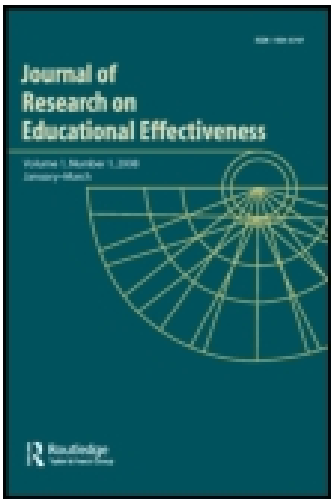

Journal of Research on Educational Effectiveness

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uree20

A Randomized Trial of an Elementary School Mathematics Software Intervention: Spatial-Temporal Math

Teomara Rutherford[a], George Farkas[a], Greg Duncan[a], Margaret Burchinal[b], Melissa Kibrick[a], Jeneen Graham[a], Lindsey Richland[c], Natalie Tran[d], Stephanie Schneider[e], Lauren Duran[e] & Michael E. Martinez[a]

[a] University of California, Irvine, Irvine, California, USA

[b] University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[c] University of Chicago, Chicago, Illinois, USA

[d] California State University, Fullerton, Fullerton, California, USA

[e] Orange County Department of Education, Costa Mesa, California, USA

Accepted author version posted online: 10 Mar 2014.Published online: 29 Sep 2014.

To cite this article: Teomara Rutherford, George Farkas, Greg Duncan, Margaret Burchinal, Melissa Kibrick, Jeneen Graham, Lindsey Richland, Natalie Tran, Stephanie Schneider, Lauren Duran & Michael E. Martinez (2014) A Randomized Trial of an Elementary School Mathematics Software Intervention: Spatial-Temporal Math, Journal of Research on Educational Effectiveness, 7:4, 358-383, DOI: 10.1080/19345747.2013.856978

To link to this article: http://dx.doi.org/10.1080/19345747.2013.856978

Journal of Research on Educational Effectiveness, 7: 358–383, 2014

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2013.856978

Routledge
Taylor & Francis Group

A Randomized Trial of an Elementary School Mathematics Software Intervention: Spatial-Temporal Math

Teomara Rutherford, George Farkas, and Greg Duncan
University of California, Irvine, Irvine, California, USA

Margaret Burchinal
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Melissa Kibrick and Jeneen Graham
University of California, Irvine, Irvine, California, USA

Lindsey Richland
University of Chicago, Chicago, Illinois, USA

Natalie Tran
California State University, Fullerton, Fullerton, California, USA

Stephanie Schneider and Lauren Duran
Orange County Department of Education, Costa Mesa, California, USA

Michael E. Martinez
University of California, Irvine, Irvine, California, USA

Abstract: Fifty-two low performing schools were randomly assigned to receive Spatial-Temporal (ST) Math, a supplemental mathematics software and instructional program, in second/third or fourth/fifth grades or to a business-as-usual control. Analyses reveal a negligible effect of ST Math on mathematics scores, which did not differ significantly across subgroups defined by prior math proficiency and English Language Learner status. Two years of program treatment produced a nonsignificant effect. Publication of evaluation results from large-scale real-world supplemental mathematics instructional implementations such as this one can provide a realistic view of the possibilities, costs, and limitations of this and other computer aided instruction supplemental interventions.

Keywords: Elementary mathematics, experimental design, computer-assisted instruction

The existing literature comprises few rigorous evaluations of mathematics curricula or instructional practices, especially those implemented on a large scale and with high quality. Thus, of the 77 reports examining interventions in elementary mathematics education within the Institute of Education Science's What Works Clearinghouse (WWC), only five met the highest WWC category of evidence, another five provided evidence meeting somewhat lower standards, and the remainder provided too little valid evidence to support claims regarding effectiveness (U.S. Department of Education WWC, 2013a). To meet the need

Address correspondence to Teomara Rutherford, University of California, Irvine, 3200 Education, Irvine, CA 92697-5500, USA. E-mail: teyarutherford@gmail.com

for rigorous research on elementary mathematics interventions, the present study reports on an independent evaluation of one computer-based supplementary mathematics instructional program, Spatial Temporal (ST) Math, based on a randomized control trial (RCT) conducted with more than 13,000 students in 52 elementary schools in Southern California. Large-scale implementation, combined with random assignment to condition, allow us to potentially detect and quantify a causal relationship between student participation in the program and educational outcomes (see Shadish, Cook, & Campbell, 2002).

PRIOR RESEARCH ON ELEMENTARY MATHEMATICS INTERVENTIONS

Of the 10 studies meeting WWC standards, only three have been designated as having "potentially positive effects," with the bulk having "no discernible" or "mixed" effects (U.S. Department of Education WWC, 2013a). No computer-assisted intervention (CAI) for elementary school mathematics has been listed within the WWC as showing positive or potentially positive effects although every year districts spend millions of dollars on these programs.

A meta-analysis conducted by Slavin and Lake in 2008 uncovered no studies providing "strong evidence" (a randomized study with at least 10 classes or schools or 250 students assigned to treatments, p. 476) of positive educational effectiveness among mathematics software programs and found little or no significant differences between treatment and control students among studies meeting their lower standard for "moderate evidence" (Slavin & Lake, 2008, p. 477). Since 2008, few published evaluations have met these criteria (for one exception see Roschelle et al., 2010). Cheung and Slavin (2013) used rigorous inclusion criteria for their updated meta-analysis on K-12 educational technology for mathematics. They included only evaluation of programs lasting longer than 12 weeks, noting the bias toward stronger effects with programs of shorter durations. Cheung and Slavin found small positive effects (effect size of .18) for CAI, contrasting these smaller effects with older meta-analyses, which they viewed as overstating effect sizes by including inappropriate studies. For studies evaluated with RCTs, the effects were even smaller, .08 (Cheung & Slavin, 2013, p. 99).

While praising CAI, the National Mathematics Advisory Panel (NMAP) also called for further research, especially on the scale-up of Integrated Learning Systems, programs that include both tutorial and drill and practice elements (NMAP, 2008a). Positive results from small controlled studies have proven especially difficult to achieve at scale given issues of curricular integration and fidelity of implementation (NMAP, 2008a). Further work with real-world implementations and rigorous evaluations of mathematics interventions would explicate the potential of CAI, for whom it might be most effective and why. Our current study presents a large-scale evaluation of ST Math to meet this need.

Evaluations of CAI programs such as ST Math may be particularly important, because these programs are being widely implemented under the belief that they provide a significant educational benefit with relatively small investments of time and money by the schools (see Slavin & Lake, 2008). Yet even optimistic projections do not claim that CAI programs are a panacea for low performance—blanket application of CAI is unlikely to raise all students to proficiency in all subjects.

Moreover, it is not clear for whom and within what situations specific CAIs have the greatest effect on student outcomes (NMAP, 2008b; e.g., Roschelle et al., 2010). In this study, we examine whether there are Student Characteristic $\times$ ST Math interaction effects on mathematics outcomes for students' language status (English Language Learner [ELL] or not) and beginning of the study mathematics proficiency. Better understanding

individual differences in response to interventions provides important information regarding the effectiveness and replicability of an intervention for different population subgroups (see NMAP, 2008b; U.S. Department of Education WWC, 2013b).

THE ST MATH INTERVENTION

Description and Use of ST Math Software

Created by the nonprofit MIND Research Institute (MIND), ST Math is designed to teach mathematical reasoning through spatial temporal representations in which key concepts are illustrated with dynamic imagery that minimizes, at least initially, mathematical symbols and technical 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 gamelike 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, multistep problem solving. Program developers report that ST Math is currently used by about 473,000 K-8 students across 1,355 schools in 24 states, with the largest concentration of schools in California, Texas, Florida, and Illinois. Eighty-three percent of the student users are eligible for free or reduced lunch. The initial student licensing fee is a maximum of $50 per student, which is, for reference, comparable to the cost of a textbook (see California Department of Education IMPL, 2013, where approved textbooks are listed for approximately $80/student), keeping in mind that ST Math is a supplement to the expense of the textbook. For subsequent years, schools pay a $35 per student renewal fee. Larger schools may choose to purchase a site license instead, saving over the individual student licensing fee by paying $49,000 for the first year and $3,750 each year thereafter.

As noted, ST Math is a supplemental program to the school's mathematics curriculum. According to the developers, full implementation requires two 45-min sessions per week in the computer lab. The program is divided into grade-level lessons designed to parallel mathematics standards for K-5 students. Linear game-play allows students to move to a higher level only after they have mastered the current level. Students have two "lives" in each level, which means they must finish the level before making two mistakes (80% mastery) or they must repeat the level. This self-paced structure ensures that the material is appropriate for the student's current abilities. The visible goal of the games is to help Jiji, an animated penguin, move from the left to the right side of the screen. Within levels, students build bridges and remove obstacles in Jiji's path by solving increasingly more advanced mathematical puzzles. These bridges and obstacles blend into the mathematical puzzles such that there is often little distinction between the game and the mathematics—in other words, the game elements *are* mathematics.

