Computers & Education 76 (2014) 215–224

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Alignment of game design features and state mathematics standards: Do results reflect intentions?☆

Katerina Schenke*, Teomara Rutherford, George Farkas

University of California, Irvine, USA

A R T I C L E I N F O

Article history:
Received 13 February 2014
Received in revised form
23 March 2014
Accepted 25 March 2014
Available online 3 April 2014

Keywords:
Mathematics education
Learning technology
Standards-based curricula
Assessment

A B S T R A C T

This paper describes the results of a randomized control trial of a standards-based mathematics software on elementary school students' (3rd–5th graders; $N = 10{,}860$) mathematics achievement. Spatial Temporal Mathematics (ST Math) engages students by presenting them with a series of game-like activities that are directly tied to the California State Standards for mathematics. We report the effects of the program on students' specific mathematics skills, as well as uncover which elements of the design of the games could be responsible for gains in achievement. We pay particular attention to the alignment of design features of the games, the standards to which developers intended to align their games, and assessments of specific mathematics skills. Results indicate a statistically significant effect of the program on students' basic number sense skills as measured by a standardized measure of mathematics achievement (effect size of 0.14). Subsequent coding of the games for elements that are related to number sense, such as the occurrence of number lines and objects that represent numbers, revealed that these design elements occur throughout the software. We discuss these findings as they relate to the development and design of standards-based mathematics curricula as well as how these features relate to their assessment.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Computer-supported learning experiences are hypothesized to increase student achievement (Ke, 2008; Kebritchi, Hirumi, & Bai, 2010; Lopez-Morteo & López, 2007). As promising as these technologies may appear, research in the area of computer-supported mathematics instruction has shown mixed results (Dynarski et al., 2007; Ke, 2008; Kebritchi et al., 2010; Slavin & Lake, 2008; Suppes, Liang, Macken, & Flickinger, 2014). Despite these inconsistent findings, mathematics learning technologies still hold promise if they can be designed, implemented, and evaluated in deliberate ways to meet educational goals appropriate and necessary for student mathematics learning. These goals should inform both the content of the curricula and the focus for evaluations of these programs—designers and researchers alike should consider this alignment between goals, content, and outcomes in the creation, implementation, and study of educational technology. The current paper takes this approach in the exploration of specific goals, implementation, and results from the study of one supplementary mathematics software program, Spatial Temporal Mathematics (ST Math), and examines improvement on specific mathematics proficiencies by elementary school students who received ST Math as compared to students who receive mathematics instruction as usual.

1.1. Mixed results of mathematics learning technology

The oft-cited paper by Dynarski et al. (2007) describes disappointing findings for educational technology as shown from the results of a randomized experiment. In a large-scale study of 51 schools, Dynarski and colleagues randomly assigned teachers to utilize different educational technology products or no technology in sixth grade mathematics or Algebra classrooms. A comparison of achievement on the

☆ This research was supported in part by grants from the Institute of Education Sciences (Grant R305A090527). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0808392.

* Corresponding author. School of Education, University of California, Irvine, 3200 Education, Irvine, CA 92697-5500, USA.
E-mail address: kschenke@uci.edu (K. Schenke).

http://dx.doi.org/10.1016/j.compedu.2014.03.019
0360-1315/© 2014 Elsevier Ltd. All rights reserved.

Stanford Achievement Test (SAT-10) for sixth graders and the End-of-Course Algebra Assessment for those in Algebra revealed no differences between the treatment and control groups. Similarly, in recent meta-analyses, Slavin and Lake (2008) and Cheung and Slavin (2013) found only small or null effects for mathematics technology in raising student achievement. Nevertheless, there remain individual evaluations presenting large positive effect sizes for mathematics software (e.g., Kebritchi et al., 2010; Page, 2002; Roschelle et al., 2010). Some researchers attribute these mixed findings to the design of the particular intervention or curriculum (e.g., Aleven, Stahl, Schworm, Fischer, & Wallace, 2003; Craig et al., 2013), or the inadequate support provided by teachers to scaffold the use of the technology by students (e.g., Wu & Pedersen, 2011). Looking outside the features of the technology and implementation, Cheung and Slavin (2013) note a variety of criteria that can influence effect sizes: they found that the rigor and design of the study, the sample size, and the measure chosen all influenced the effects seen in evaluations of mathematics technology. Their meta-analysis echoed findings by the National Mathematics Advisory Panel that short, targeted interventions often saw the largest effects (NMAP, 2008, chap. 6). These differences may be explained by one feature common to many modern evaluations of educational technology: the use of broad, standardized tests. Although these assessments are closely tied to policy and may provide useful information about some aspects of a program's effectiveness (see NMAP, 2008, chap. 6), they may be too diffuse to adequately assess the true effects of a program. The view that effect sizes from mathematics technology are "disappointing" can be better understood in light of the metric with which they are measured. The typical effect size benchmarks drawn from Cohen (1988) were broadly derived from effects across the social sciences and may not be applicable to modern educational research, especially interventions measured by existing standardized tests (see Hill, Bloom, Black, & Lipsey, 2008). We offer a few recent examples of studies of educational technology to illustrate the differences that may be brought about by the choice of assessment instrument.

Roschelle et al. (2010) developed an intervention targeting specific mathematics concepts for 7th and 8th graders. Whereas the intervention integrated educational technology, curriculum, and teacher professional development, effects were measured with researcher-developed assessments of specifically targeted mathematics skills (rate, proportionality, and linear functions). Using these researcher-created assessments, Roschelle and colleagues found effect sizes ranging from 0.50 to 0.63 on participants' achievement. Findings from this study might give an inflated view of the true effectiveness of the program for the broad umbrella of mathematics performance. In another example of a study using a narrowly defined assessment, Wilson, Dahaene, Dubois, and Favol (2009) implemented an intervention designed to increase student number sense skills. Wilson and colleagues used narrow measures of number sense such as written and verbal numerical comparisons of numbers, non-symbolic comparison tasks, and verbal counting. Their findings indicated that the intervention had a statistically significant effect on only one of these measures. Both studies indicate statistically significant effects on specific skills but do not explicate the effects of educational technology on broader mathematical competencies, therefore potentially limiting their ability to provide policy-relevant information.

