Randomized Trial of Elementary School ST Math Software Intervention Reveals Significant Efficacy

Mark Bodner and Andrew Coulson
MIND Research Institute
Research Park, Irvine, California

June 2021

1.1 Abstract

A randomized controlled trial group design study funded by IES NCR Grant R305A090527 was conducted in which 16,307 3rd, 4th, and 5th grade students in 52 school grade-level clusters were randomly assigned to receive ST Math (program revision Gen3), a supplemental mathematics software instructional intervention, or to a business-as-usual mathematics instruction control. After checking for acceptable attrition, and with control for baseline equivalency, Intent-to-treat (ITT) hierarchical linear model (HLM) findings on the California Standards Test math reveal a significant difference in differences treatment effect at the student level across all grades ($p<0.0005$). Analysis of individual grade-levels reveals larger increases in student standardized math scores in schools receiving the intervention than controls, with ANCOVA revealing significantly higher increases achieved individually for 4th and 5th grades ($p<0.001$).

Reference this paper: Bodner, M., Coulson, A. (2021). *Randomized Trial of Elementary School ST Math Software Intervention Reveals Significant Efficacy*. Irvine, CA: MIND Research Institute. Retrievable from https://www.mindresearch.org/our-research.

1.2 Introduction

The Institute of Educational Sciences (IES) funds rigorous large-scale studies of educational interventions. IES What Works Clearinghouse (WWC) specifies characteristics of study designs and analytic methods regarding attrition, baseline equivalence, significance and reporting of findings.[1] This study evaluates an IES Goal NCR Efficacy and Replication grant R305A090527 large scale randomized controlled trial (RCT) at 52 schools and evaluates the study in accordance with WWC guidance.

Previous research has been published on the longitudinal impact of ST Math in this RCT utilizing substantially different main impact analysis methods (Rutherford et al., 2014; Wendt et al. 2014; Schenke et al., 2014; Wendt et al., 2019). However, none of the foregoing were aligned with the simple group comparison IES RCT guidelines. The work in this paper fully aligns WWC guidelines (What Works Clearinghouse Standards Handbook, Version 4.1, 2020), methods, and terminology to the same RCT experiment and dataset, and finds and reports the resultant method statistics, impact significance, and magnitude.

This study addresses the primary research question of whether the ST Math intervention produced gains in mathematics on average for all students and within their respective grades, as measured by state standardized test scores and compared via random cluster assignment to a control group. This method adds to the previous longitudinal evaluation of ST Math.

The study is carried out on data from a randomized controlled trial (RCT) conducted on a large scale of 16,307 3rd, 4th, and 5th grade students in grade-level clusters in 52 elementary schools in Southern California. Utilizing an ITT analysis, we compared within the same grades, intervention students receiving ST Math with comparison students receiving business-as-usual mathematics instruction at their schools. The results of HLM significance analysis on the entire dataset of students reveals a significantly higher performance ($p<0.001$) in mathematics is produced by the ST Math Intervention, as measured by the California Standards Test for mathematics.

Further, within individual grades (3rd, 4th, and 5th) impact analysis of ST Math intervention students and comparison students controlled for baseline characteristic show the treatment group performed higher within grade than the control group, with ANCOVA analysis showing 4th and 5th grade treatment groups scoring significantly higher ($p<0.001$, $g=0.062$ 4th grade and $g=0.096$ 5th grade respectively) than their respective control groups.

1 Acknowledgement: The authors would like to acknowledge and thank the IES WWC for its guidelines and methodologies, as laid out in the What Works Clearinghouse Standards Handbook, towards the carrying out and completion of this work.

1.3 Experimental Design

The RCT experimental design used random assignment at the school level. Students were assigned at the school level, by grade-level-cluster, to two conditions—the treatment condition receiving the ST Math supplemental intervention or the control condition receiving business-as-usual mathematics instruction.

The experiment began in 2008/09 for an initial cohort of 34 schools and was repeated the following school year for a cohort of 18 schools (Table 1). Randomization was blocked by schoolwide percent English Learner (EL) and then distributed equally between all 52 schools. All of a school's grade-level-cluster of students, classrooms and teachers for a given grade as a group were assigned by grade-level either to the treatment or control condition. All students included in the analytic sample were required to have baseline CST scores for the year immediately prior to the first year of the study. As standardized test scores are acquired beginning in 2nd grade, this study included analysis of all students from 3rd through 5th grades (Table 1).

Cohort 1 Group	2007-2008 CST Baseline	2008-2009 CST Study Year	N
Treatment	2nd grade	3rd grade	1,834
Control	2nd grade	3rd grade	1,489
Treatment	3rd grade	4th grade	1,421
Control	3rd grade	4th grade	1,719
Treatment	4th grade	5th grade	1,512
Control	4th grade	5th grade	1,720

Cohort 2 Group	2008-2009 CST Baseline	2009-2010 CST Study Year	N
Treatment	2nd grade	3rd grade	660
Control	2nd grade	3rd grade	820
Treatment	3rd grade	4th grade	834
Control	3rd grade	4th grade	676
Treatment	4th grade	5th grade	714
Control	4th grade	5th grade	717

Total All Cohorts and Grades	N
Treatment	6,975
Control	7,141

Table 1: Experiment Cohort Grade-Level Random Assignment Conditions

1.4 The Intervention

Created by the non-profit MIND Research Institute (MIND), ST Math teaches mathematical reasoning through spatial temporal representation in which key concepts are illustrated with dynamic imagery that minimizes, at least initially, mathematical symbols and terminology. ST Math is delivered via computer and uses an interactive interface to present individualized instruction according to the student's pace of learning. The game-like exercises are formulated to engage and motivate students to solve mathematics problems and to advance steadily through the curriculum. Successive games present problems of increasing difficulty, eventually leading to quite challenging, multi-step problem solving (Rodgers et al., 2003; Bodner et al., 2004; Peterson and Bodner, 2009).