The content for each grade, K-5, contains several *modules* that match curricular units found in more traditional classroom instruction with focus on a mathematical concept such as *Addition and Subtraction Situations.* Each module contains several games (see Figure 1a), and within each game, there are between one and 10 levels of increasing mathematical difficulty. Each game has its own consistent scenario and rules. Figure 1b to 1d displays level one of the game "Push Box." In Push Box, students see a ramp on top of a box (the number of boxes varies with levels), and they see a bulldozer poised to push a collection of boxes on top of the ramp. Students must choose the correct sum from the boxes on the right. This sum represents where they will place a bridge to allow Jiji, waiting to ride the

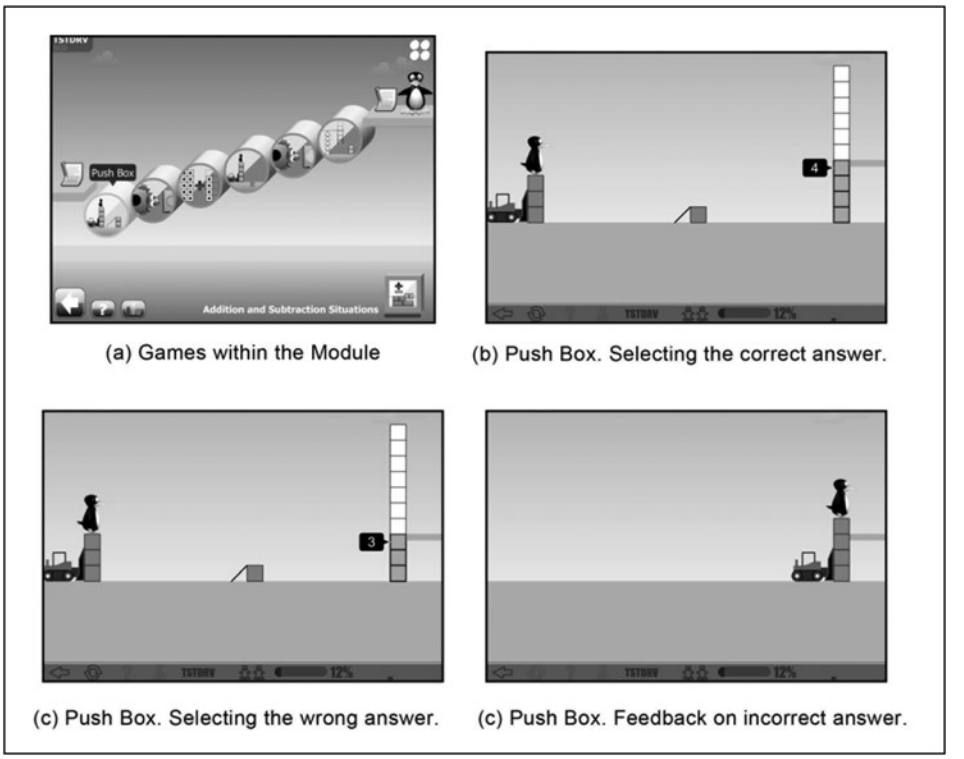

Figure 1. The first game in the second-grade curriculum, Push Box. *Note.* Students must add the box on the floor with those on the bulldozer and place the bridge on the appropriate square to allow Jiji to cross.

bulldozer on top of the boxes on the left, to cross the screen. Figure 1b illustrates the correct placement of the bridge, at four blocks, and Figure 1c illustrates an error, placing the bridge at three blocks. Once students choose a placement, the result of their choice is animated. In the case of Figure 1d, the bridge is placed below the spot needed for Jiji to walk across and the student will see that Jiji is stuck on the screen. In response to this incorrect answer, Jiji will give a puzzled look to the student, and either a new problem will start (if the child has not missed any within the current level to this point) or the child will fail the level and will be directed back to a screen to restart the level.

Push Box is the first game in the second-grade curriculum. Once students complete this game, they move on to the rest of the games in the Addition and Subtraction module and then on to the next modules, which for second grade are (in order) Place Value up to 1,000; Money; Time and Elapsed Time; Addition and Subtraction, two digits; Geometry and Measurement; Fractions; Intro to Multiplication; Intro to Division; and Addition and Subtraction, three digits. There are 10 second-grade modules, 13 third-grade modules, 12 fourth-grade modules, and 13 fifth-grade modules. Each game (within the modules) covers certain California State Standards for mathematics for the grade covered. For example, program developers relate Push Box to the Algebra and Functions standard 1.2, "Relate problem situations to . . . addition and subtraction," and note it as tangentially related to Number Sense standard 2.2., "Find the sum or difference of two whole numbers up to three digits long" (MIND Research Institute, 2007). Each game in each grade level is

similarly linked with a grade-appropriate California standard (see California State Board of Education, 2010a). The content and progression of ST Math was originally designed for California standards but has been modified somewhat for administration in states with different mathematics standards.

For the version of ST Math evaluated in this article, Generation 3, students within the same grade all began the year on the same game within the software and proceeded through the games as they solved them. Although students may have begun at the same place, variation in individual rates of progress means that, over time, students mastered different lessons within the software: different levels, different games, and even different modules. Although MIND occasionally allows for placement below grade level for special education students, non–special education students in the current study all received the software for their grade level.

MIND's Theory of Change and Benefit to Certain Student Populations

The MIND theory of change for ST Math involves coordination between teachers and software to first help students develop the ability to visualize underlying mathematics concepts, and then create links between these concepts and the types of problems students encounter in their math classroom or on a standardized test. The developers hypothesize that by learning the meaning behind algorithmic procedures through intuitive spatial relationships, students gain conceptual understanding along with procedural and computational skills—a learning process that may ultimately lead to increased mathematics competency and retention (National Research Council [NRC], 2005; Shaw & Peterson, 2000). These ideas find some support in research by educators and cognitive scientists. Geary (1995) described a biologically primary system of mathematical understanding upon which humans and other animals relied. Opfer and Siegler (2012) similarly described an implicit ability to process nonsymbolic representations of quantity. This implicit ability is related to later mathematics performance (e.g., Booth & Siegler, 2008; Opfer & Siegler, 2012) and can be improved through training, such as with number lines (e.g., Ramani & Siegler, 2008). Drawing on the relation between spatial representations and this implicit system (see Geary, 1995; see also Geary, 2011, noting the unique relationship between spatial skills and mathematics) may be an especially promising way to enhance students' sense for numerical magnitudes and aid in creating links between symbolic and nonsymbolic representations. Research exploring how students learn the mathematics of fractions has shown that visual representations can improve conceptual understanding of both the magnitude and manipulation of fractions (see Siegler et al., 2010; Siegler, Thompson, & Schneider, 2011). MIND has designed the system of visual representations within ST Math to support understanding of number magnitude and relations across the spectrum of numbers and as they increase in difficulty. Repetition across grade levels of certain spatial representations, such as the number line, may support students' analogous transitions to more complex mathematics, like problems involving fractions (see Siegler et al., 2011; Wu, 2005).

Mathematics educators have also noted that exposure to, and manipulation of, multiple representations of mathematics problems may enhance conceptual understanding (see NRC, 2001). Although the push for instruction that fosters greater conceptual understanding of mathematics is not new, instructional practices in the United States are still largely focused on learning procedures (NRC, 2005).[1] Students who are not taught the concepts behind the

[1]This may be changing with adoption and implementation of the Common Core Standards, which stress conceptual as well as procedural learning.

procedures generally have difficulty transferring procedures appropriately to new problems or identifying their errors (see NRC, 2001, 2005). Inclusive in this notion is the idea that understanding why an answer is wrong can help foster conceptual understanding (NRC, 2005). In line with these findings, ST Math is designed to provide animated representations of students' incorrect solutions. However, all students who chose the same answer within the games are shown the same animated representation of the result of this choice—this may work for some students but not for all. Hence knowledgeable teachers, actively monitoring student progress are considered an integral part and potential weakness of the CAI. Teachers must identify those students who are "stuck" and provide assistance and instruction so that they can both progress in the game and understand the concept at issue (Peterson & Patera, 2006).

English Language Learners. By drawing on innate spatial-temporal ability (Shaw & Peterson, 2000), ST Math is designed to provide access to the ST Math lessons for those students who struggle in traditional language-heavy learning environments. ELLs consistently perform below fluent English speakers in standardized tests of mathematics, and although this gap between ELLs and non-ELLs may be slowly closing, differences remain (Hemphill & Vanneman, 2011). The prospect of closing these achievement gaps may depend on the identification of curricula and instructional practices that particularly meet the needs of ELL students. Access to standard mathematics curricula requires considerable amounts of verbal or written communication—ELL students may not have the academic vocabulary necessary to make sense of traditional math lessons (Hoffert, 2009). In addition, because of the slow rate of instruction or dedication of class time to language learning, ELLs often do not have the opportunity to learn a full year's complement of math material during a regular academic year (Abedi & Herman, 2010). For these students, ST Math may meet the needs of both access and breadth of material: Language-minimal concept instruction is designed to provide access to the curriculum, and self-pacing allows students to progress individually through a year's material. Although ST Math is designed so that ELLs may master mathematical concepts without simultaneously having to master English-related peculiarities of mathematics learning, it does not teach math by excluding all language—students are gradually introduced to mathematical symbols and language after the initial language-free introduction to concepts. The designers hypothesize that this allows for the integration of language and content that Harper and de Jong (2004) claimed is necessary for ELL student success. By providing dynamic representations accompanied by visual and image-based instruction in their use, ST Math was hypothesized to support stronger math outcomes for ELL students compared to students who are proficient in English.

Student Initial Mathematics Proficiency. The hierarchical nature of mathematics implies that an understanding of foundational mathematical concepts may become critical as students advance toward higher mathematics, a proposition supported by the strong relation between early mathematics skills and later achievement (e.g., Duncan et al., 2007; Ramani & Siegler, 2011). The concept of *developmental progression*, as defined in mathematics learning, lends further support—developmental progression "describes a typical path children follow in developing understanding and skill about [a] mathematical topic" (Clements & Sarama, 2009, p. 3). Within this progression, new mathematics skills are built on previously mastered skills to form a trajectory of increasingly sophisticated thinking. The most effective instruction occurs when instructional tasks are matched to the student's skill level (Clements & Sarama, 2009). Such a system optimizes the learner's readiness, but without it, gaps may appear in student knowledge and lead to an unstable foundation

for future mathematics endeavors (NRC, 2009; Thurston, 1990). ST Math's individualized curriculum is designed to allow students to progress to new concepts only once they have mastered the foundational materials. The designers hypothesize that students who struggle with mathematics may make greater gains in mathematics outcomes than those in more typical classrooms because they can take time to review basic concepts at their own pace.