At the same time, other studies have used assessments that are too broad to provide evidence of the evaluated program's effectiveness. For example, an evaluation of a 25-week after school program using the Assessment and Learning in Knowledge Spaces[1] (ALEKS) intelligent tutoring was conducted with a group of sixth grade students (Craig et al., 2013). Craig and colleagues used the Tennessee Comprehensive Assessment Program[2] (TCAP) as their assessment for evaluating learning gains of the two groups (a group using ALEKS and a group with teacher-led instruction). They found that students in the ALEKS condition did not have statistically significantly higher scores on the standardized test than those students in the teacher-led condition. One possible explanation for the lack of findings is that the TCAP was too broad to pick up on the change in skills intended to be improved by ALEKS' designers. Programs that fail to show significant positive effects on broad standardized tests may be dismissed too quickly—although there is value in these assessments, the mathematics learning community acknowledges that they very likely fail to assess all important mathematics skills (see Schoenfeld, 2002). Other evidence can be seen in the research by Hill et al. (2008). They found that interventions measured with broad standardized tests were found to have smaller effects than those measured with narrower measures, more tightly tied to the area of the intervention.

We believe that appropriate assessment instruments should be chosen by reference to the developer's goals and intentions for the program. Many developers are mindful of the pressures placed on districts through government mandates such as in the No Child Left Behind Act (NCLB §1111, 2001[3]) and choose to evaluate their program's effects using the state standardized assessments; however, it is unlikely that all educational technology programs are closely aligned to relevant standards, and therefore such measures may fail to capture otherwise meaningful results. Research from outside of educational technology has brought up the importance of aligning mathematics instruction with valid and meaningful assessment (Krajcik, McNeill, & Reiser, 2008; Martone & Sireci, 2009; Schoenfeld, 2006). Schoenfeld (2002) has suggested that assessment is the key piece to improving mathematics education in the United States, because assessment can help drive the development of better mathematics curricula in the classroom. Moreover, current studies outside of educational technology have lauded efforts to align curriculum and instruction (Krajcik et al., 2008; Martone & Sireci, 2009; Webb, 1997). Our paper extends this work to the field of educational technology and focuses on the importance of alignment between goals, curriculum, and assessment. By way of illustration, we detail the exploration of this alignment in the evaluation of one supplementary mathematics software program for elementary school students, ST Math.

1.2. Context for the current study

The developers of ST Math drew on the California Content Standards for Elementary Mathematics in designing curricular units within the program. Based on documentation provided by the developers, we consider progress under these standards as the goals for ST Math and will discuss the standards, their creation, and prior research on other standards-based curricula in the following section.

[1] ALEKS is web-based tutoring system which adapts to learner behavior. ALEKS has designed curriculum for use in K–12 settings as well as in higher education with courses such as "prep for Precalculus" and "Essential math skills for business". More information can be found at www.aleks.com.

[2] TCAP is the statewide assessment given in the state of Tennessee to measure students' skills and progress in Reading, Language Arts, Mathematics, Science, and Social Studies.

[3] Since then, the Obama administration has passed a bill aimed at greatly narrowing the scope of NCLB and making states, not the federal government, accountable for student achievement (Rich, 2013).

1.2.1. Standards for mathematics competence

Results from the 1995 Third International Mathematics and Science Study[4] (TIMSS) point to the need for improved mathematics instruction in the United States (Beaton et al., 1996). These findings depict a consistent trend for mathematics achievement in the U.S. that continues to be evident in international comparisons (see e.g., OECD, 2013). Furthermore, studies of classroom teaching illustrate that U.S. mathematics lessons are insufficiently challenging and are often too procedurally oriented (Hiebert et al., 2003). To improve this situation, researchers and practitioners developed standards of practice to guide curriculum and assessment within mathematics (NCTM, 2000). These standards guide states in choices about their assessments, teacher preparation programs, textbooks, classroom curricula, and software programs (Common Core State Standards Initiative, 2010). Curricula based on these standards suggest not only the content that should be taught but also *how* it should be taught. For example, standards-based curricula place less emphasis on memorization, and greater emphasis on students actively "creating" mathematics through discovery or math-based activities (Kilpatrick, Swafford, & Findell, 2001; NCTM, 2000). The NCTM also defined the content areas in which students should be proficient as they progress through the school mathematics curriculum. These are: Numbers and Operations, Algebra, Geometry, Measurement, and Data Analysis and Probability.

Outside of educational technology, several studies have compared standards-based curricula with traditional curricula and have found that standards-based curricula benefit students (Fuson, Carroll, & Drueck, 2000; McCaffrey et al., 2001; Post et al., 2008; Reys, Reys, Lapan, Holliday, & Wasma, 2003; Riordan & Noyce, 2001). In one study containing 200 schools in Massachusetts, Riordan and Noyce (2001) used a quasi-experimental design and found that a standards-based curriculum had a statistically significant positive effect on students' state standardized achievement tests when compared to a traditional curriculum. Effects of standards-based curricula have also been found to extend beyond achievement on state mathematics assessments to outcomes such as number of years students elect to take mathematics courses in school and student ratings of self confidence in mathematics (Cichon & Ellis, 2003; Webb, 2003).

Although standards-based instruction has been shown to have positive effects on learning, successful implementation of such curricula may be affected by other contextual factors. For example, Polikoff (2013) investigated the association of self-reported curricular content coverage and teacher characteristics such as number of years of teaching and type of teaching credential. He found that number of years teaching and classroom size were related to alignment with state standards, suggesting that not all standards-based curricula will be implemented equally—this may have implications for the heterogeneity in positive results associated with standards-based curricula. Further exploring how standards and curricula align can inform researchers' and policy-makers' expectations for the effects achievable from the standards movement and resulting curricular reform. We make these connections within the realm of educational technology by linking the California Content Standards to the content within the ST Math program.

