Paper to appear in the *American Educational Research Journal* in 2010

Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-Scale Studies

Jeremy Roschelle[1], Nicole Shechtman[1], Deborah Tatar[2], Stephen Hegedus[3], Bill Hopkins[4], Susan Empson[5], Jennifer Knudsen[1], Larry Gallagher[1]

[1]SRI International
[2]Virginia Tech
[3]University of Massachusetts Dartmouth
[4]Charles A. Dana Center, University of Texas at Austin
[5]University of Texas at Austin

October 20, 2009

Correspondences
Please address all correspondences to Jeremy Roschelle, jeremy.roschelle@sri.com.

Acknowledgements
This material is based on work supported by the National Science Foundation under Grant No. 0437861. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank G. Haertel, G. Estrella, K. Rafanan, P. Vahey, S. Carriere, H. Javitz, M. Robidoux, S. Hull, L. Hedges, S. Goldman, H. Becker, J. Sowder, G. Harel, P. Callahan, F. Sloane, B. Fishman, K. Maier, J. Earle, R. Schorr, M. Dunn, A. Stroter, and B. McNemar for their contributions to this research. We thank the participating teachers and Texas Educational Service Center leaders from regions 1, 6, 9, 10, 11, 13, 17 and 18; this project could not have happened without them. We thank and remember Jim Kaput, who pioneered SimCalc as part of his commitment to democratizing mathematics education.

Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-Scale Studies

Abstract

We present three studies (two randomized controlled experiments and one embedded quasi-experiment) designed to evaluate the impact of replacement units targeting student learning of advanced middle school mathematics. The studies evaluated the SimCalc approach, which integrates an interactive representational technology, paper curriculum, and teacher professional development. Each study addressed both replicability of findings and robustness across Texas settings with varied teacher characteristics (backgrounds, knowledge, attitudes) and student characteristics (demographics, levels of prior mathematics knowledge). Analyses revealed statistically significant main effects, with student-level effect sizes of .63, .50, and .56. These consistent gains support the conclusion that SimCalc is effective in enabling a wide variety of teachers in a diversity of settings to extend student learning to more advanced mathematics.

Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-Scale Studies

Introduction

Middle school is an important transition point in students' school trajectories, especially with regard to mathematics (Nathan & Koellner, 2007). Starting in middle school, mathematical concepts become increasingly difficult; they are more abstract, and understanding requires making connections across algebraic and graphical representations (Leinhardt, Zaslavsky, & Stein, 1990). International comparison research shows that although U.S. fourth-grade students compare favorably, eighth-grade students fall behind their foreign peers, particularly in their mastery of more advanced mathematics (Schmidt et al., 2001). This is a cause for concern not only for students' future development in mathematics, but also for students' preparation for careers in science, engineering, and technology (Tai, Qi Liu, Maltese, & Fan, 2006).

Further evidence of the need to improve American students' mathematics achievement comes from the National Assessment of Educational Progress (NAEP) (National Center for Education Statistics, 2006). Although NAEP results reveal improvement in middle school mathematics learning over time, the trend falls far below the No Child Left Behind goal of enabling all students to achieve proficiency by 2014. NAEP results show that in eighth grade, only 47% of White students and 54% of Asian students rise above the *basic* level to *proficient* or *advanced* use of mathematical concepts and skills. The often cited evidence of an achievement gap is that only 13% of Black students, 19% of Hispanic students, and 13% of students eligible for free or reduced-price lunch demonstrate proficiency in eighth grade (National Center for Education Statistics, 2006, p. 37). Evaluating research-based curricular programs that aim to

enable all students to learn more advanced mathematics is an important goal.

This article reports results from a series of studies within a research program on scaling up and evaluating a technology-based approach to advanced mathematics called SimCalc. The SimCalc approach is intended by its developers to "democratize access to the mathematics of change and variation" by enabling a broader range of students to learn more advanced mathematical concepts and skills. Further, developers intend for SimCalc students to learn more advanced mathematics without jeopardizing progress on basic mathematics skills.

The approach was originally developed in design research conducted in just a few classrooms at a time with frequent researcher participation in daily teaching and learning activities. This method allows rapid iterative improvement of a design, but does not yield strong evaluative results. Like most design research (Design Based Research Collaborative, 2002), early SimCalc research had small convenience samples of teachers whom the research team provided with unmeasured amounts of support. With a convenience sample, research findings are subject to the "boutique critique"—that the schools, teachers, and students are special in some important way and that the results will not be generalizable (Roschelle, Tatar, & Kaput, 2008). In order to better evaluate the promise of the design, the research and developed team initiated a program of scaling up so that the design could be evaluated in experimental field trials with more careful attention to sampling populations.

Important scaling goals of the research reported in this paper were to extend the approach to over one hundred schools and teachers and thousands of students while removing the participation of researchers in the classrooms. Scaling up, however, is not merely about achieving a larger *N* (Cobern, 2002). A key element of the program was also to more tightly integrate software, curriculum and professional development so that teachers and students would

experience an aligned intervention. We have written extensively about the process of scaling up elsewhere (Roschelle, Tatar, & Kaput, 2008; Roschelle, Tatar, Shechtman, & Knudsen, 2008) with regard to process of going from design research to a well-specified intervention; this paper focuses on evaluation of SimCalc as packaged in intervention form and delivered at a reasonable scale.

We report results from two experimental comparisons and one quasi-experimental comparison within two different grade levels. By presenting all three comparisons here, we allow the examination of replicability of results. In addition to replicability, a goal of this research program was to evaluate the SimCalc approach with a wide variety of teachers in diverse demographic settings.

After providing background on research on technology in mathematics education, we give an overview of the SimCalc approach to integrating professional development, curriculum, and technology. We then describe three studies. The first study, in the seventh grade, had the largest sample. In addition to evaluation of main effects, this experiment allowed comparisons among geographical areas with strikingly different demographics. In the second study, we followed teachers from the control group in the first experiment as they implemented the SimCalc approach in the next school year. This quasi-experiment allowed us to contrast treatment and control conditions within a teacher and across student cohorts. In the third study, we examined an extension to eighth grade of the same overall SimCalc approach but with a more advanced mathematical topic. In addition, the eighth-grade experiment provided professional development via a train-the-trainers paradigm, an additional realistic element of scaling up and an additional means to minimize the potential that the program developers were directly involved in producing observed outcomes. The train-the-trainers paradigm has been used in two

reportedly successful programs for reading, Reading Recovery and Success for All (Denton, Vaughn, & Fletcher, 2003; Fletcher, Foorman, Dento, & Vaughn, 2006). By examining three sufficiently powered comparative studies, we are able to bring considerable scientific evidence to bear on the replicability and robustness of the SimCalc approach as well as the potential to deploy it via a train-the-trainers model. In our discussion, we draw conclusions based on the similar pattern of results across experiments and discuss limits to generalizability.

Technology in Mathematics Education

Technology has been used throughout the history of mathematics education. Socrates drew figures in the sand to illustrate his points (Jowett, 1875). Many ancient and modern societies used an abacus both as a procedural tool and conceptual model of arithmetic (Fauvel & Maanen, 2000). More recently, most elementary schools introduced use of physical manipulatives, such as Dienes' Blocks, for introducing place value and other arithmetic concepts (Varelas & Becker, 1997). Calculators and graphing calculators are common in secondary schools (and controversial in lower grades). Four-function and scientific calculators are typically used to simplify tedious computations, leading to greater focus on the pedagogical point, and graphing calculators can be important in students' development of their conceptual understanding of mathematical functions (Doerr & Zangor, 2000; Ellington, 2003). Thus, technology supports both *computation* and *representation.* In particular, technology can support mathematical ideas in ways that are important for conceptual understanding (Kaput, Hegedus, & Lesh, 2007).

Computer software for mathematics learning can take many forms and operate through different cognitive mechanisms. Software can provide students with opportunities for practice

and rapid feedback in a motivating environment or have higher order cognitive goals (Wenglinksy, 1998). Programming tools can provide opportunities for students to write mathematical programs, and developmental principles suggest that constructing programs can lead to constructing knowledge (Papert, 1980). Experimental research has documented positive effects from the use of the programming language Logo (National Mathematics Advisory Panel, 2008). Educational applications can build on cognitive theory (Anderson, Corbett, Koedinger, & Pelletier, 1995; Corbett, Koedinger, & Hadley, 2001) by tracing students' step-by-step progress on a problem and intervening when their performance differs from expert behavior. Wireless networks may support learning via formative assessment and/or social mechanisms that operate at classroom level (Roschelle, 2003). Given the long history of technology in mathematics education and the many differences in approach and application, useful research must now go beyond making claims about technology in general or in isolation from its use; specific approaches must be described—with their entailments, assumptions, and goals—and evaluated on their merits (National Mathematics Advisory Panel, 2008).

The intervention discussed in this paper takes a representational approach (Kaput, 1992) in which computers are seen as supporting new visualizations of and interactions with mathematical objects. Graphical representations (e.g., graphs, geometric figures, animations) are often juxtaposed with linguistic representations (e.g., text, verbal narratives, algebraic symbols). Dienes (the designer of Dienes' Blocks) provided the rationale for this juxtaposition; he argued that because mathematical concepts are abstract but human minds develop concepts from concrete experiences, we can often best come to understand an abstraction by interacting with multiple concrete embodiments (Dienes, as cited in Lesh, Cramer, Doerr, Post, & Zawojewski, 2003). Each of these concrete embodiments will be impure in some ways, but by striving to

make connections between embodiments at a conceptual level we can grasp the intended mathematical abstraction.

Research on the Representational Approach to Technology

The research base supporting a representational approach is broad but fragmented. Cognitive theory supports the approach via the multimedia principle, which has firmly established the benefits of carefully integrated presentations of the same concept in linguistic and graphical forms (Mayer, 2005). In related research, a meta-analysis that summarized findings from more than 100 research studies involving 4,000+ experimental/control group comparisons revealed that both representing knowledge graphically and using manipulatives to explore new knowledge and practice applying it had a large effect size. “The overall effect size for these techniques was .89, indicating a percentile gain of 31 points. The use of computer simulation as the vehicle with which students manipulate artifacts produced the highest effect size” (Marzano, 1998, p. 91). Despite the positive findings of this meta-analysis, one limitation is that in none of the experiments was multilevel linear modelling used to account for the nesting of students within teachers, and few carefully described the role of technology in the overall intervention. This is a serious limitation because it fails to rule out clustering effects.

As yet, randomized experiments using multilevel linear modelling to explore the use of technology and its mediators and moderators in mathematics education are rare, and no such experiments have been conducted on the representational approach. The most prominent random-assignment experiment is the National Study of the Effectiveness of Educational Technology Interventions (EETI), which found “test scores were not significantly higher in classrooms using selected reading and mathematics software products” (Dynarski et al., 2007, p.

xiii). For the EETI study, 16 mathematics and reading products were chosen on the strength of prior research and their readiness for large-scale deployment. The products were evaluated in a random-assignment experiment in hundreds of schools, with vendors providing as much support as they determined to be necessary. Nearly all the products were well used and well liked by the participating teachers. Yet the results were not different from business as usual. The EETI did not, however, included products that take a representational approach.