The idea that CAI may be effective in improving mathematics outcomes for low-performing students has been present in the literature since the initial uses of the medium (e.g., Edwards, Norton, Taylor, Weiss, & Dusseldorp, 1975). This idea persists, particularly in special education research (e.g., Li & Edmonds, 2005; Traynor, 2003), although, as noted previously, large-scale random assignment studies of ability-differential program effects are rare. We explore these ideas by investigating whether the effect of ST Math differs between students who begin the study at five different math proficiency levels defined by cut-points set by the state. It was our hypothesis that students in the three categories below the "proficient" cut-point would make greater gains in mathematics achievement than would their peers in the control classrooms and that the overall effect sizes would be larger than those of their peers in the top two proficient categories of the ST Math condition.

Effect of More Than 1 Year. New instructional technology, like ST Math, may initially boost student achievement due to engagement and motivational benefits from the novelty of the intervention (e.g., Song & Keller, 2001; Tung & Deng, 2006). Should this be the case, we might expect the effect of ST Math to level off or fade away after the 1st year of implementation. On the other hand, because implementation of ST Math requires specific teacher actions, including progress monitoring, intervention, and integration, 2nd-year program effects may be larger than those from the 1st year of implementation. Similarly, students who have experience with ST Math for more than 1 year may themselves become more adept at learning from this medium and translating what they learn to assessments outside of the software.

Alignment With Outcome Measures

ST Math was designed to align with California State Standards; therefore, in this analysis the impact of ST Math is assessed with the California Standards Tests (CSTs), a standardized test series developed to evaluate the competency of California's students with respect to these same standards. In addition to serving as a measure of ST Math's effect on student math performance, the CSTs have important policy implications for students and schools alike. For students, the mathematics CSTs measure a number of skills deemed critical for success in a complex, technical society—these skills are highlighted by California policymakers and the National Research Council (California State Board of Education, 2010a; NRC, 2001). For schools, the CSTs provide markers of students' progress and carry consequences for the schools themselves. Schools whose students perform poorly on the CSTs face the possibility of corrective action, restructuring, public scrutiny, and the loss of students through school choice (No Child Left Behind §1111, 2001).[2]

Previous research has shown correlational associations between ST Math and student achievement (Graziano, Peterson, & Shaw, 1998; Peterson et al., 2004; Martinez et al., 2008). This article extends the evaluation of ST Math to make causal inferences regarding

[2]Recent developments have seen relief from these sanctions for failing districts through waivers issued by the U.S. Secretary of Education (U.S. Department of Education, 2013).

the 1- and 2-year effects of ST Math on student achievement as measured by the California Standards Test. In addition, this article examines whether ST Math is associated with greater improvement in test scores among ELLs and students with weaker incoming mathematics skills—two groups of students in need of an intervention to increase their trajectory—both to improve their mathematics skills and to assist their schools in reaching policy-relevant proficiency benchmarks.

To summarize, this article addresses the following four research questions: (a) Does ST Math produce gains in CST scores on average for all students? (b) Does ST Math produce differential gains for students who enter the study at different proficiency levels? (c) Does ST Math produce greater gains for ELLs? (d) Does ST Math produce stronger effects for students after they have participated in the program for 1 year?

METHOD

Design and Procedure

The current RCT study used random assignment at the school level. The 52 elementary schools in the study included two cohorts with a staggered implementation design. The first cohort of study schools was drawn from schools selected to participate in MIND's Orange County Math Initiative. This countywide initiative, supported by local business partners and the Orange County Department of Education, provided ST Math without cost to low-performing schools. To determine eligibility, every school in Orange County was ranked by its Academic Performance Index, which is a weighted composite of student scores on state-mandated standardized tests. Schools that fell into the lowest three deciles (155 elementary schools) were invited to participate in the Orange County Math Initiative. After attending an informational session, 73 of the qualifying schools applied to participate, and following a site and eligibility audit, 71 schools were accepted into the Initiative. A subset of 41 schools was eligible to participate in the current study because they were not already users of ST Math. Of these 41 schools, 34 agreed to participate in the RCT.

After the recruitment of the Cohort 1 schools, a partnership involving MIND, the Orange County Department of Education, and researchers from the University of California, Irvine, obtained an Institute of Education Sciences grant to support the implementation and evaluation of ST Math within this sample. This article is a product of the evaluation conducted by the latter two, without oversight from MIND. After receipt of the grant, a second informational session was held for recruitment of an additional cohort of students. Using the same eligibility criteria as for the original sample, eighteen schools were eligible to be part of this new study cohort—all 18 agreed to participate in the RCT.

Randomization. Prior to the fall of 2008, the original 41 schools in Cohort 1 were randomly assigned to one of two conditions: 21 schools were assigned to implement ST Math at Grades 2 and 3 and not in Grades 4 and 5 (Group A), and 20 schools were assigned to implement ST Math at Grades 4 and 5 and not in Grades 2 and 3 (Group B). Although within schools the grades were split between treatment and control, the randomization occurred at the school level to either a second/third-grade implementation or a fourth/fifth-grade implementation. Thus, Grades 2 and 3 of Group B served as controls for the treated Grades 2 and 3 of Group A in the treatment year, and Grades 4 and 5 of Group A served as controls for the treated Grades 4 and 5 of Group B in the treatment year. The decision was made to assign all of a school's classrooms in a given grade as a group to either treatment or control to encourage

Table 1. Comparison of sample descriptives to county and state

	Analysis Sample *M* / %	Total Sample *M* / %	County *M* / %	California *M* / %
Math CST	351.86	351.56[a]	385.59	372.48
ELA CST	329.47	329.08[a]	356.23	346.48
Male	51%	51%	50%	49%
Free/Reduced lunch	90%	88%	46%	57%
Black	2%	2%	2%	8%
Hispanic	85%	84%	47%	50%
White	5%	6%	31%	26%
Vietnamese	4%	4%	6%	1%
Other race	4%	5%	31%	15%
English Language Learner	63%	63%	39%	32%
N	13,803	16,315	110,402	1,401,811

Note. County and California data aggregated for Grades 2 through 4 in 2007–2008 from the California STAR reporting website: http://star.cde.ca.gov/star2008 (California Department of Education, 2011b). Means and percentages from the study sample reflect data at baseline, which is 2007–2008 for Cohort 1 and 2008–2009 for Cohort 2. Demographics of all students in Grades 2 through 4 in the study schools from 2007–2008 to 2008–2009 are relatively stable. If all students were measured in 2007–2008, mean Math and ELA scores would each be 2 points lower due to statewide trends.
[a]Test scores provided for the total sample are limited to those students who had valid California Standards Tests (CST) data: math ($N = 13{,}905$) and English/Language Arts (ELA; $N = 13{,}963$).

fidelity to condition. Before receiving their assignment, seven schools excluded themselves from the study and did not sign the randomization agreement. The resulting Cohort 1 who implemented ST Math beginning with the fall of 2008 consisted of 18 Group A (Grades 2/3) and 16 Group B (Grades 4/5) schools. Cohort 2 schools were randomly assigned to a condition and began ST Math implementation at Grades 2 and 3 or Grades 4 and 5 (nine schools in each condition) at the start of the 2009–2010 school year. No schools in Cohort 2 withdrew from the study.

Sample and Participants

The study sample consisted of all second- through fifth-grade students in 52 low-performing schools within 10 districts in Southern California. Schools ranged in size and enrolled between 200 and 800 students in the study grades during a given year throughout the 3 years included in this analysis. Analyses in this article employ data from the 13,803 students who took the mathematics and English/Language Arts CSTs for their school's first study year and who had pretest data available for the immediately prior year. This represents 84% of the participating students. Descriptive statistics for the study sample are provided in Table 1 and show that the study sample is generally very similar to the total sample. There were no significant differences in the cohorts with regard to school size or initial CST scores.

Because our analysis required a pretest score, and CSTs are only offered to students beginning in second grade, the analysis sample was limited to those students who were in second through fourth grades during their pretest year. This excluded Cohort 1 students who

began second grade in 2008–2009 and students from either cohort who began second grade in 2009–2010. As measured during the first study year, 16,315 students were in the grades targeted for analysis. However, 2,091 were new to the districts for their school's 1st year of implementation, so did not have pretest information. These students were distributed evenly between treatment and control. However, students who moved were more likely to have a posttest below the proficiency cut-point: 13.63% of those below the proficiency cut-point moved into the study districts as compared with 11.89% of those above proficiency, $\chi^2(1, N = 16{,}221) = 11.01$, $p = .001$. An additional 331 students were missing data or had scores out of range for 1 or both years—these students may have taken an assessment other than the CSTs due to disability or language status. Of the 331 with out of range or missing scores, 256 had a reported diagnosed disability for at least 1 year during the study. An additional 90 students switched between the study cohorts, making their results difficult to interpret; they were excluded.

Table 2 shows descriptive statistics by treatment status for students in the study schools during the first implementation year at their school. Aggregated, the mean score of these students was 351.56, a little more than 1 point above the proficiency benchmark set by the state of California (California State Board of Education, 2010b); individual schools had between 27% and 76% of students who had not met proficiency by this point. Among ELLs in this sample, the mean mathematics CST score, averaged between treatment and control, was 334.17, a significant difference from the mean score of non-ELL, 381.09, $t(13801) = 38.38$, $p < .001$. This translated to a significantly larger number of ELL students among those who were not proficient before the start of the study $\chi^2(1, N = 13{,}904) = 950.43$, $p < .001$. Details about the CSTs and proficiency benchmarks are provided in the upcoming Variables section.