1.3. ST Math

ST Math is a mathematics software program developed by the MIND Research Institute to teach mathematical concepts through spatial representations (see Shaw, 1999). The current curriculum extends from kindergarten through fifth grade. It is unique in that it was developed to minimize the use of language in classroom mathematics instruction, and engages students by presenting them with a series of game-like activities that are directly tied to relevant state standards for Mathematics. Because of the minimal language use in the curriculum, ST Math is hypothesized to engage students of diverse linguistic background and is currently being used in 2050 schools across 35 states. Previous research has found that ST Math improves performance on a broad measure of mathematics achievement with an effect size of 0.07 (Rutherford et al., 2014). More targeted effects of the program have not been identified. In this paper we seek to understand how alignment of standards to content within ST Math affects student achievement, and more broadly, how consideration of alignment can inform evaluation of educational technology.

The version of ST Math studied in the current paper was implemented in California schools in the 2008–2009 and 2009–2010 school years. The structure of the program divides mathematical concepts into objectives, each of which contain games set within a particular arena. Within each game, students progress through one to ten levels by completing individual mathematics puzzles. Success on at least 80% of the puzzles within each level is required to move to the next level. Students complete the program at their own pace; however, teachers can control progress by allowing students to skip levels on which they are stuck, although reports indicate that this practice is uncommon.

Each game within ST Math is explicitly tied to at least one of the criteria within the strands of the California Content Standards. For example, in the game Number Funnels, students are asked to round the number to the nearest unit using place value. Thus, as shown in Fig. 1, students must round numbers to the nearest tenth and show their response on a number line. Developers of ST Math have noted that this game aligns with the Number Sense I strand in the California Content Standards, which contains questions about knowledge of decimals, fractions, and negative numbers (California Department of Education, 2009; MIND Research Institute, 2007).

1.4. A randomized study of ST Math

The present study uses data from a two-year randomized control trial of ST Math with third through fifth grade students within 52 Southern California public elementary schools. We ask whether the designer's alignment intentions are reflected in positive effects on student performance. Specifically, we note the designer's explicit alignment of games with specific California content standards and ask whether students who receive ST Math gain more than control counterparts on assessments measuring these targeted standards (Study 1). In light of our Study 1 results, we then go beyond the explicit intentions of the designers to explore the implicit representation of developer's intentions in an analysis of specific game features and how these features are reflected in student test score results (Study 2). Our goal is to present a balanced evaluation that utilizes standardized tests that are both available at scale and policy-relevant, but that pays attention to specifically theorized improvements by looking at measures finer grained than the global mathematics measure typically used with standardized assessments.

[4] The TIMSS provides international data on mathematics and science achievement of 4th and 8th graders in the United States and other countries. These data have been collected in 1995, 1999, 2003, 2007, and 2011 with the next wave of data collection planned for 2015.

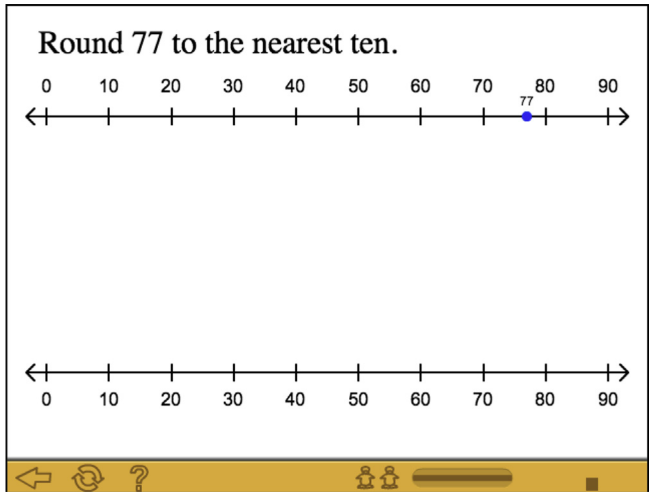


Fig. 1. Screen shot of the Number Funnels game within ST Math. Players need to indicate where on the number line 77 should be placed if rounded. This is an example of a game containing a number line.

2. Method

2.1. Participants

Southern California schools were recruited for a Goal 3 Institute of Education Sciences-funded evaluation of ST Math. Schools in the area were eligible to apply for the study if they fell in the bottom one third of the achievement distribution as measured by the mandated state standardized test. As part of a staggered design, two cohorts of selected schools were randomly assigned to treatment or control conditions at pairs of grades. During the 2008–2009 school year, eighteen cohort one schools implemented ST Math at grades two and three, and 16 schools implemented ST Math at grades four and five. In the following year, nine Cohort two schools implemented ST Math in second and third grades and nine implemented in fourth and fifth grade.

The current analysis focuses on third through fifth graders with valid pre- and posttest mathematics assessment data, resulting in a study sample of 10,860 students from 50 schools. Table 1 provides demographic information: the sample was 51% male, 85% Hispanic, 5% White, 4% Vietnamese, 2% African American, 4% other race, 91% free or reduced lunch, and 70% English language learners, reflecting the demographic makeup of the surrounding area.

2.2. Measures

2.2.1. Standardized assessment of mathematics knowledge

The California Standards Test[5] (CST) was used as a measure of mathematics knowledge. The CST is a standardized test that was given to all students in California public schools starting in second grade. Students took these tests in mathematics and English/Language Arts every year—scores had important implications for students, and also for schools under NCLB (NCLB §1111, 2001). The Mathematics CST was divided into five scores reflective of the prescribed strands within the California State Standards for mathematics. Together, these strands determined the student's overall score on the CST. These strands were: Number Sense I (NSI), Number Sense II (NSII), Algebra and Functions (A&F), Measurement and Geometry (M&G), and, Statistics, Data Analysis and Probability (SDAP). The names and general focus of these strands were consistent across grade levels but specific content differed by grade. Table 2 gives a description of the strands and provides example problems for grades 2–5. These illustrate concepts that students at the "proficient" level should have been able to demonstrate with competence. The NSI and NSII strands differed in content: NSI measured recognition of numbers and NSII measured whether students could perform basic operations with numbers.