ST Math is a supplemental program (Buschkuehl, 2020). The program is delivered as grade-level curricula aligned to mathematics standards for K-5 students. Students play through a series of modules that cover specific mathematics learning objectives, moving to higher levels only after mastering each current level. Students must complete a level's items with 100% mastery (with 1 item replay allowed) or they repeat the level until a score of 100% is attained. Initial level scaffolding ensures that the material is appropriate for the student's current understanding of the mathematics material.

The typical game-play goal is to build structures or remove obstacles to enable an animated penguin (named JiJi), to move across the computer screen (see Figure 1). Within each puzzle students arrange virtual manipulatives into structures (*e.g.* bridges). The virtual manipulatives present a standards-aligned math problem item visually. Once the student posits their solution and launches the animation, the manipulatives follow rigorous mathematical rules to show why the posited answer did, or did not, solve the math problem. As each Level is passed, program scaffolding presents increasingly more challenging puzzles, that is more advanced mathematical problems. The game elements are the mathematics itself in that these obstacles blend into the mathematical problems presented such that there is often little or no distinction between the game and the mathematics.

The content of each grade contains modules that match curricular units found in traditional classroom instruction with focus on a mathematical concept—for example multiplication and division, addition and subtraction, estimation, and so on. Each module consists of a number of games, and within each game there are generally four to five levels of increasing mathematical difficulty. Each game has its own consistent scenario and rules.

Figure 1 gives an example and displays a level of the game "JiJi Cycle." In JiJi Cycle, students see JiJi the penguin on a virtual manipulative: a unique type of cycle with wheels consisting of several disks whose circumferences can equal 1 or fractions of unity (*e.g.* a disk with slices removed forming one half of a wheel, two-thirds of a wheel, etc.) The ground in the game corresponds to a segment of the real number line. Based on the structure of the cycle (*i.e.* the number that the sum of the wheels' circumferences totals) students must estimate where to position a ballooned platform at the appropriate position on the number line, such that it corresponds to the sum of the circumferences of the wheels of the given cycle to within an estimated error (only arced portions of the circumferences of the wheels are included in the sum). The estimated error of placement has an acceptable value defined by the width of the platform.

If the platform is placed correctly to within the error allowed by its width, JiJi proceeds to the platform (using up the wheels of the cycle as it moves along the number line) and will land on the platform, being subsequently carried off of the screen by the balloons attached to the platform. If the estimation of platform placement was not within the error allowed by the platform width, then JiJi will undershoot or overshoot the platform after using up all the wheels of the cycle, and the platform subsequently takes off without JiJi aboard. The game covers visual concepts of addition of integers and non-integers on the number line, in addition to estimation and defining concretely what is a good estimation.

JiJi Cycle is the first game in the 5th grade curriculum's ninth module, "Fractions on the Number Line." Once students complete this game, they move on to the rest of the games in the module and subsequently to the next modules in the curriculum. There are thirteen 3rd-grade modules, twelve 4th-grade modules, and thirteen 5th-grade modules in the 3rd, 4th, and 5th grade ST Math version generation 3 curricula respectively.

