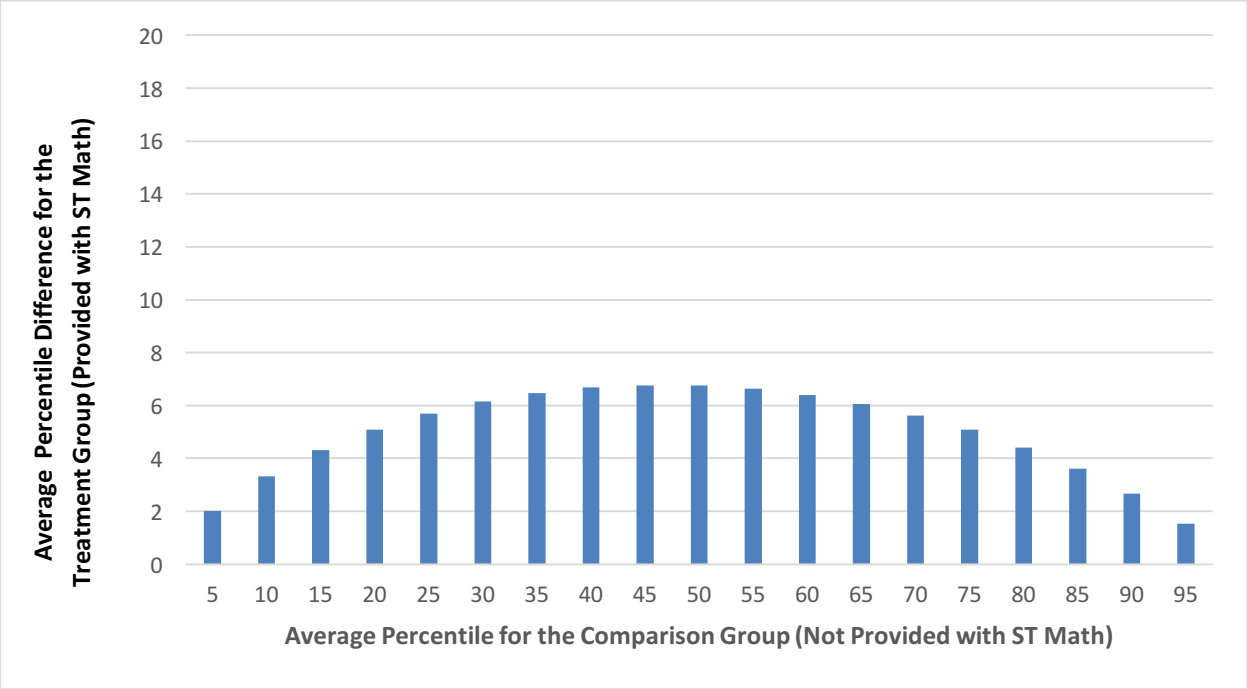
**Exhibit 4. Differences in Mathematics Performance for Schools Provided with ST Math, Across Grade Levels**

| Outcome | Adjusted mean: Treatment schools | Adjusted mean: Comparison schools | Adjusted mean difference | *z*-test | Effect size | *p*-value |
|---|---|---|---|---|---|---|
| Scale score | 0.51 | 0.39 | 0.12 | 4.65 | 0.13 | .001*† |
| % proficient | 0.62 | 0.45 | 0.17 | 5.88 | 0.17 | .01*† |

* Statistically significant at *p*-value < .05, two-tailed test.
†Statistically significant at < BH critical value correcting for the false discovery rate under multiple testing.
Note: All outcomes adjusted for baseline grade-level percentages of students at or above proficient in math; and for baseline school-level percentages of White, Asian, Latino, Native American/American Indian, and African American students, the number of students with high socioeconomic need, and the number of students enrolled. The outcomes account for the nesting of grades within schools.
Treatment N = 1,132 grade-level clusters; comparison N =1,132 grade-level clusters; across 16 states.