2.2.2. Alignment of the ST Math games to standards

We used the ST Math Scope and Sequence documents (MIND Research Institute, 2007) to indicate the goals and intentions of the game developers. For each game in the ST Math curriculum, MIND designated the direct, indirect, and tangential coverage by the game of the standards: number sense, measurement and geometry, algebra and functions, and statistics, data analysis and probability. Because MIND only provided information about the coverage of number sense broadly, we used the information on sub-standards within the MIND documents to code each game as belonging to the NSI or NSII strand. Only games that were noted as directly covered by a particular standard were considered to be aligned with that standard; games that were only indirectly or tangentially aligned with standards were considered not to be aligned. Depending on the grade, between 17 and 30% of the games were aligned with NSI, 18 and 30% with NSII, 8–15% with A&F,

[5] In Spring 2014, new assessments of mathematics achievement will be piloted with students in California as part of the Smarter Balanced Assessment initiative. These assessments were designed to measure students' progress toward college and career readiness and are based on educational goals that differ from the goals of the CST (California Department of Education, 2014).

Table 1
Descriptives for the analysis sample.

	Count	Percent
Black	175	1.61
Hispanic	9238	85.06
White	557	5.13
Vietnamese	412	3.79
Other race	480	4.42
Male	5511	50.75
Free or reduced lunch	9840	90.61
English language learner	7621	70.17
Grade 3	3118	28.71
Grade 4	3596	33.11
Grade 5	4146	38.18
Treatment	5453	50.21
N	10,860	

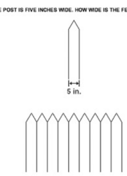
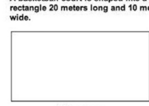
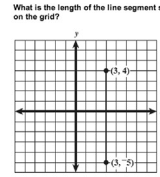
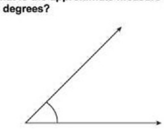
Note. Grade is grade at posttest.

4–11% with M&G and 1 and 10% with SDAP. Seventeen to 29% of the games within a given grade level were not directly covered by a standard.

2.2.3. Covariates

District records provided information on the following covariates: socioeconomic status determined by eligibility for free or reduced lunch (dichotomous), English language learner status (dichotomous), ethnicity (Hispanic, Black, White, Vietnamese and Other Ethnicity), English Language Arts score from the previous year, and gender (dichotomous).

Table 2
Description of the CST strands by grade.

	Sample Problems			
CST cluster	2nd Grade	3rd Grade	4th Grade	5th Grade
Number Sense I Understanding place value of whole numbers	[WHICH NUMBER GOES IN THE BOX?] 386 < □ < 521 297 A · 334 B · 410 C · 528 D	9000 − 3782 = A 5218 B 5328 C 6782 D 12,782	Which of these is the number 5,005,014? A five million, five hundred, fourteen B five million, five thousand, fourteen C five thousand, five hundred, fourteen D five billion, five million, fourteen	Not available
Number Sense II Solving and calculating question about he operations and fractions and decimals.	[WHICH FRACTION IS EQUAL TO ONE WHOLE?] $\frac{1}{3}$ A · $\frac{1}{8}$ B · $\frac{2}{3}$ C · $\frac{8}{8}$ D	Third-grade students went to a concert in 8 buses. Each bus took 45 students. How many students went to the concert? A 320 B 360 C 380 D 3240	$267 \div 6 =$ A 43 B 43 R3 C 44 D 44 R3	Tony had a rope 8.35 meters long. He cut off 2.6 meters. How long was the piece of rope that was left? A 5.65 meters B 5.75 meters C 6.65 meters D 6.75 meters
Algebra and Functions Interpret and solve problems with operations	[WHAT NUMBER GOES IN THE BOX TO MAKE THIS NUMBER SENTENCE TRUE?] 15 + 8 = □ + 15 7 A · 8 B · 15 C · 23 D	If Mai bought apples for $2.50 and she paid with a $10 bill, which expression shows the correct amount of change? A $10 + $2.50 B $10 − $2.50 C $10 × $2.50 D $10 ÷ $2.50	The sum of *x* plus *y* equals 26. If $x = 17$, which equation can be used to find the value of *y*? A $y - 17 = 26$ B $17 + y = 26$ C $x - y = 26$ D $x + 17 = 26$	What value for *z* makes this equation true? $8 \times 37 = (8 \times 30) + (8 \times z)$ A 7 B 8 C 30 D 37
Measurement and Geometry Using measurement to solve problems. Computing area and volume of simple objects.	[EACH FENCE POST IS FIVE INCHES WIDE. HOW WIDE IS THE FENCE IN THE PICTURE?]  30 inches A · 45 inches B · 50 inches C · 65 inches D	A basketball court is shaped like a rectangle 20 meters long and 10 meters wide.  What is the perimeter in meters of the court? A 30 meters B 50 meters C 60 meters D 200 meters	What is the length of the line segment shown on the grid?  A 9 units B 7 units C 5 units D 4 units	What is the approximate measure of this angle in degrees?  A 20° B 45° C 110° D 135°
Statistics, Data Analysis, and Probability Students interpret data on bar graphs and make predictions.	[Carrie practices the piano each day. The table shows how long she practiced each day last week. How many minutes longer did she practice on Wednesday than on Tuesday? Mark your answer.] Piano Practice Times Day: Minutes Monday: 26 Tuesday: 24 Wednesday: 30 Thursday: 35 Friday: 15 6 A · 5 B · 4 C · 2 D	Not available	Not available	Not available

2.3. Analysis plan

In order to predict which mathematics skills were affected by ST Math, we conducted five separate regression analyses (one corresponding to each CST strand) by regressing CST posttest for a given strand on participation in ST Math along with English language status, previous year's ELA score, free or reduced lunch status, gender, ethnicity, and same strand pretest score as control variables.

3. Results and discussion

Table 3 presents the descriptive statistics on mathematics pretest scores, separately for treatment and control students. Although the schools were randomly assigned to condition, pretest scores differ at a level that attains statistical significance ($p < .05$) between treatment and control groups in all grades and strands except for SDAP in fifth grade. For example, third graders in the control condition had higher pretest scores than students in the treatment condition whereas fourth graders in the treatment condition had higher pretest scores than students in the control condition. These differential start-points are controlled with CST pretest in the regression-adjusted results.