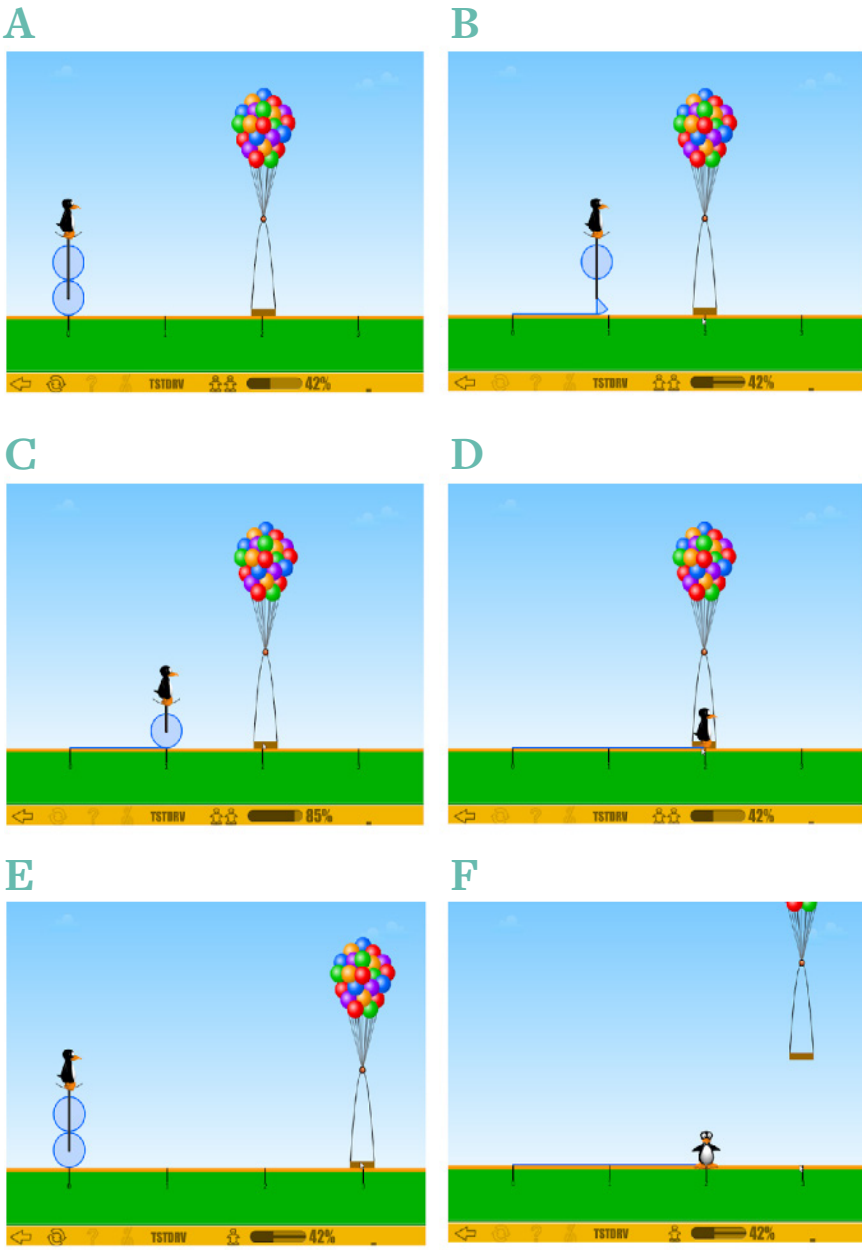

Figure 1

A puzzle in a level of the first game called JiJi Cycle in the 5th-grade curriculum's 9th module. Virtual manipulatives represent addition of integers and non-integers on the number line. A-D shows screenshots of the game animation sequence for a correct response to a puzzle. Each puzzle exhibits a different cycle whose wheels' circumferences sum to an integer or non-integer (*e.g.* cycles may possess wheels which are half, two thirds, or some other fraction of a disk).

A. The puzzle is presented in which JiJi rides atop of a cycle whose wheels consist of two disks. In this puzzle each of wheels' circumferences equals 1. Utilizing one complete rotation of each of the two wheels will enable the cycle to travel 2 units along the number line as the sum of the circumferences of the wheels is equal to 2. For the correct response, the student uses the cursor to slide the ballooned platform to the location on the number line corresponding to 2.

B. The cycle proceeds along the number line using up its wheels' circumferences as it moves.

C. The first unit wheel has been used up and has progressed JiJi and the cycle, having reached the number 1 on the number line on its way to the number 2.

D. The second unit wheel subsequently has deposited JiJi at the number 2 on the number line, landing JiJi on the ballooned platform manipulative whose center had been correctly placed at that location by the student. The platform will subsequently float off the screen with JiJi aboard.

E. An incorrect response to the same puzzle shown in (A-D). The cycle possesses two circles with unit circumference summing to the number 2, but the platform's center is placed at the number 3 on the number line.

F. The cycle progresses to and halts at the number two, using up its two unit-circumference wheels, dropping JiJi off one full unit short of the student's placement of the ballooned platform's center (outside the allowed estimation error defined by the width of the platform which is ±0.2). The platform subsequently takes off without JiJi aboard.

1.5 Sample

1.5.1 Experiment Schools

The sample consisted of all 3rd-, 4th-, and 5th-grade students in 52 low-performing schools within ten districts in Southern California. Student population descriptive academic performance index statistics of the analytic sample are given in Table 2. All schools were in the lowest three deciles of California API academic math test performance index. Schools were recruited to join the experimental study as part of a local grants-funded roll out of ST Math program licenses. The roll out of the intervention started in year one at two grade levels per school site. The experiment designated the intervention and comparison schools as type "A" and "B." Schools were randomly assigned to one of those two conditions. In year one Type "A" schools were required to implement the intervention in grades 2 and 3 (only); Type "B" schools in grades 4 and 5 (only).

	3rd Grade				4th Grade				5th Grade			
	Control		Treatment		Control		Treatment		Control		Treatment	
	%	Count	%	Count	%	Count	%	Count	%	Count	%	Count
Male	51.9	1,199	51.6	1,286	49.8	1,192	53.3	1,201	49.0	1,193	51.0	1,136
Female	48.1	1,110	48.4	1,208	50.2	1,203	46.7	1,054	51.0	1,244	49.0	1,090
Free/Reduced Lunch	83.5	1,927	88.4	2,204	86.6	2,073	85.4	1,925	88.6	2,159	82.7	1,840
Amer. Indian/ Native American	0.0	0	0.4	9	0.1	3	0.1	3	0.2	4	0.2	5
Asian	7.7	177	5.1	126	4.1	99	6.8	153	3.0	72	6.7	149
Hawaiian/Pacific Islander	0.5	11	0.4	9	0.5	13	0.4	9	0.3	6	0.5	11
Filipino	1.6	37	1.6	40	1.7	41	1.4	32	1.6	38	1.4	32
Hispanic/Latino	82.9	1,915	86.3	2,153	86.5	2,072	83.8	1,890	87.9	2,142	82.9	1,846
Black/African American	1.7	40	1.6	41	1.7	41	1.4	31	1.4	34	1.8	39
White	5.2	121	4.3	108	4.6	109	5.9	132	5.1	125	6.2	139
Other Race	0.4	8	0.3	8	0.7	17	0.2	5	0.7	16	0.2	5
EL	67.7	1,562	66.9	1,668	57.0	1,364	57.0	1,285	44.7	1088	43.8	974
Total N	2,309		2,494		2,395		2,255		2,437		2,226	