To our knowledge, the research reported in this paper is the first to examine a representational approach within a program of randomized controlled experimentation with a sufficiently large scope to use multilevel modelling. The advantages of multilevel modelling are that we can both correct for analytic errors in single level analyses and examine both the main effects and the moderating effects of teacher-, classroom-, or school-level variables on student-level outcomes.

Overview of the SimCalc Approach to Mathematics Learning

The SimCalc Project, based at the James J. Kaput Center at the University of Massachusetts, Dartmouth, has a research program with a broader scope than this series of studies, extending to more grades, more technologies, and more mathematics. In the overall project mission of “democratizing access to the mathematics of change and variation” (Kaput, 1994; Kaput, 1997), the phrase “the mathematics of change and variation” is meant to highlight the strand of mathematics relating to much of algebra and leading to calculus. Kaput used this phrase in contrast to the mathematics of uncertainty (probability and statistics) and the mathematics of space (geometry). The mathematics of change and variation emphasizes the concepts of rate and accumulation as thematic content that can be developed across many grade

levels. A foundational belief of the SimCalc project team is that reconceptualizing some of the mathematics in middle school and high school in light of the broader mathematics of change and variation can yield a more coherent and fruitful mathematical experience for both disadvantaged and advantaged learners (Kaput, 1997). The phrase “democratizing access” refers to desire to enable students in disadvantaged settings to have a better opportunity to learn advanced and important mathematics. Scaling up has been a theme throughout the SimCalc program (Roschelle, Tatar, & Kaput, 2008; Roschelle, Tatar, Shechtman, & Knudsen, 2008). Early cognitive and developmental research in the program involved small numbers of students. As the program evolved over a decade, the research shifted to tens of teachers, then tens of schools, and in these experiments a handful of geographic areas.

The hallmarks of the SimCalc approach to the mathematics of change and variation are the following:

1. Anchoring students’ efforts to make sense of conceptually rich mathematics in their experience of familiar motions, which are portrayed as computer animations;
2. Engaging students in activities to make and analyze graphs that control animations;
3. Introducing piecewise linear functions as models of everyday situations with changing rates;
4. Connecting students' mathematical understanding of rate and proportionality across key mathematical representations (algebraic expressions, tables, graphs) and familiar representations (narrative stories and animations of motion);
5. Structuring pedagogy around a cycle that asks students to make predictions, compare their predictions with mathematical reality, and explain any differences.

The SimCalc MathWorlds software provides a “representational infrastructure” (Kaput et al. 2007; Kaput & Roschelle, 1998) that is central to enabling this approach. Most distinctively, the software presents animations of motion (Figure 1). Students can control the motions of animated characters by building and editing mathematical functions in either graphical or algebraic forms. After editing the functions, students can press a *play* button to see the corresponding animation. Functions can be displayed in algebraic, graphical, and tabular form, and students are often asked to tell stories that correspond to the functions (and animations). The software is meant to be used in what Dewey described as a cycle of “doing and undergoing” (Dewey, 1938). The program developers view student use of the software and teacher explanations and teacher-led discussions as complementary activities (Lobato, Clarke, & Ellis, 2005). They expect that students can learn more from teacher-led presentations and discussions when they have had direct experience with the software, as in a preparation for future learning paradigm (Bransford & Schwartz, 1999).

[----------INSERT FIGURE 1 ABOUT HERE----------]

In addition to proportional and linear functions, students and teachers can make piecewise linear functions, which can be used to model familiar situations. In the screen in the software shown in Figure 1, the graph plots position vs. time and the animation above displays the corresponding motion. Thus the multisegment line in the graph can represent the following story:

> Two girls were having a race. One girl ran at a constant speed to the finish line. The other girl started to run across the field but then realized she dropped her baton and stopped. She walked back to get the baton and then started in the right direction again and finally ran quickly to end the race in a tie.

In the studies reported here, we used many but not all of the core features of the MathWorlds software. Features not used included devices that can collect data from real physical motions and a classroom wireless network to organize social mathematical activities (Kaput & Hegedus, 2002). Although the software runs on computers or graphing calculators, we used computers because they are more available in middle schools.

A favorite slogan of the program founder, James J. Kaput, was that "new technology without new curriculum is not worth the silicon it's written in." This representational infrastructure is made usable in classrooms by specifying particular software documents (setups), paper curricular materials, and teacher professional development. Consequently, the interventions we tested consisted of an integration of technology and curriculum with supporting teacher training. The specific curricula used were based on lesson documents and packages previously tested in design research at the University of Massachusetts but reformulated to fit the specific needs of our target state, Texas. Although the program developers believe that strong classroom practices around mathematical argumentation would result in superior implementations, the research team did not expect to be able to change teachers' existing pedagogy in the short time available for summer workshops. Therefore, first year training workshops modelled appropriate pedagogy but did not seek to change classroom discourse practices. Second year training workshops included a limited focus on appropriate teaching moves.

Several studies conducted with subsets of the data considered here reached publication earlier than this article and are used to inform the presentation here. Roschelle, Shechtman et al. (2008) examined correlations between data in teacher logs and student outcomes. In particular, they found that teachers who more frequently reported teaching goals that were cognitively

complex had students who learned more. Pierson (2008) investigated classroom discourse in a set of classrooms for which we had videotapes of the exact same lesson. She found that teachers who were more responsive to student ideas and who presented students with more challenging mathematical tasks had students who learned more. Dunn (2009) investigated the train-the-trainer model in the Eighth-Grade Experiment. She found that the trainers were successful in accomplishing simple and limited goals such as introducing teachers to the software and curriculum workbooks and preparing them to teach using those materials. She found less influence from the training workshops on pedagogy, such as how much time teachers allowed students to use the software or how student work with the software was taken up in classroom discussion. Empson (2009) conducted in-depth case studies of three teachers and found that different teachers configured the available learning resources in quite different ways; there was not a single successful approach to enacting the SimCalc materials in a classroom. For example, in one classroom the software was a more prominent resource for students; in another, teacher-led discussion was a more prominent resource. These different configurations drew on teachers' strengths in different ways.

Research Design and Methods

The core research questions of the Scaling Up SimCalc Research Program were as follows:

1. Can a wide variety of teachers use an integration of technology, curriculum, and professional development to create new opportunities for middle school students to learn complex and conceptually difficult mathematics?
2. Can these findings be extended across grade levels?

3. Do student gains persist as we reduce the presence of the research and development team?

This article primarily focuses on the first two questions. The third question is addressed in part in the experiment that tests a train-the-trainers model because the R&D team has little direct contact with teachers in this model. The third question also highlights sustainability, which will addressed in a forthcoming article that tracked teachers for an additional year.

Experimental Design

To investigate all three research questions, we implemented two randomized experiments (one of which contained an embedded quasi-experiment) with pre/post measurement. The first experiment, the Seventh-Grade Experiment Year 1, began in summer 2005 with seventh-grade content, students, and teachers. The second, the Eighth-Grade Experiment, began in summer 2006 and was designed to extend the findings of the Seventh-Grade Experiment Year 1 to eighth-grade content, students, and teachers and investigate a train-the-trainers approach. Schools were randomly assigned to either a treatment or control group at the beginning of each study.

Whereas the Eighth-Grade Experiment lasted 1 year only, the Seventh-Grade Experiment lasted 2 years and followed a delayed-treatment design. The second year of the study afforded an embedded Seventh-Grade Quasi-Experiment in which control teachers (also called the delayed-treatment teachers) began to use the SimCalc replacement unit and treatment teachers (also called the immediate-treatment teachers) continued to use it. The staggered nature of this design enabled us first to compare between-teacher results obtained in Year 1 (i.e., immediate treatment versus delayed treatment). Then we could make a within-teacher quasi-experimental comparison

between classrooms of the delayed-treatment teachers in Year 1 with the classrooms of the same teachers in Year 2, when those teachers received the SimCalc replacement unit. The delayed-treatment design has been used successfully in schools, both in our pilot and by other investigators (Campbell, Shadish, & Cook, 2001; Slavin, 2002).

These studies were preceded by a pilot experiment that is reported in detail elsewhere (Tatar et al., 2008). We found an overall main effect in the pilot study such that students of treatment teachers had both higher posttest scores ($z = 2.95$, $p < .01$) and greater learning gains ($z = 3.49$, $p < .0001$) than students of control teachers.

Site Selection: Texas

This research took place in the state of Texas, which provided a good setting for a number of reasons. First, Texas is a large state with wide regional variations in the diversity of subpopulations of teachers and students. Given that a key goal of both studies was to evaluate the SimCalc approach with a wide variety of teachers in assorted demographic settings, this diversity was important. Second, we were able to partner with the Charles A. Dana Center at the University of Texas, which has both a strong interest in increasing the number of students who progress in mathematics to advanced placement (AP) calculus and a history of providing teacher professional development at large scale throughout the state. Through its highly regarded and extensive professional development programs for mathematics teachers in Texas, the Dana Center had the ability and credibility to recruit teachers throughout the state, to help the project address concerns that potential teachers and administrators might have about participation in the project, and to facilitate workshops and provide workshop components, including a train-the-trainer model. In addition, the Dana Center had already been promoting an aligned sequence of

instruction leading from middle school through AP calculus, and SimCalc naturally fit into this sequence. Third, Texas has an established, stable, and well-aligned system of standards and accountability. This enabled us to align the SimCalc curriculum with existing Texas curricula and standards and have confidence that the standards would not change midstream in our research and that conversations with teachers about curriculum would be consistent throughout the state. Fourth, Texas conducts a yearly census of teachers, schools, and districts. This allowed us to evaluate the properties of our sample relative to more general demographic information.

Components of the Treatment Interventions

In this description of the major components of the treatment, we begin with a discussion of the intervention logic and then describe the mathematical content, curricula, and teacher professional development. Note that in our experiments we did not evaluate the strength of the contribution from each of these components, but rather the impact of the intervention as a whole.

Intervention Logic: Replacement Units That Integrate Technology, Curriculum, and Teacher Professional Development

Cohen and colleagues (Cohen, Raudenbush, & Ball, 2003) have argued that the proper focus of research on scalable improvements to education is on instruction. Following their work, which also answers important historical critiques of attempts at educational improvement (e.g., Elmore, 1996), we viewed instruction as *the interaction among teachers and students around content in environments*. Our intervention logic was to seek improvement by providing as input to this system an integration of curriculum, software, and professional development.

Design research provided evidence that influenced the way we conceptualized our experiments and implemented the representational approach in the classroom. Overall, design

research rejects the idea that technology alone can have a robust effect on student learning. Instead, researchers recommend examining interventions that integrate multiple factors including pedagogy, curriculum, professional development, assessment, and school leadership (Roschelle, Pea et al., 2000). Design research points out that these elements are not truly separable in practice and further suggests that it may not be a high priority to separate them.