Table 4 presents results from the regression of CST mathematics strand scores on treatment status after one year. ST Math had a small, positive effect on all strands. Aggregating all effects produced a total effect of 0.22 of a standard deviation. When looking at this effect by strand, it attained statistical significance only for the NSI strand, where treatment students scored nearly one-sixth of a standard deviation higher than control students ($\beta = 0.14$, $p < .0001$). Although the effects on other strands did not reach statistical significance, small positive trends were seen for all: NSII ($\beta = 0.05$, $p = .28$), A&F ($\beta = 0.01$, $p = .88$), M&G ($\beta = 0.01$, $p = .88$), and SDAP ($\beta = 0.01$, $p = .88$). Treatment by grade interactions were found to be insignificant and are therefore not included in the table or further discussion.

In considering whether these results reflect the intended alignment between ST Math and the mathematics content standards, we can compare the percent of standards coverage across the games and the percentage of contribution to the combined effect for each CST strand. This is a rough estimate, but may provide some evidence of alignment or lack thereof (see Fig. 2). The NSI strand represented 64% of the total effect of ST Math on mathematics CST scores; however, games labeled as covering NSI only accounted for between 17 and 30% of the content of ST Math, depending on grade level. NSII had a similar content coverage range (18–30%), and yet the effect of ST Math on student achievement scores for the NSII CST strand accounted for only 23% of the effect. The marginal treatment effect on the other three strands each accounted for only 5% of the aggregate effect. Differential placement of games aligned with NS strands could account for their overrepresentation in the effect of ST Math—on average, students only complete 73% of the program, the beginning of which may weight more heavily toward NS content. To address this situation, we conducted additional regressions controlling for the percent of the program completed by the students. These analyses revealed no differences in the effect of ST Math on all strands.

The results of the analyses in Study 1 reveal an effect of ST Math largely explained by the NSI strand. This effect is not in line with developer documentation regarding intended content coverage. To investigate other aspects of the program that may be producing this effect, we performed a content analysis to reveal program features that may weigh more heavily toward content or skills assessed within NSI. The results of these analyses are detailed in Study 2.

4. Study 2

To better understand the features of ST Math that may be supporting its effect on NSI, we look to the concept of "number sense," its definition, and prior research on interventions for its improvement. Number sense is an ability to understand magnitude of numbers and to approximate and manipulate numerical quantities (Dehaene, 1997). Early number sense is crucial for students' success in later mathematics

Table 3
Pretest raw scores by math CST cluster and grade.

		Grade 3		Grade 4		Grade 5	
		Control	Treatment	Control	Treatment	Control	Treatment
NSI	Min	0	0	0	1	3	1
	Max	15	15	16	16	17	17
	Mean	10.89	10.39	10.74	11.25	12.47	12.71
	SD	3.25	3.43	3.2	2.98	3.2	3.21
NSII	Min	3	2	0	1	1	0
	Max	23	23	16	16	14	14
	Mean	16.3	15.85	9.84	10.44	9.56	10.03
	SD	4.4	4.5	3.45	3.4	3.33	3.34
A&F	Min	0	0	0	0	1	2
	Max	6	6	12	12	18	18
	Mean	4.54	4.36	7.7	8.13	12.99	13.52
	SD	1.38	1.44	2.43	2.42	3.96	3.92
M&G	Min	0	0	0	2	0	0
	Max	13	14	16	16	12	12
	Mean	8.85	8.28	11.59	12	7.38	7.93
	SD	2.28	2.49	2.61	2.5	2.75	2.78
SDAP	Min	0	0	0	0	0	0
	Max	7	7	5	5	4	4
	Mean	5.05	4.74	3.77	3.9	2.69	2.73
	SD	1.61	1.69	1.2	1.13	1.1	1.09
N		1406	1712	1873	1723	2128	2018

Note. Pretest scores for third, fourth, and fifth grade are significantly different between treatment and control conditions except for fifth grade. Grade levels represent grade at year of implementation; pretest scores are those from one year prior to implementation. Max indicates how many questions are asked in each strand.

Table 4
One year effect of ST Math on CST strand scores.

	Posttest NSI	Posttest NSII	Posttest A&F	Posttest M&G	Posttest SDAP
Treatment	0.14** (0.04)	0.05 (0.04)	0.01 (0.04)	0.01 (0.04)	0.01 (0.04)
Corresponding math pretest	0.40*** (0.01)	0.42*** (0.01)	0.35*** (0.01)	0.35*** (0.01)	0.35*** (0.01)
ELA pretest	0.31*** (0.01)	0.30*** (0.01)	0.36*** (0.01)	0.36*** (0.01)	0.36*** (0.01)
English language learner	0.05* (0.02)	0.04* (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
Male	0.09*** (0.02)	0.03 (0.02)	0.10*** (0.01)	0.10*** (0.01)	0.10*** (0.01)
White	−0.01 (0.04)	−0.10* (0.05)	−0.03 (0.04)	−0.03 (0.04)	−0.03 (0.04)
Black	−0.16* (0.07)	−0.19*** (0.05)	−0.17** (0.06)	−0.17** (0.06)	−0.17** (0.06)
Vietnamese	0.22*** (0.03)	0.27*** (0.03)	0.24*** (0.03)	0.24*** (0.03)	0.24*** (0.03)
Other ethnicity	0.02 (0.03)	0.07* (0.03)	0.06 (0.03)	0.06 (0.03)	0.06 (0.03)
Grade 3	0.01 (0.05)	0.01 (0.05)	0.00 (0.04)	0.00 (0.04)	0.00 (0.04)
Grade 4	0.02 (0.05)	0.02 (0.05)	0.01 (0.04)	0.01 (0.04)	0.01 (0.04)
Free or reduced lunch	−0.04 (0.02)	−0.05 (0.03)	−0.01 (0.03)	−0.01 (0.03)	−0.01 (0.03)
Constant	−0.13 (0.07)	−0.03 (0.08)	−0.06 (0.07)	−0.06 (0.07)	−0.06 (0.07)
N	10,860	10,860	10,860	10,860	10,860
R-sq	0.43	0.43	0.43	0.43	0.43

Note. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. Standardized regression coefficients, standard errors in parentheses. Standard errors have been corrected for nesting by clustering on school, resulting in 50 clusters. NSI is number sense I, NSII is number sense II, A&F is algebra and functions, M&G is measurement and geometry, and SDAP is statistics, data analysis and probability. Students who have valid pre- and post data for all 5 strands were included in this sample. 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, fifth grade serves as the reference grade, and Hispanic as ethnic reference group.