Table 2: Student Descriptive Summary Statistics (Analytic Sample)

1.5.2 Sample Attrition

Attrition is defined as occurring when an outcome variable is not available for all subjects initially assigned to the intervention and comparison groups.

Attrition can introduce bias if the characteristics of those subjects lost are correlated to the outcome measure. Accordingly, WWC's attrition standard is based on a model that requires both overall attrition (the rate of attrition for the entire experimental sample as a percentage of the randomized sample that is lost) and differential attrition (the percentage point difference in the rates of attrition between the comparison and intervention groups) to be less than specified percentages (WWC Standards Handbook, Version 4.1, 2020: pp. 8-13; https://ies.ed.gov/ncee/wwc/Document/21).

Analysis of the attrition rate for this experiment revealed that the study qualified as a "low attrition rate" study by WWC guidelines. For the 3rd grade population, overall attrition was 12.94%, and differential attrition between groups was 1.53 percentile points. For the 4th grade population, overall attrition was 14.33%, and differential attrition was 3.06 percentile points. For the 5th grade population, overall attrition was 13.04% and differential attrition was 1.02 percentile points (Table 3). For all three grades, the attrition met the standards for a tolerable threat of bias under both optimistic and cautious assumptions (WWC Evidence Review Protocol For Elementary School Mathematics Interventions, version 2.0). None of the schools in the study dropped out of inclusion in the analytic sample as a result of attrition.

	Schools Assigned		Schools Contributing Data		Students Assigned		Students Contributing Data		Overall Attrition	Differential Attrition
	Control	Treatment	Control	Treatment	Control	Treatment	Control	Treatment		
3rd Grade	27	25	27	25	2,628	2,889	2,309	2,494	12.94	1.53
4th Grade	27	25	27	25	2,844	2,584	2,395	2,255	14.33	3.06
5th Grade	27	25	27	25	2,818	2,544	2,437	2,226	13.04	1.02

Table 3: Experimental Sample Attrition Statistics

1.6 Analysis

The current study investigates the effect of ST Math with an intent-to treat (ITT) analysis in order to preserve the integrity of the random assignment design. Therefore the measured fidelity to intervention use is not incorporated into the analysis (fidelity is described in the Discussion, see section 1.8.2).

Analysis was carried out on all 14,116 3rd, 4th, and 5th grade students meeting the outcome measures requirements for the study. Specifically, CST math scores were required to be available for the student for the baseline year immediately preceding the first year of the study, and for the study year. Since baseline CST testing was only available starting with 2nd grade, 3rd grade was the lowest grade available with baseline scores and thus the lowest grade available for inclusion in this study (Table 1).

1.6.1 Data Collection

Data collection from the districts was performed by the Orange County Department of Education, a subcontractor on the IES NCER grant. Each year OCDE collected CST data from the districts for all students in grades 3, 4, and 5.

Note: for purposes of calculating leakage and dose, the vendor provided student-level usage (minutes and lessons) at the end of each year. The vendor joined this usage to the student records with CST data (see Discussion section 1.82).

1.6.2 Outcome Measure

The impact of ST Math in this study was assessed with the California Standards Test (CST). The CST is a standardized test series that was developed to evaluate competency of California students with respect to the California State Standards (California Standards, 2010; NRC, 2001).

1.6.3 Baseline Differences

Baseline CST scores for treatment and control groups were roughly normally distributed; equality of variance tests (Levene's test) and normality checks were carried out and the assumptions met. Although as a low attrition RCT there is not a WWC requirement to control for baseline equivalence, we measured baseline equivalence on the CST and despite the random assignment detected a significant difference at each grade level between the average scores of the intervention and control groups (Table 4). This analysis revealed that a significant difference ($p<0.05$ ANOVA) of 0.11 standard deviations was present between the average baseline scores of the treatment and control groups for 3rd grade, 0.15 standard deviations for the 4th grade, and 0.11 standard deviations for the 5th grade.

Since these differences were greater than the WWC guideline for baseline equivalence of not exceeding 0.05 standard deviations, ANCOVA analyses were conducted in accordance with WWC guidelines to adjust for this difference in baseline CST scores in analyzing the impact differences between treatment and controls within each grade (WWC Standards Handbook Version 4.1, 2020: pp.13-17).