We conceptualized the SimCalc intervention as a replacement unit for several reasons. Prior large-scale research had recommended the replacement unit strategy (Cohen & Hill, 2001) because it balances the trade-offs between ambition and specificity. The goals of the research were inherently ambitious and so, too, was the use of the representational infrastructure. Replacement units were large enough and long enough to allow real change and meaningful learning consistent with these goals. At the same time, the short, contained nature of a replacement unit limited the perceived risks of the teachers and schools in participating, allowed us to provide explicit curricular content and pedagogical guidance and tight connections to existing standards, and enabled us to understand and manage the conditions of implementation.

Focal Mathematical Content: Multiple Perspectives

In preparing our experiments, we specified the mathematics that would be the focus of each intervention. Not only did we have to specify the focal content to be covered to both meet and exceed current mathematics instruction, but we also had to ensure that this was done in a way recognizable and consistent with state concerns, with research knowledge, and with best pedagogical practice. Coherence is increasingly seen as the most important quality of mathematical curricula (National Mathematics Advisory Panel, 2008). We started first by identifying the concepts at the intersection of the Texas seventh- and eighth-grade standards and

the capabilities of the SimCalc approach. This led to the identification of proportionality and linear function as our target mathematics.

Among middle school mathematical concepts, proportionality ranks high in importance, centrality, and difficulty (Hiebert & Behr, 1988; National Council of Teachers of Mathematics, 2000; Post, Cramer, Behr, Lesh, & Harel, 1993). For example, the National Council of Teachers of Mathematics (NCTM) describes proportionality and related concepts as "focal points" for learning in seventh and eighth grade (National Council of Teachers of Mathematics, 2007). From a mathematics perspective, proportionality is closely related to the important concepts of rate, linearity, slope, and covariation. In addition, proportionality offers an opportunity to introduce students to the concept of a function, through the constant of proportionality, k, that relates x and $f(x)$ in the functional equations of the form $f(x) = kx$. A deep understanding of the concept of function as it relates to rate, linearity, slope, and covarariation is central to progress in algebra and calculus. These concepts are also central to students' science learning. Without understanding rate and proportionality, students cannot master important topics and representations in high school science, such as laws (e.g., $F = ma$, $F = –kx$), graphs (e.g., of linear and piecewise linear functions), and tables (e.g., interpolating between explicit values relating the width and length of maple leaves). Mathematics education research has identified persistent difficulties in mastering these concepts and has theorized that proportionality is at the heart of the conceptually challenging shift from additive to multiplicative reasoning (Harel & Confrey, 1994; Leinhardt, Zaslavsky, & Stein, 1990; Vergnaud, 1988).

To further specify the target mathematics, the team examined textbooks used in the Texas state curricular standards, preexisting SimCalc materials, and the research literature. In conjunction with a mathematics advisory board including three mathematicians and three

mathematics educators, we developed a mathematics framework for the seventh- and then for eighth-grade intervention that abstracted the mathematical concepts (Table 1) to be used in the curricula and assessments. The SimCalc team noted that proportionality can be taught both as a formula ($\frac{a}{b} = \frac{c}{d}$) and a function $f(x) = kx$. The analysis of the latter function across algebraic, graphical, tabular, animated, and verbal forms can be the starting point for the learning progression that leads to calculus. In particular, emphasizing the conceptual links among different expressions of *rate* brings coherence to instruction that promotes an ever-deepening understanding of the mathematics of change and variation across many years of material. The particular opportunity in seventh-grade instruction is to connect the multiplicative constant *k* in the algebraic expression $y = kx$, the slope of a graphed line, the constant ratio of differences in a table comparing *y* and *x* values, and the experience of rate as *speed* in a motion. In eighth grade these connections expand to the more complex model implied by the linear function $y = mx + b$.

[----------INSERT TABLE 1 ABOUT HERE----------]

In this article, we use the symbol M_1 to refer to the mathematics that is measured on the tests used for accountability in Texas. This mathematics embodies a *formula* approach to proportionality and linearity and tends to ask students to find a number given two or three other numbers. We use the symbol M_2 to refer to mathematics that goes beyond what is tested in Texas. This mathematics embodies a *function* approach to proportionality and linear function and often asks students to consider the mapping between a domain and range and to connect such concepts as *rate* across multiple representations (e.g., *k*, in $y = kx$ and the slope in a graph of $y = kx$).

This discussion of overlapping but not entirely consistent mathematical perspectives is

important because our findings may appear different depending on which perspective a reader adopts. For example, a reader with a state standards/accountability perspective might come to a different view than a reader with a mathematics education research perspective. A teacher who wants to implement NCTM recommendations may come to a different conclusion than a teacher who wants her classroom's scores to improve on the existing state test.

Curricula

We designed two replacement units, one for the seventh grade and one for the eighth grade. Each unit covered the relevant mathematical content as outlined in Table 1. The materials for both units were student workbooks, a teacher's guide, and corresponding SimCalc MathWorlds files. The package was designed to be used daily over a 2- to 3-week period to meet all the requirements to cover an existing topic in the curriculum (i.e., rate and proportionality in seventh grade and linear function in eighth grade) while also introducing a more advanced perspective. The computer files configured the software to fit a particular lesson. Teachers were required to have access to computer laboratories or classroom computer sets, but students could share computers. Teachers could teach their unit by simply following the problems and questions posed in the workbook in the order given. These were not "scripted" curricula, but they did suggest movements between small group work, whole class discussion, and seat work. The teacher guides provided lesson plans that teachers could adapt and hints on possible student responses.

The seventh-grade curriculum, Managing the Soccer Team, addresses central concepts of proportionality: linear function, in the form $y = kx$, and rate. Speed as rate is developed through a sequence of increasingly complicated simulations. Lessons progress through representations—

from graphs, to tables, to equations—aiming to teach students to translate among all three and to connect each concept to verbal descriptions of motion or other real-world contexts. The unit's contextual theme is that students must serve as a soccer team manager—training players, ordering uniforms, planning trips to games, and negotiating their own salary.

The eighth-grade curriculum, Designing Cell Phone games, addresses linear function and average rate. Linear functions are developed as models of motion and accumulation. Students learn to use different representations of these functions for problem solving and to translate among the representations. Graphical representations are intended to enable students to efficiently solve traditionally difficult word problems about average rate. The unit's contextual theme is that students are designers of electronic games who must use mathematics to make the games functional.

Teacher Professional Development

For each of the studies, teachers were provided with professional development opportunities to strengthen their mathematical content knowledge, learn to use the curriculum materials, and/or plan specifically how to use the materials.

In all three studies, treatment teachers attended a 3-day summer workshop introducing the respective SimCalc replacement units. The teachers worked through the SimCalc materials as learners, experiencing a complete but compressed version of the entire unit. The workshop facilitators emphasized the mathematics in the replacement unit and the mathematics knowledge needed for teaching the unit. The facilitator also modelled best-practice pedagogical methods and drew attention to the techniques she used to prompt thorough exploration of mathematical ideas. Teachers had ample opportunity to practice using the software. Together, the facilitator and

participants also discussed potential classroom issues that might arise during the unit.

Treatment teachers also attended a 1-day workshop in the early fall in which they made specific plans for how and when to use the SimCalc materials in their classroom. Working primarily in pairs, the teachers wrote lesson plans and thought through their own logic for the unit.

In addition in the Seventh-Grade Experiment Year 1, before the 3-day SimCalc material workshop, treatment teachers attended a 2-day workshop called TEXTEAMS, which addressed the mathematical knowledge for teaching rate and proportionality. This workshop is described in more detail in the section below on the design of the counterfactuals.

To investigate whether student gains would persist as we reduced the presence of the research and development team, we used two different teacher professional development delivery models. For the Seventh Grade studies, the training in implementing the SimCalc replacement unit remained relatively tightly controlled across the various Texas regions. Two members of the SimCalc team—both highly experienced mathematics teacher educators—led all the professional development workshops. This model was intended to ensure consistency and quality of delivery across all the workshops. In the Eighth-Grade Experiment, we used a train-the-trainers model. The SimCalc team trained six teacher educators from five regions of Texas in a 2-day workshop. The workshop covered the learning goals, the MathWorlds software, and the curriculum workbooks. Each of these participants then returned to their home region and, at a later date, facilitated a 3-day summer workshop for teachers participating in this study. As a dissertation describes in detail (Dunn, 2009), all workshops introduced teachers to the SimCalc software and curriculum units but the pedagogical content of these workshops for teachers varied . In subsequent observational case studies of classroom implementation, Dunn found that while

the workshops were sufficient to enable teachers to teach with the SimCalc materials, the workshops did not exert much influence over teachers' existing pedagogy.

Design of the Counterfactuals

In an experimental design, the counterfactual (control condition) must be designed to allow causal inferences to be made about the impact of the treatment intervention. In these experiments, a primary goal was to investigate whether the SimCalc intervention was the causal factor in enabling a wide variety of students to learn more advanced (M_2) mathematics while maintaining gains on more basic (M_1) mathematics. Thus, the design of the counterfactual conditions needed to reduce the plausibility of alternative explanations for the cause of any differential learning gains that we might observe across groups. The single most important threat to internal validity would be the presence of a confound—some other contemporaneous but unrelated circumstance that caused the measured growth in student learning. Because of the integrated nature of our intervention, we designed the counterfactuals to encompass both curriculum implementation and teacher professional development.

For curriculum implementation, because our unit of instruction was a replacement unit (e.g., Cohen et al., 2003), the most natural counterfactual was the curriculum that was replaced—the business as usual curriculum. In both the Seventh- and Eighth-Grade Experiments, the business as usual curriculum addressed, within the same time frame as the SimCalc unit, similar basic concepts (M_1) but provided less coverage of more complex concepts (M_2). Using this as a counterfactual enabled us to examine the learning that took place when the opportunity to learn M_2 concepts was provided through an integration of technology, curriculum and professional development. Also, in the Seventh-Grade Studies, as described below, teachers in both the

treatment and control groups had equal access to TEXTEAMS materials that they could use to supplement their units to teach M_2 concepts. Thus the contrast is between (a) an integrated SimCalc intervention and (b) business-as-usual curriculum with similar M_1 coverage supplemented with teacher professional development and materials that provided M_2 content.