(Carr, 2012; Jordan, Glutting, & Ramineni, 2010; Siegler, 2009). Without more advanced conceptualizations of number, students may not be able to grasp more higher-level mathematics concepts such as place value or part-whole relationships of numbers (Carr, 2012). Number sense is developed as a result of children exploring numbers, visualizing them in a variety of contexts, and relating numbers to each other (Howden, 1989). Greeno (1991) suggests that number sense can be likened to cognitive expertise—to develop this expertise, learners must engage in substantial interaction with the conceptual environment of numbers. One activity that may support this necessary interaction is exposure and play with number lines. Interventions focusing on number lines can improve number sense as seen in the ability to compare magnitudes, estimate on number lines, count, and identify numerals (e.g., Ramani & Siegler, 2011; Siegler & Ramani, 2008, 2009; Wilson et al., 2009). Previous studies have found that in order to improve number sense, learners must have repeated exposures to visual representations such as number lines, pictorial representations of numbers, and material linking symbolic and non-symbolic representations of numbers (Kilpatrick et al., 2001; Wilson et al., 2009).

To the extent that these elements supporting number sense are present in ST Math, the curriculum can be considered a number sense intervention, and an improvement in student number sense skills should be both expected and encouraging. To this end, we sought to understand the ubiquity of these elements throughout the games, and in particular, within and outside of the games that ST Math developers had labeled as specifically designed to impact NSI, the strand containing skills related to number sense as defined above.

5. Method

5.1. Game coding

Four trained research staff coded all games within the third through fifth grade 2010–2011 and 2011–2012 ST Math curricula for the presence of two characteristics related to the development of number sense. The first code recorded whether there was a number line present in the level. This included any line that had visible notches with or without numbers (see Fig. 1). The second code recorded if blocks, animals, bills, coins, or any other countable object represented numbers in the level (see Fig. 3). Four coders each trained on the first grade curriculum until they reached 90% agreement on all codes. Then each researcher individually coded 20% of the second grade curriculum. All

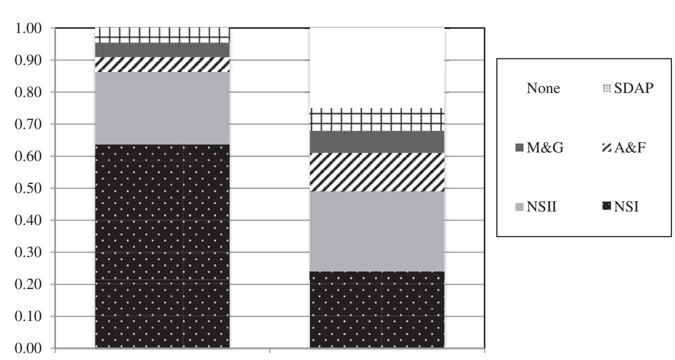


Fig. 2. Comparison of alignment of games and their effects on student achievement. Student mathematics achievement was measured using the raw scores of the California Standards Test (CST). Alignment of games to standards was determined using documentation from MIND Research Institute (2007). Effect shows the percentage of each strand on the total effect of ST Math on student achievement. Standards show the percentage of games aligned with each CST strand. Number sense I is NSI, number sense II is NSII, A&F is algebra and functions, M&G is measurement and geometry, and SDAP is statistics, data analysis and probability.

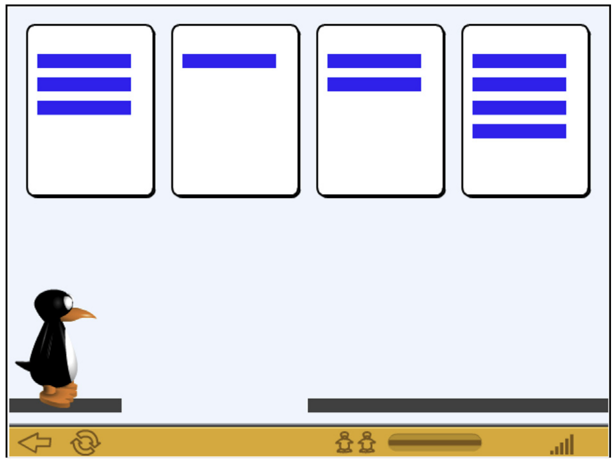


Fig. 3. Screen shot of ST Math Jiji the penguin needs to cross the gap and walk toward the right part of the screen. Students must choose the correct number of bars that can perfectly fill the gap and allow Jiji to cross. This is an example of a game that contains numbers represented as objects.

four researchers coded an additional 20% of the curriculum. Kappa was calculated for this 20% and addressed before the next grade, third grade, was begun. If there was a disagreement among the coders on the overlapping 20%, the result was determined by majority. If there was an even split, the researchers met to discuss and resolve the code. All the study grades were coded in this manner. Final Kappas for each of the two elements (number line and numbers as objects) are: 0.96 and 0.76 respectively for third grade; 0.83 and 0.70 for fourth grade; and 0.83 and 0.60 for fifth grade.

6. Results and discussion

Overall, between 147 and 222 games, depending on grade level, included at least one of the elements hypothesized to support number sense. Averaged across grades, this represents 58% of the ST Math content. Sixty-one percent of the games noted by the developers as covering NSI contained at least one number sense element and 76% of the games noted as covering NSII contained at least one number sense element. Looking outside of the stated number sense strands (NSI and NSII), a good deal of games not labeled as addressing number sense nevertheless contained number sense elements. Fig. 4 illustrates the presence of each of the elements hypothesized to support number sense across games within each of the five strands. Results are averaged across grade levels.