As seen in Table 2, gender, ethnicity, language status, and eligibility for free/reduced lunch did not significantly differ between treatment and control students. Starting mathematics CST scores differed slightly between treatment and control groups, $t(13801) = 3.35$, $p = .001$, with treatment students scoring, on average, 4 points higher than control students. Pretest scores for treatment and control were roughly normally distributed: skewness for both groups was similar and below .50; kurtosis was within .30 of three.

The 2-year gains for treatment and control students were investigated using the only randomly assigned students with 2 years of data: Cohort 1 students who began the study in third grade during the 2007–2008 school year. Within this subsample, there were some statistically significant differences in demographics and baseline CST scores. The treatment students included more Vietnamese students (7% as compared to 3%), $\chi^2(1, N = 2{,}676) = 23.13$, $p < .001$; fewer Hispanic students (81% as compared to 85%), $\chi^2(1, N = 2{,}676) = 8.59$, $p = .003$; and more male students (52% as compared to 48%), $\chi^2(1, N = 2{,}676) = 5.67$, $p = .02$. Treatment students in this subsample start with mathematics CST scores 12.36 points higher than control students, $t(2675) = 4.13$, $p < .001$, and English/Language Arts (ELA) CST scores 4.45 points higher, $t(2675) = 2.11$, $p = .04$.

Implementation. MIND liaisons worked with school and district administrators to set up ST Math for each school within currently existing computer labs. Students began attending ST Math lab sessions twice a week for 45 min each session at the beginning of the school year. Based on previous trials of the software, MIND determined that this was the frequency and duration that was both practical for schools and would allow students to complete the majority of the program by the end of the school year. A 4- to 5-hr professional development training session on how to use the software was provided to all study teachers. Study schools were also given technical support for the 1st year of implementation and for additional years

Table 2. One-year differences in math California Standards Tests scores between treatment and control, divided by grade, cohort, and subgroup

		Control					Treatment				
		Pre		Post			Pre		Post		
Grade	Cohort	*M*	*SD*	*M*	*SD*	Count	*M*	*SD*	*M*	*SD*	Count
All	Pooled	349.80[a]	71.75	359.62	77.55	6,966	353.96	74.06	368.03	78.90	6,837
All	1	347.29	71.01	358.62	76.34	4,822	349.10	72.95	364.49	78.47	4,660
All	2	355.42[a]	73.09	361.87	80.17	2,144	364.35	75.36	375.60	79.30	2,177
2nd	1	266.40	49.72	349.79	55.64	52	265.71	48.61	362.71	73.24	28
2nd	2	304.29	87.06	334.14	93.87	7	322.20	71.54	356.20	42.14	5
3rd	1	356.39[a]	74.70	372.20	80.38	1,442	345.43	75.68	362.85	80.44	1,756
3rd	2	355.66	76.03	366.51	82.28	789	348.87	70.94	370.95	81.39	632
4th	1	342.48[a]	72.81	357.36	70.79	1,659	351.73	76.08	366.57	71.00	1,386
4th	2	354.41[a]	73.74	362.90	70.60	641	368.91	76.47	388.26	75.79	797
5th	1	346.75[a]	64.10	348.42	76.93	1,669	352.55	65.66	364.54	82.76	1,490
5th	2	357.06[a]	68.60	356.49	85.37	698	374.66	75.83	367.39	80.05	694
		Subgroup Differences, Pooled Grades & Cohorts									
Group											
Male		351.23[a]	74.75	360.31	81.05	3,545	355.32	76.89	370.48	81.30	3,546
Female		348.31[a]	68.48	358.91	73.74	3,421	352.48	70.87	365.39	76.16	3,291
Reduced lunch		345.67[a]	69.74	355.65	75.66	5,871	348.70	71.56	362.83	76.38	5,721
Black		351.83	64.07	347.26	78.77	106[b]	359.00	76.26	377.53	78.77	107
Hispanic		343.52[a]	68.72	353.28	73.86	5,979[b]	346.20	69.86	359.95	74.30	5,771
White		379.14	76.28	382.40	79.40	344	386.04	81.40	395.17	86.85	372
Vietnamese		417.43	74.69	439.81	80.64	233[b]	416.13	71.06	439.78	77.78	279
Other race		387.40[a]	78.68	401.35	90.63	304	402.43	87.02	418.31	92.30	308
English Language Learner		332.99[a]	66.08	345.10	71.48	4,375	335.91	66.58	352.00	72.61	4,298

Note. Pretest is 2007–2008 for Cohort 1 and 2008–2009 for Cohort 2. Grade listed is posttest grade.

[a]Indicates pretest score is statistically significantly different ($p < .05$) between treatment and control students within that group. [b]Indicates number of students in listed subgroup is statistically significantly different ($p < .05$) between treatment and control students, pooling the grades and cohorts.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

by paying the lesser of a $3,500 or $35 per student renewal fee to MIND. All study schools continued to pay this fee and receive support for the duration of the study.

Each year in the study, schools had the option to add two treatment grades to provide multiple years of treatment to students as they progressed through the elementary grades. As a consequence, a third-grade student who was assigned to a Group A (Grades 2/3) school did not stop receiving ST Math in fourth grade so long as their school exercised the option to add additional grade levels. To date, only one school has elected to not add grades during their subsequent years in the study. The delayed treatment design utilized within this study permitted variation in the number of years and grade levels of those assigned to treatment, and supported equal engagement in the study by treatment and control teachers (e.g., Roschelle et al., 2010)—initial control teachers knew that their grade would receive the intervention within 2 years.

Because ST Math is a supplemental program, treatment students may have received an additional 90 min a week of mathematics instruction compared to control students. A survey of treatment teachers within the study schools (2011–2012) indicated the teachers used time from a variety of subjects in order to implement ST Math. When asked where the instructional time for ST Math came from, 34% of teachers reported math, 17% reported English/Language Arts, 36% reported Social Studies/Science, 9% reported Art/Music/PE, and 4% reported other computer lab time. Thus, treated students received approximately 90 additional min per week of (ST) mathematics instruction, minus the time taken from classroom mathematics instruction to attend the ST Math computer lab. It is not known how much time control group students spent in mathematics instruction compared to treatment students, but it is assumed that they spent less time in mathematics instruction.

Fidelity of Implementation. A key condition for a successful RCT is that the intervention is implemented as intended with reasonable fidelity. Otherwise, it is not clear what is being tested (see WWC guidelines). For full implementation, ST Math requires that students complete all of the software modules for their grade and that their teachers refer to and draw on student experiences with the software during classroom mathematics instruction (see Peterson & Patera, 2006). ST Math students within the current study were expected to spend two 45-min sessions each week in the computer lab for an average total of 68 sessions per year. Teachers appeared to take their classrooms to the lab as scheduled: On average, students utilized the ST Math software for 68 days during the 1st year of the study implementation and 66 days during the 2nd year as indicated by game-play data. Due to the self-paced nature of the program, utilizing ST Math the maximum number of days within a year did not necessarily mean the student completed all of the grade lessons. On average, students completed 80.97% of their grade's lessons by the end of the school year and 72.77% of the lessons by April, when the CSTs were administered. Percentage of program completion by end of year varied as a function of student initial proficiency level. The lowest performers at pretest, those who were "far below basic," on average completed only 47.82% of the program. The highest performers, those who were "advanced," averaged a completion rate of 90.36%. Those in the middle three proficiency categories, "below basic," "basic," and "proficient," averaged completion rates of 58.64%, 69.21%, and 80.54% of the program, respectively.

Observations of teacher fidelity were conducted during the 2009–10 school year. Eight observers (retired teachers or school administrators) were trained by MIND staff on a protocol developed by the Center for Elementary Mathematics and Science Education at the University of Chicago and employed this protocol to observe 102 treatment and 90 control mathematics classrooms once during class time and once during lab time (for the

treatment classes). This represents 25% of the 806 teachers in our study grades and schools in the 2009–10 school year; one teacher was randomly selected from each grade in each school for a total of 208 teachers. Six teachers opted out of the observation. The protocol aimed to capture teacher fidelity to MIND's view of what the core elements of the program were, including the use of visualizations, specific teacher questioning practices, and the drawing of connections between classroom and in-game experiences. Initial results suggest limited teacher fidelity; of the observed classrooms, only 38% of treatment teachers were mentioned the software or Jiji at all during nonlab mathematics time. Only 21% of teachers were observed to draw connections between the games and what the class was learning. Interpretation of the results should keep in mind the relatively high student attendance rate in ST Math labs and the low teacher compliance with the need to integrate classroom instruction with what is learned in the lab. It is likely that teachers implemented the classroom aspects of the intervention that would be typical in most districts.

Variables

Standardized Test Scores. Scores from the CST, administered to all California students in Grades 2 to 11 in the spring of each year. Scores were used to assess mastery of grade-level mathematics content standards. CSTs are criterion-referenced, standards-based assessments developed in alignment with the California Content Standards (California Department of Education, 2010a). For the 2007–2008 test administration, the latest year for which this information is available, Cronbach's alphas in Grade 2 and 3 CST mathematics were 0.93 and 0.94, respectively (Educational Testing Service, 2008). Scale scores ranging from 150 to 600 were calculated by the state to allow for comparison between grade levels and were provided to the researchers by the participating school districts. These scale scores are necessary because tests are designed to assess each grade's standards and therefore differ between grades; within grades, each year's test is based on the same core of standards but contains different questions. Each student in the current study had data on 1 to 3 years of CST scores (2008, 2009, 2010), depending on the grade level of the student. Across mathematics and English in all elementary grades, a scale score of 350 points indicates a student is considered by the state to be proficient in that subject's content-matter for that grade. In addition to specifying the 350-point proficiency cutoff, the state of California has designated math cutoff points for far below basic (scores less than approximately 240, depending on grade level), below basic (below 300), basic (300–350), and advanced (above 400, with the exact value depending on grade level; California Department of Education, 2010b).