	Average Baseline CST Score Treatment Group	Average Baseline CST Score Control Group	Number of Standard Deviations Difference Treatment/Control	Significance
3rd Grade	342.86	353.50	0.11	P<0.05
4th Grade	348.74	342.51	0.15	P<0.05
5th Grade	357.88	346.13	0.11	P<0.05

Table 4: Baseline Outcome Statistic Equivalence Summary Statistics

1.6.4 Statistical Analysis

The importance of utilizing nested models for analysis of individuals clustered in like-condition groups has been demonstrated (Aitkin et al., 1981; McCoach and Adelson, 2010). When analyzing the effects of an intervention implemented at the school, grade or classroom level, an analysis at the student level may increase the probability of a Type 1 error when testing significance, due to an estimated standard error of the treatment effect that is too small.

As randomization was implemented at the school grade-level cluster in this study, we carried out HLM analysis of significance on the entire analytic sample to take into account the nested structure of the experimental conditions which were carried out on entire grade-levels of teachers and classrooms at schools.

In parallel we utilize ANCOVA analysis of significance and impact for each individual grade-level 3, 4, and 5 taking into account the difference in baseline scores in treatment and control groups.

1.6.4.1 School Nesting

A two-level null model containing no predictors was first examined to determine how much variation could be attributed to the school level and the individual student level. This null model is given by the equations:

$$\text{Level 1: } \text{CST\_Scores}_{ij} = \beta_{0j} + \varepsilon_{ij}$$
$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \mu_{0j}$$

and with substitution the mixed model is:

$$\text{CST\_Scores}_{ij} = \gamma_{00} + \mu_{0j} + \varepsilon_{ij}$$

where CST_Scores_{ij} is the dependent variable and is the score obtained on the CST Math test by student i at school j, ε_{ij} is the residual error at the student level (student i, at school j), γ_{00} is the grand mean across schools of the intercepts for CST scores, and μ_{0j} is the error or variation in intercepts across schools of the grand mean CST scores. The results of the null model are shown in Table 5 below.

					95% Confidence Interval	
Parameter	Estimate	Std. Error	Wald Z	Sig.	Lower Bound	Upper Bound
Residual	6663.927417	79.471286	83.853	.001	6509.972815	6821.522897
Intercept	257.063304	56.583553	4.543	.001	166.985071	395.733236

Table 5: Estimates of covariance parameters for the two-level null HLM model determining the ICC. The dependent variable is CST math scores. The Intercept estimate parameter corresponds to variation from schools. The residual estimate parameter corresponds to the variation between students.

From the table the intraclass correlation coefficient (ICC) value may be determined to be

$$\text{ICC} = 257.063304/(6663.927417+257.063304) = 0.037$$

This ICC based on the HLM null model thus indicates that approximately 3.7% of the variation in student CST performance was due to between-school differences. Though not large, this value indicates that students within schools do exhibit some correlation with each other and thus the significance evaluation utilizing HLM to account for clustering is warranted (Yasuyo et al., 2014).

1.6.4.2 Cohort Nesting

A three-level null model was also examined to determine cluster-level impact of schools (Level 2) given that they were nested within the two different schools cohorts (Level 3) and spread across two separate experiment school years of 2008/09 and 2009/10 (Cohort 1 =0, Cohort 2=1). That is we examined the mixed model:

$$CST\_Scores_{ijk} = \gamma_{000} + \mu_{00k} + r_{0jk} + \varepsilon_{ijk}$$

Where CST_Scores_{ijk} is the dependent variable CST math scores measured for student i, in school j, nested in cohort k. γ_{000} is the grand mean across all groups, μ_{00k} is the random effects from cohort k, r_{0jk} is the random effects from school j in cohort k, and ε_{ijk} is each student's random deviation. The results of this three-level null model revealed that no significant variation occurred from the nesting of schools in different cohorts (Table 6)—cohort nesting ICC=0.0018. Thus the cohorts were aggregated and the data was analyzed utilizing a two-level HLM with nesting at schools in level 2 (Figure 2).

					95% Confidence Interval	
Parameter	Estimate	Std. Error	Wald Z	Sig.	Lower Bound	Upper Bound
Residual	6663.927417	79.471286	83.853	.001	6509.972815	6821.522897
Intercept (Cohort)	12.131981	34.568127	0.352	0.725	0.046055	3206.395968
Intercept (School)	257.283904	57.733780	4.456	0.001	165.731629	399.410830

Table 6: Three-Level (Cohort) HLM null model extending the two-level HLM to examine variability due to student nesting in schools at level 2 (Table 5), with schools nesting into two different cohorts at Level 3.

1.6.4.3 HLM Model Used

The next stage of the HLM analysis extended the null model to a two-level mixed model, adding the level 1 (student level) predictors of baseline CST scores ($BaseScore_{ij}$ a continuous variable) and grade ($Grade4_{ij}$ and $Grade5_{ij}$; binary variables denoting the grade of students normalized to 3rd grade). A level 2 predictor (school level) of treatment group was added ($treatment\_Group_j$ a binary variable; control group=0, ST Math intervention group=1)—see Figure 2. Baseline CST scores slope-variation between schools was taken into account in the mixed model.