The effect size of the z-scored percentage of students who are proficient is 0.17, which can be converted to the difference between the treatment and comparison groups by using mean percentiles. However, the difference in the mean percentiles is dependent on where the scores fall in the distribution, with larger percentile-point differences occurring in the middle. The difference in percentile points that corresponds to an effect size of 0.17 along the normal distribution can be found in Exhibit 5. For example, if the comparison group's average percentage of students who are proficient is at the 10th percentile, an effect size of 0.17 would mean that the average percentage of students who are proficient in the treatment group would be at the 13th percentile — a difference of 3 percentile points. If the average comparison group percentage of students who are proficient is at the 50th percentile, an effect size of 0.17 would mean that the average percentage of students who are proficient in the treatment group is at the 57th percentile, for a difference of 7 percentile points.

**Exhibit 5. Percentile Differences Between the Group Provided with ST Math and the Comparison Group When the Effect Size = 0.17**




## Results for Schools with at Least One Grade Level that Implemented ST Math with Fidelity

The analyses of schools with at least one grade level that implemented ST Math with fidelity found statistically significant differences for all grade levels in these schools in terms of the percentage of students at or above proficient in math, and for math scale scores on the state standardized exam (Exhibit 6). These differences were found after adjusting for several school-level characteristics and math proficiency rates (or scale scores) from the year before ST Math was provided. Specifically, students in grade-level clusters 3 through 5 in schools that implemented the ST Math program with fidelity had z-score proficiency rates in math that were, on average, from a fifth to nearly a half of a standard deviation higher than those of students in grade-level clusters that were not provided with the ST Math program. These findings remained statistically significant after applying the correction for multiple comparisons. When examining average student math scale scores on the state standardized assessment between the two groups, the findings were statistically significant for all grade levels, and the magnitudes (or effect sizes) ranged from 0.27 to 0.43 of a standard deviation. The findings remained statistically significant after applying the correction for multiple comparisons. The unadjusted baseline and follow-up means and standard deviations are included in Appendix B.

**Exhibit 6. Differences in CST Mathematics Performance When ST Math Was Implemented with Fidelity, by Grade Level**

| Grade level 3 | | | | | | |
|---|---|---|---|---|---|---|
| Outcome | Adjusted mean Treatment (N = 168) | Adjusted mean Comparison (N = 168) | Adjusted mean difference | *t*-test | Effect size | *p*-value |
| Scale score | 0.40 | 0.08 | 0.32 | 5.00 | 0.36 | .01*† |
| % proficient | 0.44 | 0.09 | 0.35 | 4.95 | 0.38 | .01*† |
| **Grade level 4** | | | | | | |
| Outcome | Adjusted mean Treatment (N = 156) | Adjusted mean Comparison (N = 156) | Adjusted mean difference | *t*-test | Effect size | *p*-value |
| Scale score | 0.77 | 0.46 | 0.31 | 5.26 | 0.36 | .01*† |
| % proficient | 1.02 | 0.64 | 0.38 | 6.13 | 0.43 | .01*† |
| **Grade level 5** | | | | | | |
| Outcome | Adjusted mean Treatment (N = 173) | Adjusted mean Comparison (N = 173) | Adjusted mean difference | *t*-test | Effect size | *p* value |
| Scale score | 0.57 | 0.36 | 0.21 | 3.44 | 0.22 | .01*† |
| % proficient | 0.63 | 0.37 | 0.26 | 3.87 | 0.27 | .01*† |

* Statistically significant at *p*-value < .05, two-tailed test.
† Statistically significant at < BH critical value correcting for the false discovery rate under multiple testing.
Note: All outcomes adjusted for baseline grade-level percentages of students at or above proficient in math (or the baseline average math scale scores); and for baseline school-level percentages of White, Asian, Latino, Native American/American Indian, and African American students, the number of students with high socioeconomic need in the school, and the number of students enrolled in the school.
Treatment N = 497 grade-level clusters; comparison N = 497 grade-level clusters; across 14 states.