School Math Curriculum. There is no specific mathematics curriculum that MIND recommends as optimal for ST Math administration. In 40 of the 52 study schools, the concurrent curriculum was Houghton Mifflin CA Math. Within a school, the same curricular provider was used across all study grades. No curriculum was disproportionately represented within a grade and treatment condition. For the current analysis, ST Math was evaluated after 1 year of program use, comparing students with 1 year versus no years, and after 2 years of program use, comparing students with 2 years versus no years.

The current study investigated the effect of ST Math with an intent to treat analysis in order to make a policy-relevant conclusion and preserve the integrity of the random assignment design (Shadish & Cook, 2009). Students and teachers in the study, as in the real world, cannot be forced to participate fully in the ST Math program. Thus, intent to

treat effect estimates more realistically capture the average program effect on the full set of students who were offered the treatment, regardless of their actual level of ST Math involvement.

Demographics. Gender, ethnicity, free/reduced lunch, and ELL status were reported by the school districts along with the CST data. Ethnicity was represented in the analysis by five groups: Hispanic, Vietnamese, Black, White, and Other, to represent the largest ethnic groups within the sample. ELL status was determined by schools as measured by the California English Language Development Test (California Department of Education, 2011a). For purposes of analysis, students were classified as ELL students if they were listed as English Language Learners in 2008–2009; those students who were Redesignated Fluent English Proficient were not labeled as ELL.

ANALYSIS

To answer the first research question (estimating the average main effect of the ST Math treatment), 1st-year posttest results were regressed on treatment, pretest scores (both mathematics and English/Language Arts), student grade, year, and demographic controls. To answer the second and third research questions (differential effects for population subgroups), this analysis was then performed separately for each proficiency level and for ELL versus non-ELL students. Coefficient differences between the groups were tested for significance. To examine the fourth question—the effect of 2 versus 1 year of ST Math—the sample was limited to those students who were in third grade in Cohort 1 during the 2007–2008 school year. For these analyses, each student contributed 2 years of outcome data (fourth and fifth grade). Students who received treatment for 2 years were compared to their same-grade and cohort 2-year-control counterparts. Effects were estimated for both the 1st and 2nd year by including a treatment by 2nd-year interaction within the equation. Although it was hypothesized that the effect of 2 years of ST Math may be stronger than 1 year alone due to teacher practices, we were unable to isolate teacher experience with ST Math. The 2nd year of ST Math for the students in our analysis was also the 2nd year of ST Math implementation within their schools. However, we were not able to eliminate the possibility that the students were taught by a teacher new to ST Math.

In evaluating education interventions, researchers must be concerned that characteristics propelling adoption of interventions are correlated with other determinants of the outcomes, leading to bias in the estimate of the intervention's effect. Given a sufficiently large sample, random assignment generally allows us to assume that student characteristics are evenly distributed between treatment and control groups, so that unbiased treatment effects can be estimated (Shadish et al., 2002). However, two issues within the current study potentially interfere with this assumption.

As noted earlier, pretest scores were not completely balanced between treatment and control group students. To control for this, mathematics and ELA pretest covariates were added to the regressions. Although other measured characteristics were balanced between treatment and control (see Table 2), it is worthwhile to increase the statistical power of our analyses by adding controls for student characteristics such as grade, year at first implementation, and demographics. These are included in the estimated models.

The study schools were chosen to have relatively similar demographic and baseline achievement, but the schools do differ both in their mean pretest scores (the range is 323.43 to 392.22, with most falling between 323.4 and 360) and the deviation of student

scores around the mean (range of the standard deviation is 50.86 to 83.81). The intraclass correlation (ICC) of .04, although below the typical nationwide school ICC of .22 (Hedges & Hedberg, 2007),[3] indicates some degree of similarity between students within schools. Considering the study's ICC, with an average school size of 419, a design effect for the current study can be calculated based on the formula provided in McCoach and Adelson (2010). The design effect of 17.72 indicates a sampling variability greater than that which would be expected from a simple random sample (McCoach & Adelson, 2010). Within this study, it implies that nesting of students within schools should be considered to more accurately estimate the significance of any treatment effect. To deal with this nesting, we report Huber-White standard errors clustered on school. These conservative adjusted standard errors were relied upon for determinations of statistical significance and were an increase over unadjusted standard errors by a factor of 2.5 to 3.5, depending on the sample.

RESULTS

On average, treatment students within the combined sample gained 14.07 points from pre- to posttest, which was 4.25 points more than gained by control students (see the top row of Table 2). This was .06 of the control group pretest standard deviation, an effect size that, as we shall see, is similar to the upcoming regression results. Table 3 presents results from the regression of mathematics CST scores on treatment status after 1 year. Columns 1 and 2 present the results by cohort, and columns 3 through 6 pool both cohorts together and present total results as well as results separately by student grade level at posttest (excluding the students who were retained and thus were in second grade twice). The overall, regression-adjusted main effect of ST Math as seen in the pooled sample (column 3) is 5.12 points, and this coefficient is marginally statistically significant ($p = .089$). Using the standard deviation of the entire sample pretest (72.93), we calculate an effect size of .07. This is negligible according to Cohen (1988) but on par with the .07 effect size found in other random control trials of elementary school interventions tested with a broad standardized test such as the mathematics CSTs analyzed here and is 14% of the average annual mathematics learning gains for a fourth grader (Hill, Bloom, Black, & Lipsey, 2008). In general, the estimated effects of ST Math were similar across the different samples in Table 3, with effect sizes ranging from .05 to .10 and with the estimate from the pooled sample falling near the middle of this range. None of these separate estimates was statistically significant. Note that standard errors were estimated using the Huber-White correction for clustering of the sample in schools, which produces larger standard errors than those computed from ordinary least squares regression.

To explore the possibility of effect moderators, we then examined ELL × Intervention and Initial Proficiency × Intervention interactions. Separate models were calculated for each pretest proficiency category and for ELL and non-ELL students (Table 4). Separate models (fully interactive) for each subgroup were chosen over additive models only including interaction terms for each Subgroup × Treatment Status because of the possibility that the mathematics and English pretests, grade, year, and other covariates in our model might have different coefficients for each of the subgroups. The 1-year estimated effect of ST Math was consistently not significantly greater than zero across the subgroups. Thus, there is

[3]This low ICC is not surprising given that the study schools were chosen for their demographic, geographic, and prior performance similarities.

Table 3. Main effect of Spatial-Temporal (ST) Math on math achievement after 1 year

		(1) Cohort 1	(2) Cohort 2	(3) Pooled	(4) 3rd Grade	(5) 4th Grade	(6) 5th Grade
ST Math	B	4.17	7.62	5.12†	3.44	6.63	5.02
	SE	(3.47)	(5.47)	(2.97)	(4.03)	(4.26)	(5.38)
	d	0.06	0.10	0.07†	0.05	0.09	0.07
Pretest Math	B	0.61***	0.60***	0.61***	0.51***	0.60***	0.74***
	SE	(0.01)	(0.02)	(0.01)	(0.03)	(0.02)	(0.03)
Pretest English	B	0.34***	0.34***	0.34***	0.51***	0.23***	0.26***
	SE	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)
Grade 2	B	25.76†	−26.19*	24.66*			
	SE	(12.74)	(9.49)	(10.96)			
Grade 4	B	5.41	8.01	5.82			
	SE	(4.27)	(5.19)	(4.23)			
Grade 5	B	−11.38**	−18.23*	−10.75**			
	SE	(4.06)	(6.57)	(4.01)			
Cohort 2	B			1.47	1.43	3.76	−8.27
	SE			(4.26)	(4.26)	(4.78)	(5.56)
Cohort 2 × GR 2	B			−48.12***			
	SE			(13.04)			
Cohort 2 × GR 4	B			1.63			
	SE			(7.07)			
Cohort 2 × GR5	B			−8.52			
	SE			(7.75)			
Constant	B	32.20***	34.61***	36.78***	12.80	83.04***	5.19
	SE	(6.39)	(8.56)	(5.39)	(8.89)	(6.63)	(11.45)
N		9,482	4,321	13,803	4,619	4,483	4,551
R^2		0.596	0.577	0.591	0.608	0.592	0.589

Note. Unstandardized regression coefficients, standard errors in parentheses. Standard errors have been corrected for nesting by clustering on school, resulting in 52 clusters. Years of analysis for Cohort 1 are 2007–2008 (pretest) and 2008–2009 (posttest), for Cohort 2 are 2008–2009 (pretest) and 2009–2010 (posttest). Within regressions (1) through (3), third grade serves as the reference grade and Hispanic as ethnic reference group. For pooled sample, Cohort 1 is the reference cohort. Control variables included in analysis but omitted from table are English Language Learner, Male, whether student failed/repeated their grade during the pretest year, National Free/Reduced Lunch program and ethnic dummy variables. Expanded tables on file with authors. Individual analysis (columns 4–6) for second graders ($N = 150$) excluded. Second graders included in analyses (1) through (3) are those who failed second grade and therefore have a pretest score.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

little support for the hypothesis that ST Math has stronger effects for the lowest performers and/or for ELLs.