That is, the equations describing the two-level HLM are:

(Student level) Level 1: $$CST\_Scores_{ij} = \beta_{0j} + \beta_{1j}*BaseScore_{ij} + + \beta_{2j}*Grade4_{ij} + \beta_{3j}*Grade5_{ij} + \varepsilon_{ij} \quad (1)$$

(School level) Level 2: $$\beta_{0j} = \gamma_{00} + \gamma_{01}*Treatment\_Group_{j} + \mu_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j} \quad (3)$$

$$\beta_{2j} = \gamma_{20} \quad (4)$$

$$\beta_{3j} = \gamma_{30} \quad (5)$$

Where μ_{1j} in equation (2) is the deviation in slope from the overall average for school j.

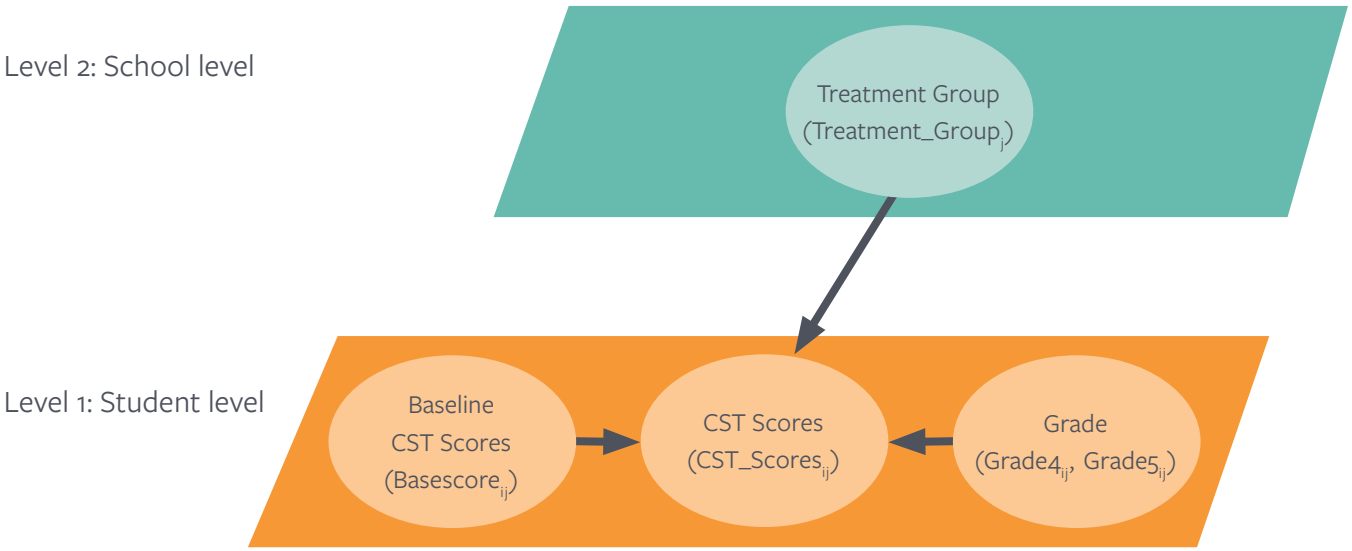

Figure 2: HLM Model: Levels and Predictors

The resulting mixed model equation is given as:

$$CST\_Scores_{ij} = \gamma_{00} + \gamma_{01}*Treatment\_Group_{j} + \mu_{0j} + \gamma_{10}*BaseScore_{ij} + \mu_{1j}*BaseScore_{ij} + \gamma_{20}*Grade4_{ij} + \gamma_{30}*Grade5_{ij} + \varepsilon_{ij}$$

The coding of the variables is given in Table 7 below. HLM analysis was carried out in SPSS (IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp).

	Base Score		TreatmentGroup		Grade4		Grade5	
	Control	Treatment	Control	Treatment	Control	Treatment	Control	Treatment
3rd Grade	Continuous	Continuous	0	1	0	0	0	0
4th Grade	Continuous	Continuous	0	1	1	1	0	0
5th Grade	Continuous	Continuous	0	1	0	0	1	1

Table 7: HLM Variables Coding

1.6.4.4 ANCOVA Analysis

ANCOVA were conducted within each grade-level 3, 4, and 5 with baseline CST math scores as the covariate, CST math scores post-intervention as the dependent variable, and group (Treatment or Control) as the independent variable. ANCOVA analysis was carried out in Matlab (Mathworks Inc, Matlab 2020a).

1.7 Significance

The results of the HLM analysis revealed a significant treatment effect from the ST Math Intervention across all grades (Tables 8). The results of the ANCOVA analysis at individual grades revealed that a significant treatment effect was present for the 4th and 5th grade populations. For 3rd grade, while the increase in math performance was marginally greater for the ST Math treatment group than the control group, that difference did not reach the level of significance (Tables 9, 10).

1.7.1 HLM Results Significance

The estimate of HLM model fixed effects shows that the ST Math intervention (Treatment_Group parameter) is seen to be significant (P<0.0005), with ST Math students producing an estimate average increase compared to the control students across grades of 4.57 points on the CST exam. Differences in Baseline CST scores (BaseScore parameter) is a significant covariate with the ST Math intervention students having a significantly higher average baseline score (p<0.0005) than control students of 0.83 points across grades. While ST Math intervention produced larger increases from baseline scores for each grade, student grade-level was a source of variability in CST scores, with 5th grade students (Grade5 parameter) showing an estimated significantly (P<0.0005) lower average CST relative to 3rd grade students of 12.31 points, and 4th grade students (Grade4 parameter) exhibited a non-significant (p=0.28) lower average CST scores of 1.16 points on average.