For math proficiency rates as well as math scale scores, the analyses that combined grades 3 through 5 revealed statistically significant differences between schools that implemented ST Math with fidelity in at least one grade-level cluster and comparison grade-level clusters that were not provided with ST Math (Exhibit 7). This was the case after adjusting for the z-score baseline grade-level percentages of students at or above proficient in math (or the z-score baseline average math scale scores); and for baseline school-level percentages of White, Asian, Latino, Native American/American Indian, and African American students, the number of students with high socioeconomic need at the school, and the number of students enrolled at the school. The analysis also accounted for the nesting of grade levels within schools and for school characteristics. Specifically, grade-level clusters that implemented ST Math with fidelity had higher math proficiency rates compared to those that were not provided with the ST Math program. The same was true for math scale scores. The magnitude (or effect size) of the difference for math proficiency rates was 0.36 of a standard deviation and was 0.31 for math scale scores.

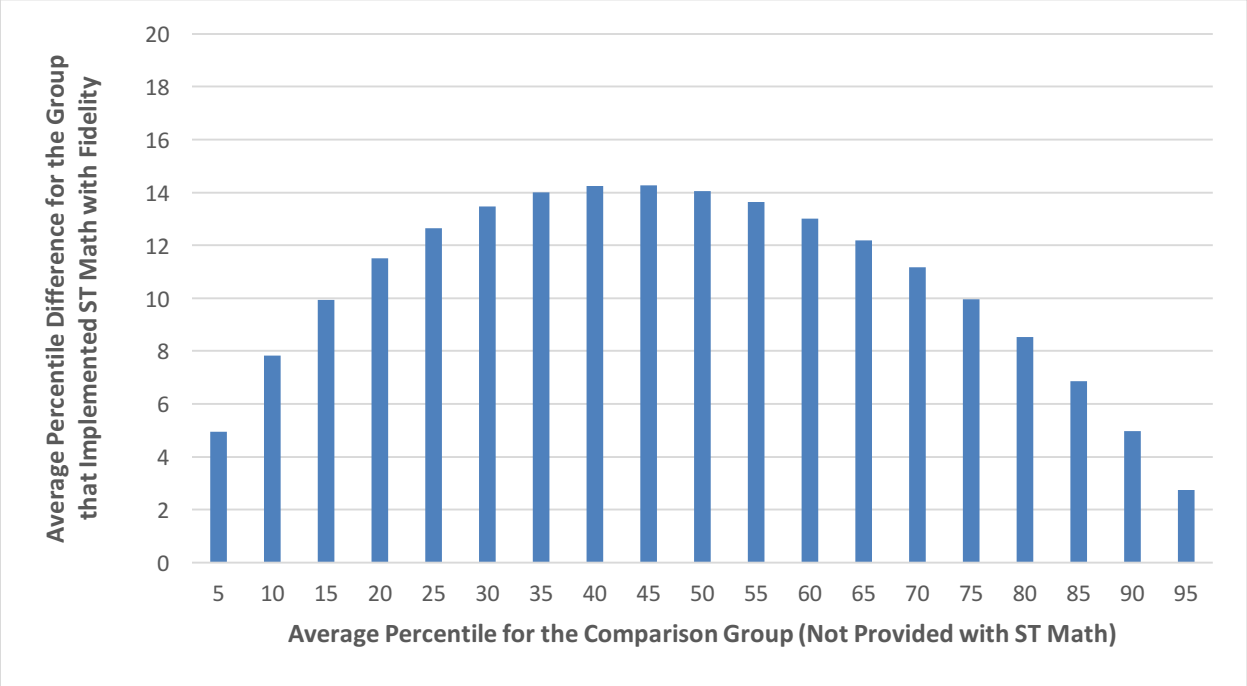
**Exhibit 7. Differences in CST Mathematics Performance for Schools with at Least One Grade Level that Implemented ST Math with Fidelity**

| Outcome | Adjusted mean | | Adjusted mean difference | *t*-test | Effect size | *p*-value |
|---|---|---|---|---|---|---|
| | Treatment | Comparison | | | | |
| Scale score | 0.59 | 0.31 | 0.28 | 7.05 | 0.31 | .01*† |
| % proficient | 0.73 | 0.40 | 0.33 | 7.74 | 0.36 | .01*† |

* Statistically significant at *p*-value < .05, two-tailed test.
† Statistically significant at < BH critical value correcting for the false discovery rate under multiple testing.
Note: All outcomes adjusted for baseline grade-level cluster percentages of students at or above proficient in math; and for baseline school-level percentages of White, Asian, Latino, Native American/American Indian, and African American students, the number of students with high socioeconomic need in the school, and the number of students enrolled in the school. The outcomes account for the nesting of grade-level clusters within schools.
Treatment N = 497 grade-level clusters; comparison N = 497 grade-level clusters; across 14 states.