The teacher professional development components were designed to address possible confounds such as the degree to which teachers might differentially across groups feel part of a new and special project, believe in the usefulness of the intervention, find the amount of work required for participation and compensation for that work acceptable, have an opportunity to interact with colleagues, or feel supported by the research team. In line with these considerations, we designed the professional development in the control interventions to parallel that in the treatment interventions along these dimensions. In the Seventh Grade Experiment, we chose a workshop called TEXTEAMS, which was developed by our partners at the Dana Center and was highly regarded in Texas. The workshop introduced both M_1 and M_2 components of rate and proportionality, and provided activities that teachers could take back into the classroom and use with their students. Both immediate and delayed treatment teachers in Year 1 received the same 2-day TEXTEAMS workshop as those in the treatment group but not the rest of the treatment professional development specific to SimCalc. In the Eighth-Grade Experiment, instead of the treatment professional development, control teachers received a 3-day summer workshop on teaching statistics called Teaching Mathematics TEKS (Texas Essential Knowledge and Skills) through Technology (TMT3). The workshop, also developed and delivered by the Dana Center, was recognized in Texas as a high-quality offering and allowed participants to learn to use technology to support student learning of statistics.

Furthermore, in the Seventh-Grade Studies, the delayed-treatment design supported equal

treatment and equal engagement among all teachers because delayed-treatment teachers knew that eventually, in Year 2, they would receive the SimCalc intervention.

Note that the emphasis in this design was on whether the intervention was effective. The counterfactual was not designed to enable us to isolate and tease apart the impact of technology as separate from curriculum. For example, one possible counterfactual might have been a SimCalc intervention without the software. We did not conduct this comparison because our intervention is fundamentally an integrated system that would have lost integrity had we removed the technology. Moreover, the counterfactual was not designed to determine whether the SimCalc intervention might be more effective than a paper-based curriculum covering similar content. Another possible counterfactual could have been non-SimCalc paper materials that addressed similar complex mathematics, written by either a third party or our own team. We could not find suitable materials and worried that if were the designers of the paper-only materials and found results in favor of the integrated technology condition, reviewers could easily argue that we purposely handicapped the paper materials. Thus we decide to leave it to further experimentation to examine whether alternative, better, or cheaper ways exist to achieve the same goal.

Table 2 summarizes key contrasts between the Seventh-Grade and Eighth-Grade Studies.

[----------INSERT TABLE 2 ABOUT HERE----------]

Assessment Design and Development

Because student achievement is the primary dependent measure for all the studies, we directed attention and resources to developing assessments that would meet rigorous standards for validity. We found that standardized tests (such as the TAKS [Texas Assessment of

Knowledge and Skills]) did not capture the conceptual depth students could reach using the SimCalc technology and curricula and thus using such tests for outcome measures would cause us to overlook potentially important impacts of the intervention. Thus, the research team decided to build its own assessments. We developed two, one for the Seventh Grade studies focusing on rate and proportionality and one for the Eighth-Grade Experiment focusing on linear function. Within each study, the identical assessment was administered at pretest and posttest.

To develop valid and reliable assessments, we followed models of best practices in assessment development (e.g., AERA, APA, NCME, 1999) and drew on the tenants of Evidence Centered Design (ECD; Almond, Steinberg, & Mislevy, 2002; Mislevy, Almond, & Lukas, 2003; Mislevy, Steinberg, & Almond, 2002).The ECD framework emphasizes the evidentiary base for specifying coherent, logical relationships among all essential assessment elements. Our assessment development process had three essential stages, as follows.

In the first stage, we established a conceptual assessment framework and assessment blueprint. The blueprint had four dimensions: (1) complete coverage of all the M_1 and M_2 topics with subscales for each (see Table 1), (2) alignment with the state content standards (the TEKS), (3) various problem contexts (i.e., motion and money), and (4) a diversity of task types (about one third each of multiple choice, short response, construction of multiple mathematical representations).

In the second stage, we developed a pool of assessment items. Using the blueprint as a guide to ensure coverage of all relevant concepts, the team drew from the instrument used in the pilot study, surveyed existing standardized tests (TAKS, NAEP, TIMSS, and other state tests) and literature for items, and created some new items. For example, on the Seventh-Grade assessment, one M_1 item asks: “If 2 / 25 = n⁄500, what is the value of n?” One of the M_2 item

asks student to construct three different representations (table, algebraic expression, and graph) of a proportional relationship between price and number of tickets purchased for a raffle.

In the third stage, we validated and refined the assessment items using empirical methods. First, we held a *formative expert panel* to review and rate items for alignment with our conceptual framework (Table 1), alignment with TEKS, and grade-level appropriateness. Members of the review panel were mathematicians, mathematics education researchers, and mathematics educators working in Texas. We used the ratings to select and refine appropriate items. Second, we conducted *student cognitive think-alouds* to obtain information about how individual students would solve the problems. This information enabled us to eliminate or revise questions that were ambiguous or that did not require the target skills to reach an appropriate answer. Third, we *tested the items in the field*. Using the refined items, we created a prototype test that satisfied all the constraints of our original blueprint. We tested the seventh-grade instrument in the field with a sample of 230 sixth- and seventh-grade students and the eighth-grade instrument in the field with a sample of 309 eighth-grade students. We used both classical test theory and item response theory to characterize the technical qualities of the items. Fourth, we held a *summative expert panel review* in which two mathematics education experts who had been part of the original panel provided summative feedback on the revised items. For each item, the experts assessed the content alignment ratings made by the formative panel and, as necessary, recommended refinements to the items for better alignment with the content framework.

The basic test specifications of the resulting assessments were as follows. The Seventh Grade rate and proportionality assessment had 30 items with an alpha of 0.86. The M_1 subscale had 11 items with an alpha of 0.73, and the M_2 subscale had 19 items and an alpha of 0.82. The

Eighth Grade linear function assessment had 36 items with an alpha of 0.91. The M_1 subscale had 18 items with an alpha of 0.79, and the M_2 subscale had 18 items and an alpha of 0.87.

In addition, assessment administration procedures were established to minimize the possibility that teaching to the test could be a substantial confound in the interpretation of any findings. Teachers were not explicitly shown the instrument at any time, and each classroom set of assessments was provided in a sealed envelope with specific instructions to open the envelope only at the time of administration. Furthermore, there were no accountability pressures that might motivate teachers to deviate from the procedures or teach to the test. While some teachers may nonetheless be predisposed on their own to teach to the test, given random assignment, there was no reason that this predisposition would be more likely in either experimental group.

Demographic and Implementation Measures

To investigate our research question about robustness across diverse settings and to help contextualize any findings for mathematics learning, we collected data on student demographics and classroom implementation. Before teaching their units, teachers were asked to fill out a roster of the students in their classroom. For each student, teachers reported gender, ethnicity, and their subjective rating of the student's prior achievement level as low, medium, or high. To allow this variable to reflect realistically teachers' perceptions of students, teachers were left to determine their own criteria for their ratings of student achievement level. For each day the unit was taught, the teacher filled out a log page probing various aspects of implementation (e.g., pages covered in the workbook, mathematical topics covered, whether class was conducted in the classroom or the computer lab). In addition, school-level data were obtained through the Public Education Information Management System (PEIMS), a publicly available database

maintained and distributed by the Texas Education Agency, the state department of education.

We measured several other variables, which are reported on elsewhere. Through implementation logs, surveys, and phone interviews, we measured many attributes of teachers (including professional background, beliefs, attitudes, teaching goals, and mathematics knowledge) and their experiences in the program. We also collected rich qualitative data through interviews with students and classroom observations to understand more fully student learning and variations in classroom implementation.

Analysis Methods and Procedures

These studies sampled intact classrooms (clusters of students), meaning that classic statistical models such as the *t* test or multiple regression models would be inappropriate without modification. In particular, with cluster sampling, standard models tend to underestimate the standard errors of key statistics and overestimate the statistical significance of results. These models can be corrected in several ways to account for the clustering of students within classrooms. In this study, we used multilevel modelling (MLM), specifically hierarchical linear modelling, to estimate the effects of the treatment (Raudenbush & Bryk, 2002). MLM accounts for measurement and sampling error at both the student and classroom level, resulting in correctly adjusted standard errors for the treatment effect.

To model or conduct significance testing for our student achievement variables, one would normally fit two-level MLM models (students nested within classrooms, classrooms nested within schools), with the student scores as the outcome variable and the treatment group as a predictor at the school level. The school level is necessary to account for the proper degrees of freedom in the test statistics, since schools are the primary sampling unit. Within schools,

students are not independent but share common classroom characteristics with one another. In our case, however, over 70% of schools had only one teacher, rendering any estimate of the level 2 variance components unreliable. We therefore collapsed our models to two levels (students nested within schools).

In analyses of student achievement, one important decision is whether to focus on student gain scores (i.e., pretest score subtracted from posttest score) or posttest scores adjusted for pretests. In a randomized experiment, both methods yield unbiased estimates of the treatment effect (Maris, 1998). One analysis expressed the choice as follows:

> In their tribute to Fredric Lord, Holland and Rubin (1983) noted that the basis for Lord's Paradox is that an analysis of difference scores and an analysis of covariance are designed to answer different questions. An analysis of difference scores answers the question about whether students changed from the pretest to the posttest, whereas an analysis of covariance answers the question of whether students who have the same pretest scores will have different posttest scores. These are not the same questions and it is unrealistic to expect them to provide the same answers. (Campbell & Kenney, 1999)

Because the treatment condition was randomly assigned (and therefore expected to have zero covariance with other predictors), either choice would be expected to yield an unbiased estimate of the treatment effect. However, when pretest scores potentially covary with other predictors (e.g., student gender), estimates of the impact of those predictors will be biased. In our analyses, we wanted to examine the gains of students across a variety of categories such as gender, and pretest scores were significantly correlated with many such student categories. In our research, the goal was to answer the question of whether students' performance changed, so the

use of gain scores was more appropriate.

We constructed a multilevel model as follows. The first level predicted student gain scores as a function of a school-specific intercept and *P* student level covariates.

Level 1 (Student): $Y_{ij} = \beta_{0j} + \sum_{p \in P} \beta_{pj} X_{ij}^{(p)} + r_{ij}$

At level 2, the school-specific intercept was modelled as the sum of a grand mean, a fixed effect for treatment assignment T_j, *Q* school-level covariates and a random deviation.

Level 2 (School): $\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \sum_{q \in Q} \gamma_{0q} W_j^{(q)} + u_{0j}$

As it turns out, tests for random slopes for all student-level covariates were non-significant, so all β_{pj} in the Level 1 equation are modelled as fixed effects (set equal to the corresponding γ_{p0}).

All models were fit using the *xtmixed* procedure within Stata version 9 and restricted maximum likelihood estimation. Continuous covariates were grand-mean centered, whereas categorical variables were represented as 0/1 indicators. In testing the impact of mediating variables (i.e., student gender, student ethnicity, teachers' ratings of student prior achievement levels, location in Region 1, and percentage of students in the school receiving free or reduced-price lunch), we fit multiple models, each adding a single fixed covariate (at the student or school level) and interaction with the treatment indicator to the model. We managed the risk of inflated Type I error rates by using the false discovery rate procedure of Benjamini and Hochberg (1995). This procedure ensures that fewer than 5% of the reported statistically significant results within a logical family of comparisons will be due to Type I error.