As depicted in Fig. 4, we found that of the games aligned with NSI, 26% contained a number line and 35% contained numbers as objects. Of the games aligned with NSII, 38% contained number lines and 38% contained numbers as objects. Of the games aligned with A&F, 19% contained number Lines and 43% contained numbers as objects. Of the games aligned with M&G, 2% contained number lines and 32% contained numbers as objects. Of the games aligned with SDAP, 36% contained number lines and 60% contained numbers as objects. Finally, of games not aligned with a standard, 23% contained a number line and 23% contained numbers as objects. Thus, there is evidence that games coded as NSII, A&F, M&G, and SDAP did in fact contain content related to student development of number sense skills. Note that even games MIND did not align with a particular standard contained elements known to increase students' number sense. Within ST Math, a common game format (present across 24% of the games overall) represented problems and/or answer choices along a number line, providing an environment ripe with opportunities for students to explore number relationships.

7. Summary and concluding discussion

This paper presents results of a randomized control trial evaluation of a supplementary mathematics software, ST Math, and its effect on specific mathematics skills. This paper also evaluates the design features present within the program and their relation to an important

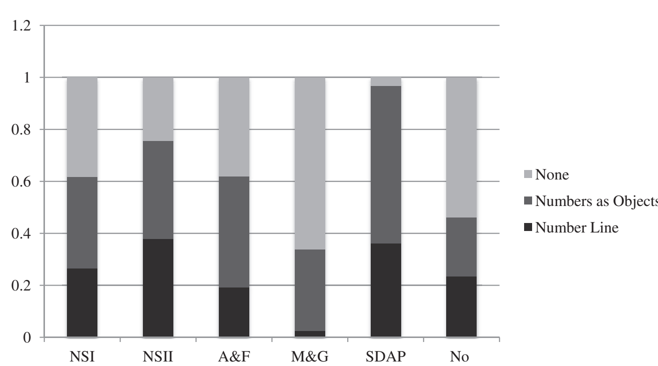


Fig. 4. Distribution of number sense elements contained within games aligned to each standard averaged across grades 3–5. Alignment of games to standards was determined using documentation from MIND Research Institute (2007). Four coders coded each of the games that contain number lines and numbers as objects.

mathematical skill, number sense. In study 1, we found a statistically significant positive effect ($\beta = 0.14$, $p < .0001$) of ST Math on NSI skills. This effect size was twice that of the effect of ST Math on the CST mathematics scale score, which factored in the lower effects on the other strands (Rutherford et al., 2014 and replicated with the current sample). Additionally, when looking at the alignment between the developer's intentions and the distribution of the effect, we found that a large portion of the effect was on NSI, rather than distributed across the strands as the developer claimed. To investigate this further, we conducted qualitative coding to understand the elements within the game that could contribute to an effect on number sense as measured by NSI. In study 2, we found that many of the elements that could be responsible for this increase were present throughout the ST Math curriculum, suggesting a misalignment between the developer's stated intentions and the actual content of the mathematics software.

7.1. Alignment between assessment and intervention

In order to fully understand the effectiveness of mathematics software, it is useful to examine the match between the instructional material and the items used in assessment. For a subject such as mathematics, where more advanced skills require the mastery of basic skills, assessment of specific skills may be vital to identify critical areas of need and improvement. However, as researchers move toward readily available broad standardized tests, analyses of targeted effects may become rare. Within the present study, relying only on the CST scale score, a global measure of math achievement (as in Rutherford et al., 2014), would not have revealed important information about the effect of ST Math on specific mathematics competencies—useful information for both the developer and the consumer of ST Math. Policy-relevant assessments are important for developers and practitioners; however, the alignment between assessment and developer goals is critical for ascertaining whether a software product or other educational tool has met its intended results.

7.2. Alignment between intentions and intervention

In study 2, we found that the games indicated by the developers as covering skills outside of number sense still presented students with opportunities to engage in tasks known to affect number sense, such as interaction with number lines and numbers represented as objects. The developers had not indicated this overarching focus on number sense in their documentation, and therefore, an evaluator may have failed to adequately assess for an effect that was nevertheless in line with the content of the intervention. Educational technology offers numerous opportunities to present multiple representations for mathematics (see NCTM, 2000)—an analysis of the affordances inherent in a particular game's mechanics or system of representation might add value when determining a developer's intended results. Evaluators should be mindful of these possibilities; however, the best chance for a developer's intentions to be reflected in assessment is to explicitly document goals and desired outcomes drawn from all aspects of the program, including those implicit in the format or structure.

7.3. Limitations and future directions

The present study relied upon the freely available California standardized tests. Although the ST Math developers stated that the content standards upon which these assessments were based were aligned with the ST Math software, it is possible that an assessment more closely aligned to the actual content of ST Math may have produced different results. Although somewhat targeted, the length of time between assessments (the CST is offered only once per year), does not offer information on ST Math's success in teaching *specific* mathematics skills such as those tied to individual objectives, especially because not all students receive all content within the ST Math curriculum. Some students failed to progress to the later objectives within the program, and others received a differently ordered curriculum due to teacher control. We were unable to measure these variations in the mathematics content taught. Additionally, although the CST was a policy-relevant assessment at the time of this study (2007–2010), as the state moves to the Common Core Standards, new tests are being piloted (CA Dept. of Education, 2014).

We found that ST Math's strongest effect was on our measure of number sense and that a potential explanation for this effect was the prevalence of elements supporting number sense development within the software. The results from Study 1 informed our decision to code for number sense elements—it is possible that elements that support other mathematics skills were present and were not identified by our measures in Study 1 or our coding in Study 2. Future studies could take a different approach and let the actual (as compared to claimed) content of the software drive the coding and assessment. Additionally, further conversations with the developers or additional document review could inform analyses that more closely couple goals and assessment.

In addition to evaluating the effect of ST Math on specific mathematics competencies, this study aimed to provide developers of educational technology with a set of considerations for designing learning experiences and documenting their goals. By considering the alignment between what skills they intend their software to affect, the elements contained within the games, and student achievement in specific content areas, developers can be better positioned to produce mathematics educational technology with implications for learning that are both meaningful and measurable.