The effect size of the z-scored percentage of students who are proficient is 0.36, which can be converted to the difference between the treatment and comparison groups by using mean percentiles. However, the difference in the mean percentiles is dependent on where the scores fall in the distribution, with larger percentile-point differences occurring in the middle. The difference in percentile points that corresponds to an effect size of 0.36 along the normal distribution can be found in Exhibit 8. For example, if the average comparison grade's percentage of students who are proficient is at the 10th percentile, an effect size of 0.36 would mean that the average percentage of students who are proficient in grades that were provided with ST Math would be at the 18th percentile — a difference of 8 percentile points. If the average comparison grade's percentage of students who are proficient is at the 50th percentile, an effect size of 0.36 would mean that the average percentage of students who are proficient in grades that were provided with ST Math is at the 64th percentile, for a difference of 14 percentile points.

**Exhibit 8. Percentile Differences Between the Group that Implemented ST Math with Fidelity and the Comparison Group When the Effect Size = 0.36**




## Discussion and Study Limitations

Although the matching procedure and quasi-experimental design provided additional rigor to the study, there are some limitations of the current evaluation. One limitation is that schools that participated in the ST Math program, or that implemented it with fidelity, may have been different in some unknown way(s) in relation to the comparison schools. Districts that had grade levels that implemented ST Math opted in to adopt the program; therefore, there may have been underlying factors contributing to improvements in math scores, such as an emphasis on improving math, or a focus on integrating technology into the classroom, relative to the comparison groups that were not in districts that adopted ST Math. Further, schools that were included in the fidelity analyses implemented ST Math at a minimum level of fidelity, as defined by MIND. These schools might have differed from other schools in ways that were not measured, such as teacher quality, support from principals, or in implementing other schoolwide math reforms that occurred concurrently with ST Math. For example, the unadjusted baseline z-scores are larger for the subgroup that implemented with fidelity compared to the larger group that implemented ST Math with or without fidelity; however, the comparison groups for the full sample and the subsample had similar unadjusted baseline z-scores.

A second limitation is that schools in the treatment group received varying years of ST Math implementation. The number of years a school participated in the ST Math program could have had

an impact on mathematics outcomes. Future research should investigate the impact of receiving multiple years of ST Math.

Future research on the cross-state impact of ST Math could be strengthened in two ways. First, despite the careful matching of treatment and comparison groups on observable characteristics, it is possible that differences existed between the two groups and that these differences contributed (in whole or part) to the positive findings for ST Math. Without randomization, the possibility that the groups differed on other characteristics besides exposure to ST Math impedes causal conclusion (Shadish, Cook, & Campbell, 2002). Second, obtaining individual student-level math outcomes would allow for a more precise estimate of standard errors and would allow researchers to assess any impacts of the program on individual students over time, either due to multiple years of exposure or to long-term effects after exposure ends.

## References


Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, *57,* 289–300.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.

Hanushek, E. A. (2012). Education quality and economic growth. In B. Miniter (Ed.), *The 4% solution: Unleashing the economic growth America needs* (pp. 226–239; 324–326). New York: Crown Business.

National Center for Education Statistics. (2015). *A first look: 2015 mathematics and reading.* Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Provasnik, S., Malley, L., Stephens, M., Landeros, K., Perkins, R., & Tang, J. H. (2016). *Highlights from TIMSS and TIMSS Advanced 2015: Mathematics and science achievement of U.S. students in grades 4 and 8 and in advanced courses at the end of high school in an international context* (NCES 2017-002). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences*. Boston, MA: Houghton Mifflin.