Recruitment and Assignment to Condition

As discussed above, we recruited through the Dana Center and regional Education Service Centers (ESCs) throughout Texas. ESCs are public organizations (affiliated with the Texas Education Agency) that provide supports for schools and districts in their region. By working with the Dana Center and with ESCs, the SimCalc project team could use the existing network of professional development service providers with strong connections to teachers and a positive track record in the eyes of Texas teachers.

In both studies, we sought to recruit teacher volunteers whose students reflected the regional, ethnic, and socioeconomic diversity of the state. In sampling broadly, to overcome a possible threat to validity that might have resulted from biased attention to the schools the ESCs already had relationships with, the ESCs were given a protocol and instructions about how to approach districts, school mathematics coordinators, principals, and teachers in an unbiased, systematic way. ESCs were instructed to recruit as many applicants as possible to ensure a diverse sample of teachers and students. A school had to meet two requirements for its teachers to be invited to participate: it had to have enough computers (we could not afford to buy computers for schools) and its leadership needed to give consent for teachers to participate.

We performed selection and random assignment at the school level; that is, if we accepted one mathematics teacher from a school, we would accept all applicant mathematics teachers from that school and assign them all to the same condition. There were three rationales for this. First, best practices in professional development provide teachers with an in-school community to integrate new materials and practices. Second, with respect to the research design, we sought to avoid problems of cross-contamination among groups by having teachers who work closely together communicate about interventions in the different groups. Third, the majority of

schools had only one teacher anyway (77% of the 73 schools in the Seventh Grade studies and 72% of the 56 schools in the Eighth-Grade Experiment).

Once a pool of applicants was generated, we randomly selected teachers by (1) creating a randomly ordered list of all applicant schools and (2) selecting schools from this list, alternating assignment to the treatment group and the control group until we met our quota for sample size in each group (140 in the Seventh-Grade Experiment; 80 in the Eighth Grade Study).

We decided not to recruit for the Eighth-Grade Experiment in schools already participating in the Seventh-Grade Experiment; therefore, none of the students participating in the eighth-grade SimCalc replacement unit had studied the seventh-grade unit in the same school.

Participants

The Appendix shows the sample characteristics, illustrating the diversity of regions, teacher demographics, and student demographics. A technical report (Tatar & Stroter, 2009) examined the diversity of the seventh- and eighth-grade samples, as well as their representativeness relative to broader populations. The samples were diverse in terms of campus poverty levels, school size, and campus ethnicity. They were also diverse in terms of teachers' gender, ethnicity, years of teaching experience, highest degree obtained, and mathematical knowledge. Comparisons were made to the population in the Texas regions in which the experiments were conducted, as well as to the state of Texas as a whole. For all variables for which we had data at the regional and state levels, the ranges and means were similar among our samples and the middle school mathematics teaching population by region and in the state. Note that the low percentages of African-American teachers and students, as well as schools from

large urban settings, reflect their small populations in the regions in which the experiments were conducted. Further studies may be needed to examine generalizability of findings to those populations.

Whereas seven of the 20 geographical regions in Texas participated in the studies, of particular note is the participation of Region 1 because of its unique demographic and socioeconomic characteristics. Region 1 is in the Rio Grande Valley adjacent to the Mexican border. It has one of the highest poverty levels in the United States and is predominantly Hispanic. Region 1 participated in the Seventh-Grade Studies; however, because of a shift in local circumstances in the year between recruitment for the two experiments (i.e., the region received a large grant for a major reform in mathematics instruction), the region did not participate in the Eighth-Grade Experiment.

While there was overall attrition in each of the studies, there is evidence that attrition was not differential across experimental groups. In the Seventh-Grade Studies, 140 teachers were accepted into the study, 117 attended the workshop (16% attrition), 95 teachers completed Year 1 (23% attrition), and 67 teachers completed Year 2 (29% attrition). When asked why they dropped from the program, teachers reported reasons that were not related to the project itself (e.g., reassignment or promotion, personal reasons, relocation). In the sample of 95 classrooms that completed the Year 1 experiment, there were no statistically significant differences between groups on any of the student, teacher, or school level variables we examined.

For the Seventh-Grade Quasi-Experiment, we considered data from only the 30 delayed-treatment teachers who finished both Year 1 and Year 2. In a quasi-experiment in which participants are not randomly assigned to treatment groups, the primary internal threat to validity is the possibility of nonequivalence of groups, which we examined between years at the student

level (the teachers and schools were the same each year). The groups were equivalent with respect to all variables except gender. This difference is not a strong threat to the validity of the study; as shown below, the baseline assessment scores were equivalent across the groups, and gender was not shown to significantly predict student learning.

In the Eighth-Grade Experiment, 88 teachers were accepted into the study, 63 attended the workshop (28% attrition), 56 teachers completed the study (11% attrition). In the sample of 56 classrooms that completed the study, the treatment and control groups were equivalent on all variables except student ethnicity, in which there was a higher percentage of Hispanic students in the treatment group. This small difference is not a strong threat to the internal validity of the study; again, as shown below, the baseline assessment scores were equivalent across the groups, and ethnicity did not significantly predict student learning in either experiment. Also, in the Eighth-Grade Experiment, the greater number of teachers in the treatment group was an artifact of teachers' scheduling conflicts with the workshops to which they were assigned. Because teachers were not informed about the workshop type until the workshop actually occurred, the consequences for randomization and thus the validity of the experiment are minimal.

Experimental Procedure

In each study, we used tightly controlled experimental procedures to minimize the possibility of bias across groups. We designed the treatment and control procedures to be almost identical with the exception of which unit was implemented. Each year, teachers attended their designated workshop(s) at their regional ESC. To ensure that they all had a consistent understanding of the research, all teachers were shown a video at the beginning of the summer workshop that explained the research project and procedures. Early in the school year, teachers

received a package that contained classroom sets of relevant instructional materials and assessments, a logbook, and supplies to mail the completed materials back to the research team. To establish the focal instructional unit, teachers were asked to determine which unit in their curriculum was most pertinent to the mathematical content of the replacement unit. In the treatment groups, teachers were asked to replace that unit with the SimCalc replacement unit. In the control groups, teachers were asked to teach the unit as they usually would. To establish a target classroom for the research, the research team randomly assigned a period number to each teacher. In the log and follow-up interviews, teachers reported that they actually did collect data with their selected target class (i.e., rather than selecting their own).Teachers administered the student assessments immediately before teaching the unit and immediately after teaching it. Each day of the unit, they filled out a page in the logbook. After completion of the postunit assessment, teachers returned the assessments and daily logs to the research team via mail.

Within each experiment, teachers received the same stipend regardless of which condition they participated in. In the Seventh-Grade Studies, teachers received a stipend of $1,000 per year, and in the Eighth-Grade Experiment, teachers received a stipend of $500 for the year.

Results

The results are reported in three sections: (1) the main effects of the treatment, (2) the robustness of the main effects across participant groups, and (3) implementation variables.

Main Effects of Treatment

Two-level MLM analyses were used in all three studies to show that the main effect was statistically significant, demonstrating that students who had the SimCalc intervention learned

more than control students who had the business as usual curricula. Table 3 shows that in all three studies, although the treatment and control groups began with similar pretest scores, treatment students had significantly higher gains from pretest to posttest. In all three studies, the effect sizes were large and educationally significant, particularly for the M_2 portion of the tests. As Figure 2 illustrates, the gains differences between the two groups in all three studies occurred mostly on the M_2 portion of the tests.

[----------INSERT TABLE 3 ABOUT HERE----------]

[----------FIGURE 2 ABOUT HERE----------]

Robustness of Learning Gains Across Participant Groups

To what extent are these findings robust across subpopulations and settings? To address this question, we examined whether the intervention was effective across five policy-relevant demographic factors: student gender, student ethnicity, teachers' ratings of student prior achievement levels, whether the school was located in Region 1, and percentage of students in the school receiving free or reduced-price lunch..

We began by first examining the extent to which some students in these groups may have begun at a relative disadvantage. Within each study, we ran a series of two-level MLM models predicting student M_2 pretest scores, one for each of the five demographic factors, entering the factor independently as a covariate at the appropriate level. Overall, we found that all of the factors, except being located in Region 1, significantly predicted M_2 pretest in all three studies at a significance level of $p < 0.01$ or lower, indicating baseline disadvantages for traditionally underserved populations. Specifically, girls started lower than boys, Hispanic students started lower than other students, students rated as low or high achieving by their teachers started lower

or higher respectively than those rated as medium achieving, and the higher the percentage of students qualifying for lunch programs in the school, the lower the pretest scores.

We then examined the extent to which students in these groups may have had differential gains. Within each study, we ran a series of two-level MLM models predicting student M_2 gain scores, one for each of the five demographic factors, entering the factor independently as a covariate at the appropriate level. These models also included as covariates an indicator for the experimental group and the factor by group interaction.

Table 4 and Figure 3 summarize the gain models. In the two main experiments, population factors did not predict student learning gains except for those students rated as low achievers. However, in the Seventh-Grade Quasi-Experiment, ethnicity, region, and percentage receiving free or reduced-price lunch in the school negatively predicted learning gains. The specific findings were as follows:

1. Student gender. Whereas boys started out with higher pretest scores, there were no main effects or interactions for the learning gains.

2. Student ethnicity. Although Hispanic students started out with lower pretest scores, there were no main effects or interactions for learning gains in the two main experiments. In the Seventh-Grade Quasi-Experiment, however, there was an interaction such that Hispanic students using SimCalc in Year 2 had lower learning gains than their non-Hispanic counterparts.

3. Teachers' ratings of student prior achievement levels (low, medium, and high). In all three studies, students at all three achievement levels gained more in the SimCalc replacement unit than their peers studying the ordinary curriculum; however, there were

also interactions in the seventh-grade studies (but not the Eighth-Grade Experiment) such that students in the SimCalc replacement units rated as low had lower gain scores than students rated as medium or high.

4. Region 1 (Seventh-Grade studies only). In the Year 1 experiment, there was no main effect and no interaction. In the Seventh-Grade Quasi-Experiment, however, there was an interaction such that Region 1 students using business as usual curriculum in Year 1 had higher learning gains than their counterparts in other regions, and students using SimCalc in Year 2 had lower learning gains than their counterparts in other regions.

5. Percentage receiving free or reduced-price lunch. Although this variable was a strong negative predictor of pretest scores, there was no main effect or interaction for the learning gains in the main experiments. In the Seventh-Grade Quasi-Experiment, however, there was an interaction such that in Year 2 when students used SimCalc, this variable was a negative predictor of learning gains.

[----------INSERT TABLE 4 ABOUT HERE----------]

[----------INSERT FIGURE 3 ABOUT HERE----------]

Implementation Variables

We collected a large set of implementation measures. We created a correlation matrix and found that most variables had either no correlation to student outcomes, or had inconsistent correlation to student outcomes across studies. Because presenting this data and discussing possible explanations of the patterns does not bear on the main effects reported herein, we plan to report this information in a subsequent technical report. We have chosen to report herein only three variables, each of which bears on SimCalc program theory: use of technology, days spent

on the unit, and topic coverage. By considering these variables, we are in a better position to ask whether the intervention worked for the reasons its developers espoused.