Table 5 presents results from the 2-year analysis of ST Math using students who began the study in third grade in 2007–2008 and who, after 2 years, are in fifth grade. Comparison-group students had ST Math in neither year. Column 1 of this table shows that the estimated effect for this student subsample in fourth grade (after 1 year of the program) is not significantly different from zero. This is lower than estimated for the full sample, suggesting that the 2-year sample results may be affected by differences in sample composition between the 1- and 2-year samples.

Table 4. One-year effect of Spatial-Temporal (ST) Math on math achievement by proficiency and English language categories

		(1) Far Below Basic	(2) Below Basic	(3) Basic	(4) Proficient	(5) Advanced	(6) Non ELL	(7) ELL
ST Math	B	3.06	4.71	3.86	5.32	5.29	5.77	4.56
	SE	(4.74)	(2.93)	(3.46)	(3.76)	(4.42)	(3.66)	(3.10)
	d	0.04	0.06	0.05	0.07	0.07	0.08	0.06
Math pretest	B	0.34*	0.63***	0.66***	0.69***	0.43***	0.62***	0.60***
	SE	(0.16)	(0.06)	(0.05)	(0.06)	(0.03)	(0.02)	(0.01)
English pretest	B	0.14†	0.19***	0.27***	0.34***	0.48***	0.36***	0.32***
	SE	(0.08)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.02)
Grade 4	B	4.42	7.16†	8.48	6.43	−2.83	−0.60	8.75*
	SE	(5.60)	(3.79)	(5.28)	(4.56)	(6.71)	(6.51)	(3.73)
Grade 5	B	−9.01	−15.73***	−16.20***	−7.89	0.41	−11.24*	−12.85**
	SE	(8.41)	(4.20)	(4.10)	(5.19)	(6.17)	(4.84)	(4.28)
Cohort 2	B	5.50	0.67	–1.33	–2.03	7.19	0.33	1.84
	SE	(8.82)	(4.10)	(6.82)	(5.14)	(7.49)	(5.04)	(4.44)
Cohort 2 × Grade 4	B	–8.55	0.33	8.44	2.14	1.92	3.21	0.53
	SE	(11.94)	(7.65)	(9.87)	(8.71)	(11.07)	(8.74)	(7.11)
Cohort 2 × Grade 5	B	9.78	1.06	–8.68	–9.57	–22.91†	–11.36	–4.72
	SE	(14.48)	(7.49)	(8.32)	(11.26)	(11.89)	(8.03)	(8.26)
Constant	B	156.61***	75.69***	35.26†	3.92	63.77**	22.97**	47.65***
	SE	(40.46)	(17.01)	(19.00)	(24.55)	(21.17)	(8.43)	(6.12)
N		591	2695	3632	3914	2879	5130	8673
R^2		0.043	0.115	0.142	0.155	0.276	0.592	0.542

Note. Unstandardized regression coefficients, standard errors in parentheses. Standard errors have been corrected for nesting by clustering on school, resulting in 52 clusters. Years of analysis for Cohort 1 are 2007–2008 (pretest) and 2008–2009 (posttest), for Cohort 2 are 2008–2009 (pretest) and 2009–2010 (posttest). Within all regressions, third grade serves as the reference grade, Hispanic as ethnic reference group, and Cohort 1 as the reference cohort. Control variables included in analysis but omitted from table are English Language Learner, male, whether student failed/repeated pretest grade, National Free/Reduced Lunch program and ethnic dummy variables. Expanded tables on file with authors. Second graders excluded from these analyses because of low sample sizes for certain groups. ELL = English Language Learner.

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 5. Two-year effect of Spatial-Temporal (ST) Math on math achievement for Cohort 1 students who start in third grade

		(1) Main Effect	(2) Far Below Basic	(3) Below Basic	(4) Basic	(5) Proficient	(6) Advanced	(7) Non-ELL	(8) ELL
ST Math	B	1.95	−7.65	8.75†	4.07	4.33	−7.18	−1.79	4.65
	SE	(5.16)	(7.35)	(4.96)	(6.11)	(6.74)	(8.14)	(6.87)	(5.19)
	d	0.03	−0.1	0.12	0.06	0.06	−0.1	−0.02	0.06
ST Math × 5th Grade	B	9.59	6.36	4.98	7.61	7.64	15.47*	15.87*	5.38
	SE	(5.99)	(12.21)	(6.74)	(7.97)	(8.34)	(6.66)	(7.34)	(6.42)
	d	0.13	0.09	0.07	0.10	0.10	0.21*	0.22*	0.07
5th grade	B	−0.06	8.68	−2.25	−7.50	−0.84	14.41**	−1.11	0.64
	SE	(3.73)	(8.62)	(3.23)	(4.99)	(5.85)	(5.02)	(4.90)	(4.19)
Math pretest (centered)	B	0.63***	0.49†	0.57***	0.54***	0.55***	0.54***	0.62***	0.63***
	SE	(0.02)	(0.26)	(0.09)	(0.13)	(0.10)	(0.05)	(0.03)	(0.03)
ELA pretest (centered)	B	0.24***	−0.02	0.11	0.19**	0.28***	0.33***	0.30***	0.18***
	SE	(0.03)	(0.29)	(0.07)	(0.06)	(0.05)	(0.07)	(0.04)	(0.05)
Constant	B	367.29***	276.79***	315.34***	330.71***	380.77***	459.24***	397.38***	343.95***
	SE	(7.59)	(15.17)	(12.50)	(11.32)	(12.81)	(14.65)	(8.04)	(13.51)
N		2,677	134	555	758	727	503	1,068	1,609
R^2		0.549	0.047	0.069	0.056	0.097	0.298	0.580	0.448
Fifth-grade score regressed on ST Math, third-grade pretests, and demographic controls									
ST Math	B	10.81	−1.35	14.34*	11.57	12.03	6.53	13.72	9.51
	SE	(7.82)	(12.24)	(6.92)	(10.94)	(10.68)	(9.81)	(8.59)	(8.84)
	d	0.15	−0.02	0.20*	0.16	0.16	0.09	0.19†	0.13
R2		0.523	0.053	0.048	0.046	0.076	0.312	0.576	0.408

Note. Unstandardized regression coefficients, standard errors in parentheses. In top analysis, each student ($N = 2{,}677$) has 2 years of observations for a total of 5,354 observations. Control variables included in analysis but omitted from table are English Language Learner, male, National Free/Reduced Lunch program and ethnic dummy variables. Pretest for both sets of observations per student is the 2007–2008 California Standards Tests (CST), before study implementation at cohort one schools. Continuous variables are group mean centered, dummy variables are centered (–.5, .5). Bottom table is fifth-grade math CST regressed on third-grade pretest with controls as above. In both tables; standard errors are clustered on school (34 schools within this cohort). ELL = English Language Learner.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

After fifth grade (2 years in the program), the 2-year effect for this sample was 11.48 points (adding together the coefficients of ST Math and ST Math times fifth grade, $d = .16$) and was not significantly different from zero ($p = .12$). There was also evidence of an initial Mathematics Skill × Treatment interaction for 2 years of ST Math. The intervention was least effective for students starting the study with far below basic skills on the CST ($d = –.02$) and most effective for students beginning with below basic, basic, and proficient mathematics skills ($d = .20$, .16, and .16, respectively).

The bottom half of Table 5 estimates these results in a different way: regressing fifth-grade score on the treatment variable and third-grade controls. For each column, the numbers are consistently close to the calculations from the sum of ST and ST × Fifth-Grade Coefficients in the top panel of the table. In particular, the 2-year main effect for the total subsample in Table 5 was not significant. Overall, the fact that the 1-year effect for the 2-year sample differs substantially from the 1-year effect estimated with the (largest possible) pooled sample suggests that the estimated 2-year effect should be viewed as no more than exploratory.

ST Math is a supplemental program, and as such requires an additional $1\frac{1}{2}$ hr of (computer lab) mathematics instruction each week. This time must come from somewhere; it is possible that instructional time for other subjects is being compromised to accommodate ST Math (evidence from our 2011–2012 surveys supports this). To test for achievement consequences of this, we analyzed student ELA CSTs scores (pooled sample pretest $M = 329.47$, $SD = 74.49$). After 1 year of ST Math, there is a small and nonsignificant negative association with treatment ($d = –.02$, $p = .204$, table on file with authors). The effect does not vary significantly across proficiency and language categories, nor does it achieve significance for any one group. Some variation is seen in the 2-year effect (Table 6; effect sizes range from –.17 to +.02) calculated by regressing fifth-grade test score performance on third-grade controls as is done with mathematics scores at the bottom of Table 5; however, none of the ST Math coefficients reach significance.

Robustness Check

Although less of a concern because of the random assignment of schools to conditions, unmeasured characteristics of schools may nevertheless bias our results. To control for unmeasured characteristics of schools, we employed a school fixed-effects model as a check of our simple OLS regression results. The fixed-effects estimates were largely consistent with the results reported; coefficients were within 1 point of those reported, and the direction and significance of results were unchanged.

For the analyses just reported, we included students who were second graders at their school's 1st year of implementation and, because they had failed and repeated second grade, therefore have both pre- and posttest results. We conducted sensitivity analyses, and no significant differences were found between models that included and excluded these students.