						95% Confidence Interval	
Parameter	Estimate	Std. Error	df	t	Sig.	Lower Bound	Upper Bound
Intercept	74.406343	2.434218	336.629	30.567	.0005	69.618148	79.194537
Grade4	-1.160961	1.075436	14057.342	-1.080	.280	-3.268957	.947036
Grade5	-12.304550	1.076142	14103.375	-11.434	.0005	-14.413930	-10.195169
BaseScore	.827279	.006614	270.446	125.081	.0005	.814258	.840301
Treatment_ Group	4.566090	.924502	13735.585	4.939	.0005	2.753941	6.378239

Table 8: Estimate of HLM model fixed effects: The results show a significant effect from the ST Math intervention across grades (Treatment_Group: p<0.0005).

1.7.2 ANCOVA Results

Impact: The ANCOVA analysis of the entire 3rd grade population showed that the increase in CST math scores from baseline for the ST Math treatment group was higher than for the control group (treatment group increase in mean score 18.18 points, control group increase 14.09 points, difference of differences 4.09 points—see Tables 9, 10). This difference did not reach significance [F(1,4802)=1.977843, p=0.16, g=.072].

The ANCOVA analysis of the entire 4th grade population showed that the increase in CST math scores from baseline for the ST Math treatment group was higher than for the control group (treatment group increase in mean score 14.62 points, control group increase 13.05 points—difference of differences 1.57 points—see Tables 9, 10). This difference was significant [F(1, 4649)=19.01858, p<0.001; g=.072)].

The ANCOVA analysis of the entire 5th grade population showed that the increase in CST math scores from baseline for the ST Math treatment group was higher than for the control group (treatment group increase in mean score 6.12 points, control group increase 0.88 points—difference of differences 5.24 points—see Tables 9, 10). This difference was significant [F(1,4662)=16.38027, p < 0.001; g=.096].

	Baseline CST Scores		Post Treatment CST Scores		Effect Size
	Mean	Std.	Mean	Std.	
3rd Grade Control	353.50	78.82	367.59	84.76	g=.072
3rd Grade Treatment	342.86	79.72	361.04	86.28	
4th Grade Control	342.51	78.00	355.56	76.64	g=.062
4th Grade Treatment	348.74	79.28	363.36	77.99	
5th Grade Control	346.13	72.85	347.01	85.08	g=.096
5th Grade Treatment	357.88	71.63	364.00	83.78	

Table 9: Average Scores and Effect Sizes (Hedge's g) 3rd, 4th, and 5th Grades

		Sum of squares	df	Mean square	F	Significance
3rd grade	Between groups	5998.545	1	5988.545	1.978	P=0.16
	Error	14557783	4800	3032.872		
	Total	14563782	4803			
4th grade	Between groups	44943.23	1	44943.23	19.01858	p<0.001
	Error	10981427	4647	2363.122		
	Total	11026370	4650			
5th grade	Between groups	47764.57	1	47764.67	16.38927	P<0.001
	Error	13588475	4660	2915.982		
	Total	13636239	4663			

Table 10: ANCOVA Results 3rd, 4th, and 5th Grades

1.7.3 Statistical Significance and Impact Summary

This study demonstrated that the ST Math intervention (version Generation 3) increased math performance at significant levels ($p < 0.0005$) across grades as determined from HLM Analysis (Tables 8 and 9). Analysis of individual grades carried out (ANCOVA analysis) reveal that the ST Math intervention increased math performance across 3rd, 4th, and 5th grades, and at significant levels ($p<0.001$) for both the 4th and 5th grades (Table 10).

1.8 Discussion

Recap

This is a large scale RCT study, with assignment of grade-level clusters, stratified before randomization by school percent EL, and the unit of analysis being individual student scaled CST scores. Baseline equivalence was factored by ANCOVA and HLM. Attrition of grade-level clusters lost and individuals lost fell within WWC boundaries for having tolerable threat of bias under both optimistic and cautious assumptions. The increased math performance, as measured by the standardized CST, was produced across grades.

1.8.1 Discussion and Contrast to Prior Study

In the present study we examined the effect of the intervention within one year—the first school year of the intervention and across grades, as opposed to examining changes in efficacy or cumulative effects across years and grade-levels as examined in previous work (Rutherford et al., 2014). Those results examining the change across years and grades of the ST Math intervention on CST performance revealed marginal significance on cumulative increase in math performance was obtained across grades.
These findings are consistent with our analysis insofar as our analysis of difference of differences revealed increases in math mean CST scale score from baseline to post-treatment between treatment and control groups for each grade, but not an increasing differentials in scores from grade to subsequent grade: 3rd grade difference between control and treatment groups increase baseline to post-intervention 4.09 points (effect size g=.077), 4th grade 1.57 points (effect size g=.062), and 5th grade 5.24 points (effect size g=.096). These values for effect size are similar to those found for other RCTs (Cheung and Slavin, 2013).