Use of Technology

Syntheses of prior research support the role of technology-based representations in student learning of cognitively demanding mathematics, especially relative to conceptual understanding (Heid & Blume, 2008). It was not a purpose of these experiments to isolate the contribution of technology to student learning but to consider whether implementation data were consistent with prior research on the role of technology.

We created a proxy measure for the amount of computer use by counting the days students spent in the computer lab as recorded by teachers in their daily log. This is admittedly a coarse approximation of actual computer use. Yet case studies and follow-up interviews confirmed that teachers used the time in the computer lab to engage students in using the MathWorlds software. Our computer use measure was collected for only the two main experiments. As would be expected, students in the SimCalc replacement unit spent much more time in the computer lab than those in the control in both the Seventh-Grade Experiment Year 1 [an average of 41.5% and 3.5% of the days in the treatment and control groups, respectively; $t(93) = 8.0$, $p < .0001$] and the Eighth-Grade Experiment [an average of 72.8% and 6.6%, respectively; $t(54) = 9.4$, $p < .0001$]. Further, the amount of time spent in the computer lab was a predictor of student learning gains in the Eighth-Grade Experiment ($z = 2.5$, $p < .05$ for the interaction term of days in computer lab by condition). However, this finding was not replicated in the Seventh-Grade Experiment ($z = 0.24$, $p = .81$, n.s.). Technology use is an important aspect of SimCalc program theory, and our findings are consistent with this feature of the theory.

Days Spent on the Unit

By considering the number of days teachers spent on the unit and their self-report of topic coverage, we can ask whether student learning gains with the SimCalc intervention followed merely from spending more time on the content or whether it resulted from using time more effectively. We measured the number of days teachers spent on the unit by triangulating among three data sources: a calendar they used to mark the days they taught the unit, the number of log pages they filled out, and the dates the pretest and posttest were administered.

Across the studies, differences between groups were small and inconsistent. In the Seventh-Grade Experiment Year 1, immediate-treatment teachers spent a mean of 14.9 days (*SD* = 8.6) teaching the replacement unit and delayed treatment spent a mean of 12.0 days (*SD* = 4.6) on their business as usual curriculum. The difference between groups was significant ($t(93) = 2.0$, $p < .05$). When one outlier teacher (who reported spending 66 days on the SimCalc unit) was dropped from this analysis, the mean difference between groups dropped from 2.9 to 1.8, but was still significant at the $p = .05$ level. In the Seventh-Grade Quasi-Experiment, there was a nonsignificant trend that the delayed treatment teachers spent less time teaching the SimCalc unit in Year 2 (10.9 days, *SD* = 2.9) than they had spent teaching their business as usual curriculum in Year 1 (12.0 days, *SD* = 4.6; $t(29) = 1.2$, $p = .23$, ns). In the Eighth-Grade Experiment, there was also a nonsignificant trend that teachers spent less time teaching the SimCalc replacement unit (12.4 days, *SD* = 4.1) than their counterparts spent teaching the business as usual curriculum (15.2 days, *SD* = 7.3; $t(54) = 1.9$, $p = .07$, ns).

Also, across all three studies, there was not a significant correlation between the number of days spent teaching the unit and M_2 gains. These findings provide evidence against a claim that time on task explains the main effects.

Topic Coverage

We also asked teachers to report the topics they covered each day in their logbook. These topics were aligned with the topics covered in the curriculum and assessment, as outlined in Table 1. As appropriate to each study, in the Seventh-Grade Studies, teachers were given a list of 12 topics, and in the Eighth Grade study, teachers were given a list of 5 topics. Specifically, teachers were asked to answer the question, *To what extent did your class focus on the following topics?* Teachers checked boxes on a 4-point Likert scale ranging from (*1*) *not at all* to (*4*) *a major focus*. We considered teachers as covering a topic in a given day if they selected a 3 or 4 on the scale. Teachers were not limited in the number of topics they could rate as a 3 or 4. We then counted the number of days a teacher covered each topic.

In the Seventh-Grade Studies, we saw a pattern of results that fits the program theory (Figure 4). First, teachers in both groups reported spending many days on basic operations (e.g., how to calculate; how to do basic arithmetic), and the only significant difference was that teachers spent more days on this in Year 1 of the quasi-experiment when they did their business as usual curriculum. This suggests that for the most part the material did not substitute advanced for basic mathematics; basic mathematics was still a major focus. Second, note the shift from emphasizing $a/b = c/d$ to $y = kx$ that was significant in the experiment and a strong trend in the quasi-experiment. Accomplishing this shift was a major intention of SimCalc program, yet *all teachers* were encouraged to make this shift in the preliminary TEXTEAMS workshop they attended. It appears that the SimCalc program helped the teachers implement the desired shift. Further, when teacher used the SimCalc intervention they reported considerably more coverage of topics that are signature aspects of the SimCalc unit—mainly the use of multiple representations and reasoning comparatively about more than one function. These findings show

that when teachers used the SimCalc intervention they interacted with the SimCalc materials in a way that concentrated on more advanced topics without neglecting basic operations.

[----------INSERT FIGURE 4 ABOUT HERE----------]

We found a similar pattern of results in the Eighth-Grade Experiment. Teachers in both groups reported spending about the same number of days on categorizing functions as proportional, linear, or nonlinear (Figure 5).They also reported similar time on using algebraic, tabular, or graphical representations of linear functions, for example, to find unknowns, identify a point on a graph, or add a missing value to a table. In addition, they reported similar time on translating across representations. These findings support the theory that the SimCalc materials afforded teachers an opportunity to concentrate on more advanced topics without neglecting the core of the state standards.

[----------INSERT FIGURE 5 ABOUT HERE----------]

Discussion

In these two randomized experiments and a quasi-experiment, we found a causal relationship between classroom implementation of a SimCalc replacement unit and student learning of more advanced mathematics. Several findings held true across all studies. SimCalc students learned advanced aspects of the target mathematics concepts (M_2) without sacrificing gains on the mathematics measured by the state test, in two studies without spending more time on the material. Indeed, for the simpler aspects of the target concepts (M_1), students of teachers who used the SimCalc replacement unit showed a trend toward greater gains that was nonsignificant in the two experiments and statistically significant in the quasi-experiment. These findings are consistent with the SimCalc program philosophy of increasing opportunities to learn

advanced mathematics within the context of the topics already included in the curriculum.

The counterfactuals in the two experiments and quasi-experiment were designed to minimize threats to the internal validity of the study and enable causal attributions of the effects on student learning due to implementation of the SimCalc intervention (Campbell et al., 2001). Because the seventh-grade studies provided all teachers with equal information about the target student learning outcomes (i.e., the functional approach to rate and proportionality), provided all teachers with materials they could use to address these outcomes (SimCalc or TEXTEAMS), and measured student learning before and after the classrooms covered relevant content, we reduced the possibility that differences in teacher awareness of instructional goals or lack of availability of relevant materials confound the findings. In particular, Texas educators consider the TEXTEAMS workshop to be useful materials for enabling teachers to teach the measured mathematics concepts. Note that our study is not a comparison of SimCalc and TEXTEAMS because we provided all teachers with the same TEXTEAMS workshop and materials; the experimental group received the TEXTEAMS training immediately before receiving the SimCalc training. Because the Eighth-Grade Experiment provided equal-duration summer workshops for all teachers, introduced all teachers to representational technology, and emphasized advanced mathematics, we reduced the possibility that differences in the duration of professional development confound the findings; the specific SimCalc materials supplied to teachers for classroom use are implicated in the observed effects.

Another possible threat to internal validity might have been the Hawthorne effect. Brown (1992) describes the Hawthorne effect as the conjecture that any intervention tends to have positive effects merely because of the attention participants get from the researchers. If one group feels particularly singled out in a positive or negative way, its performance may be

influenced. To counter the possibility of the Hawthorne Effect, we were careful to design counterfactuals to give teachers high-quality professional development experiences and materials. Further, we designed all presentations and materials for teachers to emphasize equally the importance of participation in this research.

Note that in the literature, pilot study, and subsequent data collection in the Seventh-Grade Studies, we found no evidence that suggested that teachers would favor one condition over the other. Mathematics teachers are less likely than any other teachers to use computing technology in instruction (Becker & Anderson, 1998). Computers are widely seen as too difficult to use, not worth the time commitment, and even extraneous to "real" mathematics. In the phone interviews we conducted during the pilot study, many control teachers expressed relief that they were not in the 5-day workshop and did not have to bother with technology and the computer lab. Yet in interviews in the pilot with teachers about their experience as being part of the control or the treatment group, we found no substantial differences. This increased confidence that we would not be testing a differential effect of attractiveness or unattractiveness between groups in the full experiment. In addition, close coupling of the cognitive demands of the SimCalc intervention and student assessment would argue against the Hawthorne effect. As Brown (1992) points out, a true Hawthorne effect is a concern only if the outcome measure is extremely general, such as "feeling in control," and not tightly aligned with the intervention's cognitive goals.

A possible threat to external validity and generalizability is that our outcome measures were developed within the project to be aligned with the project goals. Methodologically, we chose to develop our own assessment because the Texas state test would not have assessed knowledge of the target M_2 content. While there is a danger of overalignment between our

intervention and measures, as we described in our assessment development process, we have been explicit in the development of our conceptual assessment framework, vetting the content with experts in the field and collecting several sources of empirical evidence (expert panel review, cognitive think-alouds, field testing) to support the validity of our assessment argument. While we may not have examined outcomes with other instruments, we have extensive empirical support for the specification of the knowledge, skills, and abilities that were tested.

Finally, we see the consistent replication of our experiments with variations in sample and setting (a wide variety of teachers and schools around the state of Texas), treatments (replacement units), and outcomes (assessments) as good cause to reject the idea that the findings result from experimental artifacts.

In addition, we found the main effects to be robust. Our sample included the more cosmopolitan Dallas-Fort Worth and Austin areas as well as the uniquely Texan western and border regions of the state. Schools within these regions varied in poverty and prior achievement levels. Within those schools were teachers with different backgrounds, practices, beliefs, and attitudes. And within the schools, were boys and girls who came from White, Hispanic, and other ethnic backgrounds and had different levels of prior achievement.

In all comparisons in the Seventh- and Eighth-Grade Studies, we found that while gender, ethnicity, and socioeconomic status were associated with students' baseline test scores, learning gains were equitable across all subpopulations. In the Seventh-Grade Quasi-Experiment, however, ethnicity, region, and socioeconomic status were associated with learning gains. An important shift in the population occurred in the Seventh-Grade Quasi-Experiment; many teachers in Region 1 dropped out. While other poor and Hispanic campuses remained in the study, these campuses may differ from the campuses in Region 1. Another possible explanation

is suggested by teacher interviews: after teaching the unit a first time, teachers reported a belief that it was more appropriate for high achieving students (a belief which is not supported by our data). Teachers in the quasi-experiment were teaching with SimCalc a second time and may have oriented their teaching away from traditionally underachieving students.