DISCUSSION

The Main Effect of ST Math

We found that 1 year of ST Math produced very modest ($d = 0.07\ SD$), marginally significant gains in mathematics CST scores among third- through fifth-grade students within 52

Table 6. Two-year effect of Spatial-Temporal (ST) Math on English language arts achievement for Cohort 1 students who start in third grade

		(1) Main Effect	(2) Far Below Basic	(3) Below Basic	(4) Basic	(5) Proficient	(6) Advanced	(7) Non-ELL	(8) ELL
ST Math	B	−3.69	−12.53	1.54	−5.75	−2.04	−5.91	−3.98	−3.53
	SE	(2.49)	(7.81)	(4.31)	(4.01)	(2.99)	(4.10)	(2.77)	(2.84)
	d	−0.05	−0.17	0.02	−0.08	−0.03	−0.08	−0.05	−0.05
Math pretest	B	0.14***	0.35	0.10	−0.02	0.04	0.13***	0.15***	0.13***
	SE	(0.02)	(0.32)	(0.11)	(0.09)	(0.07)	(0.03)	(0.02)	(0.02)
English pretest	B	0.57***	0.45†	0.51***	0.55***	0.64***	0.52***	0.57***	0.56***
	SE	(0.02)	(0.24)	(0.05)	(0.05)	(0.04)	(0.06)	(0.03)	(0.03)
Constant	B	351.34***	260.78***	322.32***	341.75***	361.51***	397.40***	373.85***	339.32***
	SE	(3.15)	(17.25)	(14.20)	(6.37)	(4.39)	(6.49)	(3.75)	(6.04)
N		2,675	134	553	758	727	503	1,068	1,607
R^2		0.586	0.081	0.193	0.274	0.411	0.486	0.603	0.425

Note. Unstandardized regression coefficients, standard errors in parentheses. Estimates fifth grade math California Standards Tests (CST) scores regressed on controls as above with standard errors clustered on schools (34 schools within this cohort). Only students in Cohort 1 who progress from third to fourth to fifth grade from 2007–08 to 2009–10 are included in this analysis. Control variables included in analysis but omitted from table are English Language Learner, male, National Free/Reduced Lunch program and ethnic dummy variables. Pretest is the 2007–2008 CST, before study implementation at cohort one schools. Expanded tables on file with author. Continuous variables are group mean centered, dummy variables are centered (−.5, .5).

***$p < .001$.

low-performing schools in Orange County, California. Estimates of 2-year effects were larger ($d = 0.15$ *SD*) but not statistically significant. Although disappointing, such null findings are important to consider given the cost of the CAI and the supplementary time required. This evidence can be classified as approaching "strong" according to Slavin and Lake (2008), because of the randomized design of the study, the large sample size (both of students and schools), and the use of clustered standard errors to account for the nesting of student observations within schools. Neither effect size reaches Slavin and Lake's (2008) +0.20 effect size threshold for an effect to be important (p. 476). The WWC uses a +.25 effect size threshold for "substantive importance." To put these results in context, the average 1-year effect size from the 10 CAI studies vetted by the WWC is .14, and among the three with enough evidence to be given the label of "potentially positive results for mathematics achievement," the effect sizes range from .04 to .27 (U.S. Department of Education WWC, 2013a). The effect size for ST Math is at the lower end of this range.

How do these results compare to those of rigorous recent studies of the effect sizes of interventions based on innovative educational technology? The two WWC-vetted studies of educational technology with positive effects have effect sizes well beyond ST Math's 1-year effect. Roschelle and colleagues (2007) find an effect of .87 for SimCalc; however, the intervention is in the shortest duration category of interventions as reported by the National Mathematics Advisory Panel (2008a)—the category most likely to have the largest effect. Such short, targeted interventions typically have assessments that are strongly related to the intervention (see Cheung & Slavin, 2013). Testing a yearlong intervention with a standardized statewide end-of-year assessment leaves more room for mis-targeting and diffusion of the effect but may provide more relevant evidence for schools in light of mandated state assessment and accountability systems.

The Barrow, Markman, and Rouse (2009) study's .17 1-year effect size is calculated from a limited sample standard deviation. When the national standard deviation is used, the effect drops to .09 (p. 67). By comparison, the ST Math effect drops from .07 to .06 when the statewide standard deviation for the appropriate grade levels (Educational Testing Service, 2009) is used in the calculation. ST Math's effect is similar to effects within the randomized experiment evaluations of educational technology in Cheung and Slavin (2013), where an average effect of .08 was found. Hence, in general, and including ST Math, it appears that when CAI is added to mathematics instruction, the learning gains are relatively small in magnitude.

Assessment of these results should also consider the costs of the program (Duncan & Magnuson, 2007). As a supplemental intervention, ST Math represents a cost over and above the cost of the resources already available to schools. An attempt to create a cost–benefit calculation runs up against the following considerations: How should the benefit be valued (e.g., per test score point increase, per percentage of students brought over proficiency)? How should the cost be calculated (e.g., with which pricing structure, for a single year or amortized over multiple years)?

In addition, monetary expense is not the only cost of the program. Where did the 90 min per week for the ST Math computer lab come from? The answer varied, with teachers drawing time from ELA, Social Studies or Science, and other areas. Although the treatment effects on ELA scores for student subgroups were not statistically significant, six out of seven of the subgroup coefficients were negative (Table 6). The largest negative effect size (–.17) for those students with ELA scores Far Below Basic suggests the possibility that, as a result of ST Math implementation, these students suffered from less instructional time devoted to ELA.

The Effect of ST Math on ELLs

Standard mathematics curricula require the comprehension and production of academic vocabulary that may be beyond the capacity of ELLs. A key feature of the ST Math curriculum is that, by providing language-minimal concept instruction, it should be particularly accessible to ELLs (Shaw & Peterson, 2000). Program designers hypothesize that ELLs can be gradually introduced to math symbols and language after the initial language-free introduction to concepts.

Yet we did not find the expected larger effect for ELL students within our study. This expectation of an ELL × Treatment interaction assumed that the ST Math instructional model is particularly well aligned with the needs of non-English Learners. Hence the theory of change model is not supported in this regard and may need to be revised. One possibility is that the language-free instruction offered by ST Math is not able to provide the conceptual understanding that ELLs need. Another possibility is that ST Math *does* provide such conceptual understanding but that, given the language challenges faced by ELLs, this understanding is not sufficient to provide larger effects for this group.

The Effect of ST Math Depending on Initial Math Proficiency Status

The developers of ST Math hypothesized that the individualized scaffolding and feedback inherent in ST Math would produce larger effects for low-performing students, similar to a *compensatory* hypothesis (Sameroff & Chandler, 1975). However, we found little support for this in either the effect estimates after 1 year (Table 4) or the 2-year effect estimates in the bottom panel of Table 5. Instead, the magnitudes of the 2-year effects for the five proficiency categories, as seen near the bottom of Table 6, find the smallest effects for the two extreme groups—the lowest and highest categories and significant effects only for students with below basic mathematics skills. Caution should be exercised when interpreting these results, and subgroup differences should be considered exploratory, as small sample sizes led to insufficient power to adequately test these hypotheses (Bloom, 2010).

What might account for the failure to observe stronger effects for the lowest students? One possibility is that, because students began ST Math at a place determined by their grade level, the lowest performing students were not able to access the curriculum—they may have needed below-grade-level content. A second possibility is that the failure of teachers to link program content and classroom curricula reduced the effectiveness of the program. More generally, it may be that the type of mathematics skill improvement and transfer envisioned by the ST Math developers does not occur as easily as they imagine. ST Math resembles a video game, and as such it is engaging for the students. But the ability of such video game activities to teach skills that transfer to the mathematics classroom, and mathematics standardized tests, appears to be lower than expected by program developers. Future research should focus in more sharply on which specific mathematics skills are successfully inculcated by ST Math, and the extent to which each of these translates into increased performance on standardized tests. It may be that such transfer requires explicit reinforcement of these skills by the classroom teacher, immediately after they are learned by ST Math play. Such linkage and reinforcement was absent in the program implementation examined here.

Within the current study, we were unable to disentangle the effect of ST Math from that of extra school time devoted to mathematics instruction, which is a limitation. It appeared that around two thirds of teachers took time away from other subjects to devote

to mathematics instruction via ST Math. To our knowledge, there has not been a random assignment study to determine the effects of additional mathematics instruction on mathematics achievement in elementary school.

IMPLICATIONS FOR ST MATH AND CAI DEVELOPMENT AND RESEARCH

Although this study was conducted to evaluate one CAI mathematics intervention, ST Math contains many features noted as important in designing CAI for underperforming students and ELLs (Freeman & Crawford, 2008; Li & Edmonds, 2005; Seo & Bryant, 2009). The highly scaffolded, visual approach was expected to produce stronger effects for these subgroups than English-speaking and higher performing peers and it did not. Future studies of other similarly featured programs may be informed by our results and set out a priori to investigate comparisons in line with our findings suggesting a pattern of strong effects for middle-performing students. In light of our results, programs serving students who are performing far below or above grade level may be more effective in increasing students' mathematics gains by including content outside of the student's school grade and/or provide a truly adaptive program with different start-points. The active involvement of the teacher might also be required. Some of these adjustments have been included in the latest generation of ST Math software: Teachers are offered more flexibility in determining which games students skip or repeat and the order in which they are introduced.

In addition, ST Math's games were designed to capitalize on the unique relation between spatial skills and mathematics (see Geary, 1995, 2011) in producing mathematics conceptual understanding. This approach did not result in statistically significantly higher standardized math scores than those produced by the business-as-usual instruction received by control students. The standardized test used as the measure of effectiveness included a range of questions on grade-level mathematics but did not include spatial mathematical representations as were used in ST Math. Developers may wish to assess their products with more tightly related measures and/or may wish to consider carefully how their software will translate to more classroom-relevant measures of performance.