1.8.2 The Implementation

1.8.2.1 Leakage

In the present study we estimated the ITT effect, that is, the effect of having been assigned to the intervention rather than the effect of actually receiving the intervention. However, it is of interest to note the degree of leakage between treatment and control groups. For the present study treatment group compliers were those assigned to that group completing any percentage of the ST Math intervention by April when the CST test was administered, with non-compliers being those assigned to the treatment group completing 0% of the ST Math intervention by that date. Similarly, for those assigned to the control group, non-compliers were considered as those completing any percentage of the intervention before the April date. The amount of leakage in this study is indicated in Table 11.

For 3rd grade there was 70.8% compliance (29.2% non-compliers) for the treatment group and 97.1% compliance (2.1% non-compliers) for the control group. For 4th grade there was 68.0% compliance (32.0% non-compliers) for the treatment group and 95.8% compliance (4.2% non-compliers) for the control group. For 5th grade there was 69.8% compliance (30.2% con-compliers) for the treatment group and 99.7% compliance (0.3% non-compliers) for the control group.

	Treatment Group		Control Group		Treatment Group
	Percent Compliers	Average Completion	Percent Compliers	Average Completion	Average Completion Compliers
3rd Grade	70.8	47.8	97.1	2.4	67.6
4th Grade	68.0	51.0	95.8	3.0	75.0
5th Grade	69.8	44.7	99.7	0.2	64.0

Table 11: Experimental Condition Compliance & Program Completion Percent

1.8.2.2 Dose

For the version of ST Math evaluated in this paper (generation 3), students within the same grade all began the year on the same first game within the software and proceeded through a subsequent fixed sequence of games as they solved them. For optimal implementation ST Math requires that students complete all the software modules in the curriculum (Peterson and Patera, 2006). In the current study, students were to spend two 45-min sessions/week on the program for an average (34 weeks) total of 68 sessions/year. Because the program requires a Level-pass to progress and allows as many Level-attempts as each student requires, students realize individual rates of progress meaning that, for any given amount of minutes, students progress to different places within the software—different levels, different games, and different modules. Thus, students at the testing date achieved different completion percentages of the program (*i.e.* received a range of "dosages" of the ST Math intervention).

For the purposes of this study dosage was considered as the percent of program full curriculum completion by April when the CST test was administered. Program completion percentages may have increased in many schools after April, but this was not considered in the calculation of dosage as it could have no effect on performance on the CST exam as measured in this study. On average, across the entire analytic sample, students completed 47.9% of the curriculum by April when the CST test was administered. These averages included 30.4% of the treatment group who were non-compliers (non-compliers consisting of students assigned to the treatment group with 0% completion by April when the CST test was administered). Compliers in the treatment group on average completed 68.8% of the curriculum. Average completion overall and for compliers for each individual grade-level is given in Table 11.

1.8.3 Future Work

Future work will examine these effects further. There was material leakage of assigned condition (Table 11), so specifically, an examination of the compliers and non-compliers with a Complier Average Causal Effect (CACE) correction (What Works Clearinghouse Standards Handbook Version 4.1) will be carried out to refine the impact estimate of the ST Math intervention. Furthermore, we will examine the effects of the continuous dose variable of ST Math intervention on math performance.

References

Aitkin, M., Anderson, D., Hinde, J. (1981). Statistical modelling of data on teaching styles. Journal of the Royal Statistical Society, Series A, 144: 419-461.

Bodner, M., Peterson, M.R., Shaw, G.L. (2004). System and method for analysis and feedback of student performance. (United States Patent No. US2004 0180317A1).

Bushkuehl, M. (2020). Spatial-Temporal Math: Underlying scientific concepts and mechanisms. MIND Research Institute white paper.

California Standards, 2010; NRC, 2001.

Cheung, A.C., Slavin, R.E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in k-12 classrooms: A meta-analysis. Educational Research Review, 9 88-113.

IBM SPSS Statistics for Windows, Version 27.0 Armonk, NY: IBM Corp.

Mathworks Inc, Matlab 2020a.

McCoach, D.B., Adelson, J.L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. Gifted Child Quarterly 54(2): 152-155.

Peterson, M.R., Patera, J. (2006). Non-language-based instruction in mathematics. Paper presented at the conference of the International Commission for the Study and Improvement of Mathematics Education.

Peterson, M.R., Bodner, M. (2009). System and method for training with a virtual apparatus (United States Patent No. US20090325137A1).

Rodgers, L., Bodner, M., Shaw, G.L. (2003). Method and system for teaching vocabulary (United States Patent No. US20030165800A1).

Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J., Richland, L., Tran, N., Schneider, S., Duran, L., Martinez, M.E. (2014). A randomized trial of an elementary school mathematics software intervention: Spatial-Temporal Math. Journal of Research on Educational Effectiveness 7.

Schenke, K., Rutherford, T., Farkas, G. (2014). Alignment of game design features and state mathematics standards: Do results reflect intentions? Computers & Education 76: 215-224.

Wendt, S., Rice, J., Nakamoto, J. (2014). Evaluation of the MIND Research Institute's Spatial-Temporal Math (ST Math) program in California. WestEd.

Wendt, S., Rice, J., Nakamoto, J. (2019). A cross-state evaluation of MIND Research Institute's ST Math program and math performance. WestEd.

What Works Clearinghouse Standards Handbook Version 4.1 (2020).

Yasuyo, A., Gee, K.A. (2014). Sensitivity analyses for clustered data: An illustration from a large-scale clustered randomized controlled trial in education. Evaluation and Program Planning, 47: 26-34.