As in any experiment, these findings should be interpreted with caution. First, the gains applied to more advanced (M_2) mathematics. Consequently, schools may not see benefits unless they assess more advanced reasoning. If a school's only goal is to increase scores on the basics, the SimCalc intervention may not be appropriate. Second, the results were obtained in Texas, a state with a long record of a stable standards-based educational system and an ability to implement a train-the-trainer model across regions. Results may vary in states with different contexts. Third, although we view replacement units as a good strategy to fit within school constraints, the tested replacement units occupied only a modest amount of instructional time. We do not yet know the consequences of more extended uses of such units and do not necessarily recommend using software every day; software use may be most useful when targeted specifically at the conceptually advanced aspects of mathematics learning. Fourth, our samples lacked any majority African American school. Fifth, we worked with volunteer teachers and do not know how well nonvolunteer teachers would fare. Sixth, we required schools to have access to a classroom set of computers, but not all schools have suitable computer facilities. Seventh, we tested an intervention that incorporated only one kind of software and not others. Other software and hardware technologies emphasize dynamic representations, including graphing calculators, dynamic geometry software (e.g., The Geometer's Sketchpad, Cabri Géomètre), and dynamic statistics packages (e.g., TinkerPlots, Fathom). But there are also many technologies for mathematics learning that are not included in this family. We do not know

whether these results will generalize within or beyond the category of representational tools or dynamic mathematics tools.

Refinements to SimCalc Program Theory

The slogan of the SimCalc program is "democratizing access to the mathematics of change and variation." Given the robustness findings, it is fair to say that the materials provide students in a wide variety of settings with access to more advanced mathematics while providing ample opportunity for them to make progress on the basics for which schools are most accountable. The intervention might have greater impact with more attention on the interaction between teacher-reported achievement level and student learning gains within classrooms. In both the seventh- and eighth-grade experiments, teacher-reported achievement expectations correlated with student gains. In interviews after implementation of the intervention, we noted that many teachers reported a belief that these materials are more appropriate for their high-achieving students. To the contrary, our findings suggest that the materials are better than the existing materials for students in all teacher-reported achievement categories. It could be that with further professional development, teachers could learn to more effectively use these materials with students they believe are low or medium achievers. In case studies conducted within the context of our experiments, we are examining this possibility.

We discussed three implementation variables that suggest how the intervention worked. An important student resource dimension of the intervention is use of computer software. Although the data reported here are only a proxy for actual student use of SimCalc MathWorlds software, we did see that students in the intervention went to the computer lab more frequently and that more days in the computer lab led to greater student learning gains. We are further

pursuing understanding the role of the technology through case studies, which are based on video recordings of classrooms, and analyses of the teacher interviews conducted after implementation. So far, these data are consonant with the program theory, which is that the SimCalc MathWorlds software provides a representational infrastructure that better supports student learning (Empson, 2009; Dunn, 2009). We hasten to add that this study should not be interpreted as a technology/no-technology comparison. SimCalc program theory emphasizes the integration of professional development, curriculum, and representational technology, and our experiments were designed to test the integration and not to isolate the effect of one component.

An important teacher-resource dimension was teachers' self-reports of their topical focus for each day's class. We found that teachers in the two groups reported emphasizing different topics. In both seventh and eighth grade, teachers in both groups emphasized basic skills and the content covered by the Texas state test to approximately the same degree. In addition, in both seventh and eighth grade, SimCalc teachers placed greater emphasis on advanced mathematics. Thus, SimCalc teachers reported expanding the range of mathematics they covered to include more advanced concepts; they did not report neglecting more basic topics in favor of more advanced ones. In the Seventh-Grade Experiment Year 1, in a previous paper we also examined teacher self-reports of their daily teaching emphasis in terms of cognitive demand, ranging from a focus on facts and routine procedures (low demand) to a focus on conceptual understanding and nonroutine problem solving (high demand). We found that teachers in the SimCalc condition reported emphasizing high-demand tasks more frequently and that the more teachers reported this emphasis, the more their students learned (Roschelle, Shechtman et al., 2008). This is particularly important in light of the finding that teachers in both groups spent about the same number of days on the target unit, because it suggests an *intensification* of mathematics learning

within the same number of days. This, too, is consonant with SimCalc program theory.

While we are further investigating teacher-student interactions in case studies using video, we have little reason to suspect that the SimCalc intervention operates by changing how teachers interact with students (Empson, 2009; Dunn, 2009). Within SimCalc classrooms, one dissertation researcher using videotapes from the Seventh-Grade Experiment Year 1 has found a correlation between teacher behavior and student learning. Teachers who were more responsive to student ideas and presented students with more challenging mathematical tasks were likely to have higher learning gains (Pierson, 2008). Our interpretation of these data is that most teachers' existing pedagogies are good enough to permit a successful SimCalc implementation in the first year, but high-quality teacher-student interactions can augment the impact.

Some pedagogies might lead to greater learning with the SimCalc materials. Indeed, the program developers have strong beliefs that highly interactive pedagogies are more beneficial to students than teacher-centered pedagogies. Yet one set of case studies (Empson, 2009) found that different configurations of SimCalc learning resources can be successful; it is not empirically clear that there is one best pedagogy for teaching with these materials. In addition, more professional development is likely to be necessary to sustain and expand implementations across many years. The program developers believe that the software and curriculum materials can support strong mathematical argumentation in the classroom; teachers may need long-term professional development to improve their support for mathematical argumentation practices.

The project team is working to mine additional findings from the large dataset. Our future work includes case studies in selected classrooms, using MLM to model the impact of background and implementation variables, analysis of interview data from teachers, analysis of the mathematics at the level of particular items on the test, and many other more detailed

inquiries. We are also examining sustainability, although this particular set of experiments was optimized to examine other research questions, not to examine or increase the odds of long-term adoption in Texas schools. We are particularly interested in understanding how variation at the school and teacher levels affects student learning but observe that robustness runs counter to finding strong moderating variables. Future studies in settings with other populations (e.g., more African American students, more urban students) and longitudinal studies would complement our research to date.

Conclusions

We designed a series of randomized controlled experiments to evaluate an integration of teacher professional development, paper curriculum, and representational software. The materials were developed via design research methods and previously had been evaluated only in small numbers of classrooms. We addressed robustness not only by including many teachers with different backgrounds, attitudes and levels of mathematical knowledge in our study, but also by testing the intervention in a variety of settings. Further, we based our work on existing scaling mechanisms by which innovations are spread from a central point to many regions, schools, teachers, and students within Texas. In addition, we gave teachers a modest amount of support (i.e., about 3–5 days of professional development), well within the range of what many school districts could afford. Our eighth-grade experiment used a train-the-trainers model of professional development delivery, another realistic element at scale.

To our knowledge, this is the first series of randomized controlled experiments to examine the effects of *representational* technologies in improving student learning using multilevel modelling. Other research has involved different kinds of technology or research

methodologies with less power to support causal claims. In particular, the prominent Dynarski et al. (2007) study was widely interpreted to suggest that technology has no effect on mathematics learning. Similarly, the National Mathematics Advisory Panel (2008) found little scientific evidence for the effectiveness of technology in mathematics learning. In this context, the replication of our findings across three studies in both seventh and eighth grade contributes to the literature by providing evidence from three experiments that a different approach to technology can produce robust effects. The SimCalc approach is distinguished from the interventions considered in the Dynarski et al. (2007) studies in that *representational* technology was the focus and the SimCalc program emphasized tight integration of curriculum, technology and teacher professional development. The kind of technology and the level of curricular integration may matter a great deal in the effects of incorporating technology in mathematics education.

We are well aware that this result was obtained for a curriculum unit of limited duration; a logical next step would be to expand to cover more of seventh and eighth grade mathematics. The SimCalc MathWorlds software is one instance of software that takes a *representational* approach. Dynamic Geometry computer software (e.g., The Geometer's Sketchpad, Cabri Géomètre) is also well established and has been deeply theorized (Lehrer & Chazen, 1998; Laborde, 2000) but has not been subjected to experimental trials across settings. Dynamic representational software for statistics is also available (Konold, 2002). We observe that the combination of dynamic algebra, geometry and statistics software could cover the central topics in middle school mathematics, for example as described in the NCTM Focal Points (National Council of Teachers of Mathematics, 2007). Thus a logical next step would be to expand the approach of integrating software, curriculum and teacher professional development to cover the key ideas in algebra, geometry, and statistics in all of seventh and eighth grade.

It is perhaps particularly interesting that this approach enabled students to *both* learn the basics as required by federal and state mandates *and* learn more advanced mathematics on the pathway to Algebra, an important policy goal. If we had only measured the basic skills required in Texas, we would have obtained a null result. Technology may be particularly valuable in mathematics education when educators seek to go beyond the basics. Educators who wish to go beyond the basics may be able to use representational technology to intensify instruction and thus cover both the basics and more advanced skills and concepts.

In terms of broader recommendations to the field, we see this work as suggesting that less emphasis should be placed on the value of technology alone and more on interventions that deeply integrate professional development, curriculum materials, and software in a unified curricular activity system. We select the word "activity" with care based on our observation that all elements of the SimCalc intervention align around enacting particular activities in the classroom (in contrast to a focus on lessons, assessments, or projects). Through our research we observed the complexity and variability in implementing these activities in classrooms. More research is needed to understand the design features of curricular activities that allow for adaptation to different student populations and teaching styles without undermining effectiveness.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5).

Anderson, J. R., Corbett, A. T., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, (4), 167–207.

Becker, H. J., & Anderson, R. E. (1998). Teaching, learning, and computing: 1998. Teacher's Survey: Combined Versions, 1–4. University of California-Irvine.

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

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 61–100). Washington, DC: American Educational Research Association.

Brown, A. L. (1992). Design experiments. Theoretical and methodological challenges in evaluating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*(2), 141–178.

Campbell, D. T., & Kenney, D. A. (1999). *A primer on regression artifacts*. New York: Guilford.

Campbell, D. T., Shadish, W. R., & Cook, T. D. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin Co.

Cobern, C. E. (2002). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3–12.

Cohen, D. K., & Hill, H. (2001). Learning policy: When state education reform works. New Haven, CT: Yale University Press.

Cohen, D. K., Raudenbush, S., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 1–24.

Corbett, A. T., Koedinger, K. R., & Hadley, W. H. (2001). Cognitive tutors: From the research classroom to all classrooms. In P. S. Goodman (Ed.), *Technology enhanced learning*: *Opportunities for change* (pp. 235–263). Mahwah, NJ: Lawrence Erlbaum Associates.