U.S. students are struggling to prepare for a world increasingly dependent on mathematics skill and the critical thinking and problem-solving skills that may be supported by mathematics education (National Academies Press, 2010; National Science Foundation, 2010). The students in our study cohort of Orange County, California, schools are no exception, yet only 49% of those in our study schools obtained mathematics scores at or above a proficient level during the year prior to study entry. Traditional mathematics instruction alone was not producing the gains necessary for our study students to succeed in STEM-related advanced study and careers. Although ST Math claims to provide a distinctive approach by emphasizing intuitive spatial representations, a game context for learning, and individualized instruction, the results of this study suggest that these kinds of supplemental interventions may not provide an educationally important impact on students' mathematics learning. As discussed, ST Math requires investments of both money and time for districts. Hence, publication of evaluation results from large-scale real-world supplemental mathematics instructional implementations such as this one can provide a realistic view of the possibilities and limitations of this and other CAI supplemental interventions.

REFERENCES

Abedi, J., & Herman, J. (2010). Assessing English Language Learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record*, *112*, 723–746.

Barrow, L., Markman, L., & Rouse, C. E. (2009). Technology's edge: The educational benefits of Computer-Aided Instruction. *American Economic Journal: Economic Policy*, *1*, 52–74. doi:10.2307/25760027

Bloom, H. S. (2010). *When is the story in the subgroups? Strategies for interpreting and reporting intervention effects for subgroups* (Working Paper). Retrieved from MDRC website: http://www.mdrc.org/publications/551/full.pdf

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development*, *79*, 1016–1031. doi:10.1111/j.1467-8624.2008.01173.x

California Department of Education. (2011a). California English Language Development Test [Test website]. Retrieved from http://www.cde.ca.gov/ta/tg/el/

California Department of Education. (2011b). *Standardized Testing and Reporting (STAR) Results* [Data files]. Retrieved from http://star.cde.ca.gov/

California Department of Education. (2013). *Instructional Materials Price List (IMPL)*. Retrieved from http://www3.cde.ca.gov/impricelist/implsearch.aspx

California State Board of Education. (2010a). *K-12 California's common core content standards for mathematics*. Retrieved from http://www.scoe.net/castandards/agenda/2010/math_ccs_recommendations.pdf

California State Board of Education. (2010b). *Technical Q & A: Percent proficient*. Retrieved from http://www.cde.ca.gov/ta/ac/ap/techqa06b.asp

Cheung, A. C. K., & 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. doi:10.1016/j.edurev.2013.01.001

Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach*. New York, NY: Taylor & Francis. doi:10.1007/0-306-48085-9_3

Cohen, J. (1988).*Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

Duncan, G. J., & Magnuson, K. (2007). Penny wise and effect size foolish. *Child Development Perspectives*, *1*, 46–51. doi:10.1111/j.1750-8606.2007.00009.x

Educational Testing Service. (2008). *California Standards Tests (CSTs) technical report, spring 2007 administration*. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt07.pdf

Educational Testing Service. (2009). *California Standards Tests (CSTs) technical report, spring 2008 administration*. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt08.pdf

Edwards, J., Norton, S., Taylor, S., Weiss, M., & Dusseldorp, R. (1975). How effective is CAI? A review of the research. *Educational Leadership*, *33*, 147–153.

Freeman, B., & Crawford, L. (2008). Creating a middle school mathematics curriculum for English Language Learners. *Remedial and Special Education*, *29*, 9–19. doi:10.1177/0741932507309717

Geary, D. C. (1995). Reflections of evolution and culture in children's cognition. Implications for mathematical development and instruction. *The American Psychologist*, *50*, 24–37.

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology*, *47*, 1539–1552. doi:10.1037/a0025510

Graziano, A. B., Peterson, M., & Shaw, G. L. (1999). Enhanced learning of proportional math through music training and spatial temporal training. *Neurological Research*, *21*, 139–152.

Harper, C., & de Jong, E. (2004). Misconceptions about teaching English-language learners. *Journal of Adolescent & Adult Literacy*, *48*, 152–162. doi:10.1598/JAAL.48.2.6

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. doi:10.3102/0162373707299706

Hemphill, F. C., & Vanneman, A. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment*

of Educational Progress (NCES 2011-459). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177. doi:10.1111/j.1750-8606.2008.00061.x

Hoffert, S. B. (2009). Mathematics: The universal language? *Mathematics Teacher*, *103*, 130–139.

Li, Q., & Edmonds, K. A. (2005). Mathematics and at-risk adult learners: Would technology help? *Journal of Research on Technology in Education*, *38*, 143–166.

Martinez, M. E., Peterson, M., Bodner, M., Coulson, A., Vuong, S., Hu, W., ... Shaw, G. L. (2008). Music training and mathematics achievement: A multiyear iterative project designed to enhance students' learning. In A. E. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* (pp. 396–409). New York, NY: Routledge.

McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly*, *54*, 152–155. doi:10.1177/0016986210363076

MIND Research Institute. (2007). *ST Math scope and sequence, second grade, 2007 California edition*. Santa Ana, CA: Mind Research Institute.

National Academies Press. (2010). *Rising above the gathering storm, revisited: Rapidly approaching Category 5*. Retrieved from http://books.nap.edu/catalog.php?record_id = 12999

National Mathematics Advisory Panel (2008a). *Chapter 6: Report of the task group on instructional practices*. Washington, DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/about/bdscomm/list/mathpanel/report/instructional-practices.pdf

National Mathematics Advisory Panel. (2008b). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

National Research Council. (2001). Adding it up: Helping children learn mathematics. In J. Kilpatrick, J. Swafford, & B. Findell (Eds.), *Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education.* Washington, DC: National Academy Press. Retrieved from http://www.nap.edu/openbook.php?isbn = 0309069955

National Research Council. (2005). How students learn: Mathematics in the classroom. In M.S. Donovan & J. D. Bransford (Eds.), *Committee on how people learn. A targeted report for teachers* (Division of Behavioral and Social Sciences and Education). Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id = 11101

National Research Council. (2009). Mathematics learning in early childhood: Paths toward excellence and equity. In C. T. Cross, T. A. Woods, & H. Schweingruber (Eds.), *Report of the Committee on Early Childhood Mathematics* (Center for Education, Division of Behavioral and Social Sciences and Education). Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id = 12519&page = R1

National Science Foundation. (2010). *Preparing the next generation of STEM innovators: Identifying and developing our nation's human capital*. Retrieved from http://www.nsf.gov/nsb/publications/2010/nsb1033.pdf

No Child Left Behind Act of 2001, Pub. L. No. 107–110, § 115, Stat. 1425 (2002).

Opfer, J. E., & Siegler, R. S. (2012). Development of quantitative thinking. In K. Holyoak & R. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 585–605). Cambridge, UK: Oxford University Press.

Peterson, M. R., Balzarini, D., Bodner, M., Jones, E. G., Phillips, T., Richardson, D., & Shaw, G. L. (2004). Innate spatial-temporal reasoning and the identification of genius. *Neurological Research*, *26*, 2–8). doi:10.1179/016164104773026471

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

Petrosino, A. (2000). Mediators and moderators in the evaluation of programs for children. *Evaluation Review*, *24*, 47–72. doi:10.1177/0193841X0002400102

Ramani, G. B., & Siegler, R. S. (2011). Reducing the gap in numerical knowledge between low- and middle-income preschoolers. *Journal of Applied Developmental Psychology*, *32*, 146–159. doi:10.1016/j.appdev.2011.02.005

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... Gallagher, L. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics. *American Educational Research Journal*, *47*, 833–878. doi:10.3102/0002831210367426

Roschelle, J., Tatar, D., Shechtman, N., Hegedus, S., Hopkins, B., Knudsen, J., & Stroter, A. (2007). *Can a technology-enhanced curriculum improve student learning of important mathematics?* (SimCalc Tech. Rep. 01). Menlo Park, CA: SRI International.

Sameroff, A. J., & Chandler, M. J. (1975). Reproductive risk and the continuum of caretaker casualty. In F. D. Horowitz (Ed.), *Review of child development research* (Vol. 4) (pp. 119–149). Chicago: University of Chicago Press.

Seo, Y.-J., & Bryant, D. P. (2009). Analysis of studies of the effects of computer-assisted instruction on the mathematics performance of students with learning disabilities. *Computers & Education*, *53*, 913–928. doi:10.1016/j.compedu.2009.05.002

Shadish, W. R., & Cook, T. D. (2009). The Renaissance of Field Experimentation in Evaluating Interventions. *Annual Review of Psychology*, *60*, 607–629. doi:10.1146/annurev.psych.60.110707.163544

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

Shaw, G., & Peterson, M. (2000). *Keeping Mozart in mind*. San Diego Calif.: Academic.

Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., ... Wray, J. (2010). *Developing effective fractions instruction for Kindergarten through 8th grade* (NCEE #2010-4039). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/fractions_pg_093010.pdf

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*, 273–296.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*, 427–515. doi:10.3102/0034654308317473

Song, S., & Keller, J. (2001). Effectiveness of motivationally adaptive computer-assisted instruction on the dynamic aspects of motivation. *Educational Technology Research and Development*, *49*(2), 5–22. doi:10.1007/BF02504925

Thurston, W. P. (1990). Mathematical education. *Notices of the American Mathematical Society*, *37*, 844–850. Retrieved from http://arxiv.org/abs/math/0503081

Traynor, P. L. (2003). Effects of computer-assisted-instruction on different learners. *Journal of Instructional Psychology*, *30*, 137.

Tung, F.-W., & Deng, Y.-S. (2006). Designing social presence in e-learning environments: Testing the effect of interactivity on children. *Interactive Learning Environments*, *14*, 251–264. doi:10.1080/10494820600924750

U.S. Department of Education. (2013). *Elementary & secondary education, ESEA flexibility*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. (2013a). *What Works Clearinghouse*. Retrieved from http://ies.ed.gov/ncee/wwc/

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. (2013b). *What Works Clearinghouse procedures and standards handbook (Version 2.1)*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Wu, H. (2005, April). *Key mathematical ideas in Grades 5–8*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Anaheim, CA.