# Appendix A. Baseline Comparisons

### Exhibit A1. Entire Sample – Grade Level 3

| | Comparison | | Treatment | | | | |
|---|---|---|---|---|---|---|---|
| Outcome | M | SD | M | SD | *t* | *p* | *d* |
| Percent proficient (z-score) | -0.06 | 1.04 | -0.06 | 0.95 | -0.06 | 0.95 | 0.00 |
| Scale score (z-score) | -0.04 | 1.01 | -0.04 | 0.95 | 0.04 | 0.97 | 0.00 |
| Student enrollment | 545.60 | 224.70 | 561.78 | 231.53 | 0.99 | 0.32 | 0.07 |
| Percent White | 45.88 | 32.65 | 44.45 | 30.94 | -0.63 | 0.53 | -0.04 |
| Percent African American | 13.92 | 20.21 | 13.16 | 16.15 | -0.58 | 0.56 | -0.04 |
| Percent American Indian | 0.40 | 0.76 | 0.38 | 0.60 | -0.42 | 0.68 | -0.03 |
| Percent Latino | 31.11 | 27.76 | 32.21 | 28.94 | 0.54 | 0.59 | 0.04 |
| Percent Asian | 7.11 | 11.95 | 7.96 | 13.02 | 0.96 | 0.34 | 0.07 |
| Percent high socioeconomic need | 49.73 | 28.31 | 51.44 | 28.17 | 0.85 | 0.40 | 0.06 |

Note: M = mean; SD = standard deviation; *t* = *t*-test statistic; *p* = *p*-value; *d* = effect size.
Treatment N = 392 grade-level clusters; comparison N = 392 grade-level clusters.

### Exhibit A2. Entire Sample – Grade Level 4

| | Comparison | | Treatment | | | | |
|---|---|---|---|---|---|---|---|
| Outcome | M | SD | M | SD | *t* | *p* | *d* |
| Percent proficient (z-score) | -0.07 | 0.96 | -0.07 | 0.91 | -0.32 | 0.75 | 0.00 |
| Scale score (z-score) | -0.10 | 0.88 | -0.10 | 0.92 | 0.38 | 0.71 | 0.00 |
| Student enrollment | 549.15 | 277.01 | 575.28 | 232.86 | 1.38 | 0.17 | 0.10 |
| Percent White | 45.28 | 34.13 | 45.11 | 30.76 | -0.07 | 0.94 | -0.01 |
| Percent African American | 15.75 | 24.17 | 12.93 | 15.64 | -1.87 | 0.06 | -0.14 |
| Percent American Indian | 0.31 | 0.60 | 0.39 | 0.60 | 1.78 | 0.08 | 0.13 |
| Percent Latino | 30.14 | 28.70 | 32.45 | 28.88 | 1.08 | 0.28 | 0.08 |
| Percent Asian | 7.66 | 12.55 | 7.96 | 12.81 | 0.33 | 0.74 | 0.02 |
| Percent high socioeconomic need | 49.53 | 27.98 | 51.57 | 27.83 | 0.99 | 0.32 | 0.07 |

Note: M = mean; SD = standard deviation; *t* = *t*-test statistic; *p* = *p*-value; *d* = effect size.
Treatment N = 366 grade-level clusters; comparison N = 366 grade-level clusters.

## Exhibit A3. Entire Sample – Grade Level 5

| | Comparison | | Treatment | | | | |
|---|---|---|---|---|---|---|---|
| Outcome | M | SD | M | SD | *t* | *p* | *d* |
| Percent proficient (z-score) | -0.12 | 1.06 | -0.03 | 0.99 | 1.24 | 0.21 | 0.09 |
| Scale score (z-score) | -0.10 | 1.04 | 0.00 | 1.01 | 1.33 | 0.18 | 0.10 |
| Student enrollment | 544.28 | 282.46 | 563.87 | 230.16 | 1.04 | 0.30 | 0.08 |
| Percent White | 44.98 | 34.60 | 43.73 | 32.00 | -0.52 | 0.61 | -0.04 |
| Percent African American | 14.14 | 21.68 | 13.88 | 17.65 | -0.18 | 0.86 | -0.01 |
| Percent American Indian | 0.35 | 0.63 | 0.43 | 1.08 | 1.24 | 0.22 | 0.09 |
| Percent Latino | 31.17 | 30.70 | 33.89 | 31.06 | 1.07 | 0.28 | 0.09 |
| Percent Asian | 7.99 | 15.30 | 7.01 | 12.37 | -0.97 | 0.33 | -0.07 |
| Percent high socioeconomic need | 50.88 | 29.61 | 52.81 | 28.73 | 0.90 | 0.37 | 0.07 |

Note: M = mean; SD = standard deviation; *t* = *t*-test statistic; *p* = *p*-value; *d* = effect size.
Treatment N = 374 grade-level clusters; comparison N = 374 grade-level clusters.