Conflict of interest

The second author has recently begun working as a part-time contract employee for MIND Research Institute on an unrelated project. Her work began after the completion of data collection and analysis. MIND Research had no involvement in the analysis or interpretation of results, or in the preparation, review, or approval of this manuscript.

Acknowledgments

We would like to thank Melissa Kibrick, Dao Vu and Anai Lepe for their assistance in the qualitative coding of this study, the MIND Research Institute, Orange County Department of Education, and the students and communities of the participating elementary schools.

References

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research, 73*(3), 277–320.
Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study*. Chestnut Hill, MA: Boston College.
California Department of Education. (2009). *Released test items*. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/cstrtqmath4.pdf.
California Department of Education. (2014, February 3). *Spring 2014 smarter balance field test*. Retrieved from http://www.cde.ca.gov/ta/tg/sa/smarterfieldtest.asp.
Carr, M. (2012). Critical transitions: arithmetic to algebra. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 3. Applications to learning and teaching* (pp. 229–255). Washington, DC: American Psychological Association.
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.
Cichon, D., & Ellis, J. G. (2003). Connections on student achievement, confidence, and perception. In S. L. Senk, & D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?*. Mahwah, NJ: Erlbaum.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Common Core State Standards Initiative. (2010). *Common Core State Standards for mathematics.* Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf.
Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., et al. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education, 68*, 495–504.
Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.
Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., et al. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort*. Washington, DC: U.S. Department of Education.
Fuson, K. C., Carrol, W. M., & Drueck, J. V. (2000). Achievement results for second and third graders using the standards-based curriculum everyday mathematics. *Journal for Research in Mathematics Education*, 277–295.
Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education., 22*(3), 170–218.
Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., et al. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. NCES 2003-013. Washington, DC: U.S. Department of Education, National Center for Education Statistics.
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*(3), 172–177. http://dx.doi.org/10.1111/j.1750-8606.2008.00061.x.
Howden, H. (1989). Teaching number sense. *The Arithmetic Teacher, 36*, 6–11.
Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences, 20*(2), 82–88.
Ke, F. (2008). A case study of computer gaming for math: engaged learning from gameplay? *Computers & Education, 51*(4), 1609–1620.
Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education, 55*(2), 427–443.
Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.
Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*(1), 1–32.
Lopez-Morteo, G., & López, G. (2007). Computer support for learning mathematics: a learning environment based on recreational learning objects. *Computers & Education, 48*(4), 618–641.
Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332–1361.
McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P., Bugliari, D., & Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: the case of standards-based high school mathematics. *Journal for Research in Mathematics Education*, 493–517.
MIND Research Institute. (2007). *ST math scope and sequence*. [Alignment of games to California State Standards grades 3–5]. Available from. Irvine, CA: MIND Research Institute.
National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
National Mathematics Advisory Panel. (2008). *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. No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107–110, § 115, Stat. 1425(2002).
OECD. (2013). *Lessons from PISA 2012 for the United States, strong performers and successful reformers in education*. OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264207585-en.
Page, M. S. (2002). Technology-enriched classrooms: effects on students of low socioeconomic status. *Journal of Research on Technology in Education, 34*(4).
Polikoff, M. S. (2013). Teacher education, experience, and the practice of aligned instruction. *Journal of Teacher Education, 64*(3), 212–225.
Post, T. R., Harwell, M. R., Davis, J. D., Maeda, Y., Cutler, A., Andersen, E., et al. (2008). Standards-based mathematics curricula and middle-grades students' performance on standardized achievement tests. *Journal for Research in Mathematics Education*, 184–212.
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*(3), 146–159.
Reys, R., Reys, B., Lapan, R., Holliday, G., & Wasman, D. (2003). Assessing the impact of "standards"-based middle grades mathematics curriculum materials on student achievement. *Journal for Research in Mathematics Education*, 74–95.
Rich, M. (2013, July 23). Education overhaul faces a test of partisanship. *The New York Times*. Retrieved from http://www.nytimes.com/2013/07/24/us/politics/education-overhaul-faces-a-test-of-partisanship.html?_r=0.
Riordan, J. E., & Noyce, P. E. (2001). The impact of two standards-based mathematics curricula on student achievement in Massachusetts. *Journal for Research in Mathematics Education*, 368–398.
Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., et al. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics. *American Educational Research Journal, 47*(4), 833–878. http://dx.doi.org/10.3102/0002831210367426.
Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Graham, J., Kibrick, M., et al. (2014). A randomized trial of an elementary school mathematics software intervention: Spatial–Temporal (ST) Math. *Journal of Research on Educational Effectiveness*. http://dx.doi.org/10.1080/19345747.2013.856978.
Schoenfeld, A. H. (2002). Making mathematics work for all children: issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13–25.
Schoenfeld, A. H. (2006). What doesn't work: the challenge and failure of the what works clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*(2), 13–21.
Shaw, G. L. (1999). *Keeping Mozart in mind* (1st ed.). Academic Press.
Siegler, R. S. (2009). Improving the numerical understanding of children from low-income families. *Child Development, 3*(2), 118–124.
Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science, 11*(5), 655–661.
Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology, 101*(3), 545.
Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: a best-evidence synthesis. *Review of Educational Research, 78*(3), 427–515.
Suppes, P., Liang, T., Macken, E. E., & Flickinger, D. P. (2014). Positive technological and negative pre-test-score effects in a four-year assessment of low socioeconomic status K–8 student learning in computer-based math and language arts courses. *Computers & Education, 71*, 23–32.
Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics and education*. Madison, WI: National Institute for Science Education.
Webb, N. L. (2003). Interactive mathematics program on student. In S. L. Senk, & D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?*. Mahwah, NJ: Erlbaum.
Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an adaptive game intervention on accessing number sense in low-socioeconomic-status kindergarten children. *Mind, Brain, and Education, 3*(4), 224–234.
Wu, H. L., & Pedersen, S. (2011). Integrating computer-and teacher-based scaffolds in science inquiry. *Computers & Education, 57*(4), 2352–2363.