## Exhibit A4. Fidelity Sample – Grade Level 3

| | Comparison | | Treatment | | | | |
|---|---|---|---|---|---|---|---|
| Outcome | M | SD | M | SD | *t* | *p* | *d* |
| Percent proficient (z-score) | -0.01 | 1.13 | 0.05 | 0.87 | 0.41 | 0.69 | 0.06 |
| Scale score (z-score) | -0.01 | 1.09 | 0.04 | 0.86 | 0.45 | 0.65 | 0.05 |
| Student enrollment | 504.59 | 222.56 | 542.21 | 228.94 | 1.52 | 0.13 | 0.17 |
| Percent White | 54.73 | 34.03 | 55.95 | 29.25 | 0.35 | 0.73 | 0.04 |
| Percent African American | 15.15 | 22.21 | 13.89 | 16.15 | -0.60 | 0.55 | -0.07 |
| Percent American Indian | 0.37 | 0.79 | 0.41 | 0.64 | 0.50 | 0.62 | 0.06 |
| Percent Latino | 23.37 | 25.50 | 24.65 | 24.71 | 0.47 | 0.64 | 0.05 |
| Percent Asian | 3.95 | 7.11 | 3.83 | 5.44 | -0.17 | 0.86 | -0.02 |
| Percent high socioeconomic need | 48.14 | 27.19 | 47.14 | 24.90 | -0.35 | 0.73 | -0.04 |

Note: M = mean; SD = standard deviation; *t* = *t*-test statistic; *p* = *p*-value; *d* = effect size.
Treatment N = 168 grade-level clusters; comparison N = 168 grade-level clusters.

## Exhibit A5. Fidelity Sample – Grade Level 4

| | Comparison | | Treatment | | | | |
|---|---|---|---|---|---|---|---|
| Outcome | M | SD | M | SD | *t* | *p* | *d* |
| Percent proficient (z-score) | 0.12 | 0.98 | 0.03 | 0.89 | -0.86 | 0.39 | -0.10 |
| Scale score (z-score) | 0.13 | 1.00 | 0.05 | 0.88 | -0.71 | 0.48 | -0.09 |
| Student enrollment | 523.85 | 265.44 | 554.00 | 206.18 | 1.12 | 0.26 | 0.13 |
| Percent White | 58.70 | 27.31 | 59.55 | 26.73 | 0.26 | 0.80 | 0.03 |
| Percent African American | 13.74 | 19.91 | 12.94 | 14.16 | -0.41 | 0.68 | -0.05 |
| Percent American Indian | 0.38 | 0.72 | 0.41 | 0.68 | 0.42 | 0.67 | 0.04 |
| Percent Latino | 21.83 | 24.35 | 21.92 | 20.63 | 0.03 | 0.97 | 0.00 |
| Percent Asian | 4.69 | 9.49 | 4.50 | 5.79 | -0.21 | 0.84 | -0.02 |
| Percent high socioeconomic need | 48.79 | 27.31 | 46.60 | 23.94 | -0.75 | 0.45 | -0.09 |

Note: M = mean; SD = standard deviation; *t* = *t*-test statistic; *p* = *p*-value; *d* = effect size.
Treatment N = 156 grade-level clusters; comparison N = 156 grade-level clusters.

## Exhibit A6. Fidelity Sample – Grade Level 5

| | Comparison | | Treatment | | | | |
|---|---|---|---|---|---|---|---|
| Outcome | M | SD | M | SD | *t* | *p* | *d* |
| Percent proficient (z-score) | 0.09 | 1.02 | 0.07 | 0.90 | -0.17 | 0.86 | -0.02 |
| Scale score (z-score) | 0.10 | 1.04 | 0.10 | 0.91 | -0.04 | 0.97 | 0.00 |
| Student enrollment | 540.20 | 264.92 | 556.10 | 206.56 | 0.62 | 0.53 | 0.07 |
| Percent White | 57.35 | 33.27 | 55.60 | 29.30 | -0.52 | 0.61 | -0.06 |
| Percent African American | 14.37 | 20.66 | 14.21 | 15.69 | -0.08 | 0.94 | -0.01 |
| Percent American Indian | 0.37 | 0.97 | 0.39 | 0.57 | 0.27 | 0.79 | 0.03 |
| Percent Latino | 21.59 | 23.33 | 23.93 | 24.50 | 0.91 | 0.37 | 0.10 |
| Percent Asian | 4.55 | 7.73 | 4.12 | 5.87 | -0.58 | 0.56 | -0.06 |
| Percent high socioeconomic need | 49.23 | 28.65 | 49.27 | 25.64 | 0.01 | 0.99 | 0.00 |

Note: M = mean; SD = standard deviation; *t* = *t*-test statistic; *p* = *p*-value; *d* = effect size.
Treatment N = 173 grade-level clusters; comparison N = 173 grade-level clusters.

# Appendix B. Outcomes at Baseline and Follow-Up

**Exhibit B1. Unadjusted Mathematics Z-Score Test Performance at Baseline and Follow-Up for All Members of the Treatment Group and the Comparison Group**

| | | Baseline | | Follow-up | |
|---|---|---|---|---|---|
| | | Treatment | Comparison | Treatment | Comparison |
| Outcome | Grade level | M (SD) | M (SD) | M (SD) | M (SD) |
| Scale score | 3 | -0.04 (0.95) | -0.04 (1.01) | 0.10 (0.79) | -.01 (0.98) |
| Percent proficient | 3 | -0.06 (0.95) | -0.06 (1.04) | 0.15 (0.97) | 0.03 (1.08) |
| Scale score | 4 | -0.07 (0.91) | -0.10 (0.88) | 0.14 (0.85) | -0.03 (0.87) |
| Percent proficient | 4 | -0.10 (0.92) | -0.07 (0.96) | 0.21 (0.94) | -0.02 (1.00) |
| Scale score | 5 | 0.00 (1.01) | -0.10 (1.04) | 0.11 (0.87) | -0.02 (0.97) |
| Percent proficient | 5 | -0.03 (0.99) | -0.12 (1.06) | 0.15 (0.98) | -0.03 (1.09) |

Treatment N = 1,132 grade-level clusters; comparison N = 1,132 grade-level clusters; across 16 states.
Note: M = mean; SD = standard deviation.

**Exhibit B2. Unadjusted Mathematics Test Z-Score Performance at Baseline and Follow-Up for Members of the Treatment Group that Implemented ST Math with Fidelity and the Comparison Group, by Grade Level**

| | | Baseline | | Follow-up | |
|---|---|---|---|---|---|
| Outcome | Grade level | Treatment M (SD) | Comparison M (SD) | Treatment M (SD) | Comparison M (SD) |
| Scale score | 3 | 0.04 (0.86) | -0.01 (1.09) | 0.34 (0.78) | -0.04 (1.01) |
| Percent proficient | 3 | 0.05 (0.87) | 0.01 (1.13) | 0.41 (0.84) | 0.02 (1.09) |
| Scale score | 4 | 0.05 (0.88) | 0.13 (0.98) | 0.40 (0.73) | 0.11 (0.97) |
| Percent proficient | 4 | 0.03 (0.90) | 0.12 (0.98) | 0.50 (0.77) | 0.12 (0.99) |
| Scale score | 5 | 0.10 (0.91) | 0.10 (1.04) | 0.31 (0.86) | 0.10 (1.01) |
| Percent proficient | 5 | 0.07 (0.90) | 0.10 (1.02) | 0.34 (0.89) | 0.09 (1.04) |

Treatment N = 497 grade-level clusters; comparison N = 497 grade-level clusters; across 14 states.
Note: M = mean; SD = standard deviation.

# Appendix C. Sample Selection Flow