Denton, C., Vaughn, S., & Fletcher, J. (2003). Bringing research-based practice in reading intervention to scale. *Learning Disabilities Research and Practice*, *18*(3), 201–211.

Design Based Research Collaborative. (2002). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, *32*(1), 5–8.

Dewey, J. (1938). *Experience and education*. New York: Macmillan Company.

Doerr, H. M., & Zangor, R. (2000). Creating meaning for and with the graphing calculator. *Educational Studies in Mathematics*, *41*(2), 143-163.

Dunn, M. B. (2009). *Investigating variation in teaching with technology-rich interventions: What matters in training and teaching at scale?* Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ.

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: National Center for Educational Evaluation.

Ellington, A. J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education*, *34*(5), 433–463.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1-25.

Empson, S. B., Greenstein, S., & Maldonado, L. (2009). Scaling Up Innovative Mathematics in the Middle Grades: Case Studies of "Good Enough" Enactments. Manuscript submitted for publication.

Fauvel, J., & Maanen, J. A. (Eds.). (2000). *History in mathematics education: The ICMI study*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Fletcher, J., Foorman, B., Denton, C., & Vaughn, S. (2006). Scaling research on beginning reading: Consensus and conflict. In M. A. Constas & R. J. Sternberg (Eds.), *Translating theory and research into educational practice: Developments in content domains, large-scale reform, and intellectual capacity* (pp. 53–75). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gutierrez, R. G., Carter, S., & Drukker, D. M. (2001). On boundary-value likelihood-ratio tests [Electronic Version]. *Stata Technical Bulletin, STB-60,* 16-18. Retrieved October 9, 2009, from http://www.stata.com/products/stb/journals/stb60.pdf

Harel, G., & Confrey, J. (1994). *The development of multiplicative reasoning in the learning of mathematics*. Albany, NY: State University of New York Press.

Heid, M. K., & Blume, G. W. (2008). Algebra and function development. In M. K. Heid & G. W. Blume (Eds.), *Research on technology and the teaching and learning of mathematics: Research syntheses* (Vol. 1, pp. 55–108). Charlotte, NC: Information Age.

Hiebert, J., & Behr, M. (1988). *Number concepts and operations in the middle grades*. Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jowett, B. (1875). *The dialogues of Plato* (2nd ed. Vol. 1). Oxford, UK: Claredon Press.

Kaput, J. (1992). Technology and mathematics education. In D. Grouws (Ed.), *A handbook of research on mathematics teaching and learning* (pp. 515–556). New York: Macmillan.

Kaput, J. (1994). Democratizing access to calculus: New routes using old roots. In A. Schoenfeld (Ed.), *Mathematical thinking and problem solving* (pp. 77–155). Hillsdale, NJ: Erlbaum.

Kaput, J. (1997). Rethinking calculus: Learning and thinking. *The American Mathematical Monthly*, *104*(8), 731–737.

Kaput, J., & Hegedus, S. (2002). *Exploiting classroom connectivity by aggregating student constructions to create new learning opportunities*. 26th Conference of the International Group for the Psychology of Mathematics Education, Norwich, UK.

Kaput, J., Hegedus, S., & Lesh, R. (2007). Technology becoming infrastructural in mathematics education. In R. Lesh, E. Hamilton, & J. Kaput (Eds.), *Foundations for the future in mathematics education* (pp. 173–192). Mahwah, NJ: Lawrence Erlbaum Associates.

Kaput, J., & Roschelle, J. (1998). The mathematics of change and variation from a millennial perspective: New content, new context. In C. Hoyles, C. Morgan, & G. Woodhouse (Eds.), *Rethinking the mathematics curriculum*. London: Falmer Press.

Konold, C. (2002). Teaching concepts rather than conventions. *New England Journal of Mathematics, 34(2),* 69-81.

Laborde, C. (2000). Dynamic geometry environments as a source of rich learning contexts for the complex activity of proving. *Educational Studies in Mathematics*, *44*(1), 151–161.

Lehrer, R., & Chazen, D. (Eds.). (1998). *Designing learning environments for developing understanding of geometry and space*. Mahwah, NJ: Lawrence Erlbaum Associates.

Leinhardt, G., Zaslavsky, O., & Stein, M. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60,* 1–64.

Lesh, R., Cramer, K., Doerr, H. M., Post, T., & Zawojewski, J. (2003). Model development sequences. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: A models & modelling perspective on mathematics problem solving, learning and teaching* (pp. 35–58). Mahwah, NJ: Lawrence Erlbaum Associates.

Lobato, J., Clarke, D., & Ellis, A. B. (2005). Initiating and eliciting in teaching: A reformulation of telling. *Journal for Research in Mathematics Education, 36*(2), 101–13636.

Maris, E. (1998). Covariance adjustment versus gain scores - revisited. *Psychological Methods, 3*(3), 309-327.

Marzano, R. J. (1998). *A theory-based meta-analysis of research on instruction*. Aurora, CO: Mid-continent Research for Education and Learning.

Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design. CRESST Technical Paper Series*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Nathan, M. J., & Koellner, K. (2007). A framework for understanding and cultivating the transition from arithmetic to algebraic reasoning. *Mathematical Thinking and Learning, 9*(3), 179–192.

National Center for Education Statistics. (2006). *The nation's report card: Mathematics 2005* (No. NCES 2001–571). Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (2007). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: National Council of Teachers of Mathematics.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: U. S. Department of Education.

Papert, S. (1980). *Mindstorms: Computers, children, and powerful ideas*. New York: Basic Books.

Pierson, J. (2008). *The relationship between patterns of classroom discourse and mathematics learning*. Unpublished doctoral dissertation, University of Texas at Austin.

Post, T. R., Cramer, K. A., Behr, M., Lesh, R., & Harel, G. (1993). *Curriculum implications of research on the learning, teaching, and assessing of rational number concepts. Rational numbers: An integration of research* (pp. 327–362). Hillsdale, NJ: Erlbaum.

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

Roschelle, J. (2003). Unlocking the learning value of wireless mobile devices. *Journal of Computer Assisted Learning, 19*(3), 260–272.

Roschelle, J., Kaput, J., Stroup, W., & Kahn, T. (1998). *Scaleable integration of educational software: Exploring the promise of component architectures*. Retrieved January 8, 2003, 2003, from http://www-jime.open.ac.uk/98/6/

Roschelle, J., Pea, R., Hoadley, C., Gordin, D., & Means, B. (2000). Changing how and what children learn in school with computer-based technologies. *The Future of Children, 10*(2), 76–101.

Roschelle, J., Shechtman, N., Hegedus, S., Pierson, J., McLeese, M., & Tatar, D. (2008). *Cognitive complexity in mathematics teaching and learning: Emerging findings in a large-scale experiment*. Paper presented at the International Conference of the Learning Sciences, Utrecht, Netherlands.

Roschelle, J., Tatar, D., & Kaput, J. (2008). Getting to scale with innovations that deeply restructure how students come to know mathematics. In A. E. Kelly, R. Lesh & J. Y. Baek (Eds.), *Handbook of design research methods in education* (pp. 369–395). New York: Routledge.

Roschelle, J., Tatar, D., Shechtman N., & Knudsen, J. (2008). The role of scaling up research in designing for and evaluating robustness. *Educational Studies in Mathematics*, *68*, 149-170.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. C., Wiley, D. E., Cogan, L. S., et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21.

Tai, R. H., Qi Liu, C., Maltese, A. V., & Fan, X. (2006). Career choice enhanced: Planning early for careers in science. *Science, 312,* 1143–1144).

Tatar, D., Roschelle, J., Knudsen, J., Shechtman, N., Kaput, J., & Hopkins, B. (2008). Scaling up technology-based innovative mathematics. *Journal of the Learning Sciences, 17*(2), 248–286.

Tatar, D., & Stroter, A. (2009). *Recruitment strategies, outcomes, and implications for a randomized controlled experiment with teachers* (SimCalc Technical Report 3). Menlo Park CA: Center for Technical and Learning, SRI International.

Varelas, M., & Becker, J. (1997). Children's developing understanding of place value: Semiotic aspects. *Cognition and Instruction, 15*(2), 265–286.

Vergnaud, G. (1988). Multiplicative structures. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 141–161). Reston, VA: National Council of Teachers of Mathematics.

Wenglinksy, H. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics*. Princeton, NJ: Educational Testing Service.

APPENDIX

SAMPLE CHARACTERISTICS

Sample Sizes by Study and Group

Seventh-Grade Year 1 Experiment			Seventh-Grade Quasi-Experiment			Eight-Grade Experiment		
Group	$N_{Teachers}$	$N_{Students}$*	Group	$N_{Teachers}$	$N_{Students}$*	Group	$N_{Teachers}$	$N_{Students}$*
Delayed	47	825	Year 1	30	510	Control	23	303
Immediate	48	796	Year 2		538	Treatment	33	522
Total	95	1,621	Total	30	1,048	Total	56	825

*Only students for whom we have complete data (both pretest and posttest) are included here.

Teacher Characteristics

	Seventh-Grade Year 1 Experiment		Seventh-Grade Quasi-Experiment	Eighth-Grade Experiment	
Variable	**Delayed**	**Immediate**	**Delayed-treatment teachers who completed Years 1 and 2**	**Control**	**Treatment**
Total count	47	48	30	23	33
Teachers by region					
Region 1 (Edinburg)	11	8	6	--	--
Region 9 (Witchita Falls)	--	--	--	4	3
Region 10 (Dallas)	--	--	--	4	8
Region 11 (Fort Worth)	13	14	8	--	--
Region 13 (Austin)	13	11	8	10	13
Region 17 (Lubbock)	--	--	--	5	6
Region 18 (Midland)	10	15	8	0	3
Female (%)	81	77	80	82.6	84.8
Years teaching total					
Mean	10.5	12.4	10.3	9.6	7.9
Range	1–29	1–40	1–27 (+1 in year 2)	0–27	0–31
Years teaching mathematics					
Mean	9.5	11.0	9.0	9.9	8.2
Range	1–29	1–40	1–27 (+1 in year 2)	0–27	1–32
Teacher ethnicity (%)					
White	70.2	77.1	70.0	87.0	78.8
Hispanic	25.5	20.8	23.3	8.7	15.1
Asian	4.3	0	6.7	0	0
African American	0	2.1	0	4.3	6.0
Master's degree (%)	17.0	18.8	16.7	26.1 *	6.0

* $p = .06$

Note: Significance tests compared groups within study using a two-level MLM model. Within each study, no significant differences existed between groups on any of these variables.

School Characteristics

Variable	Seventh-Grade Year 1 Experiment		Seventh-Grade Quasi-Experiment	Eighth-Grade Experiment	
	Delayed	Immediate	Delayed-treatment teachers who completed Years 1 and 2	Control	Treatment
Total count of schools	37	36	25	19	23
Total campus enrollment					
Mean	612	557	584	569	550
Range	71–1490	102–1119	71–1490	104– 2245	121–1375
Free/reduced-price lunch (%)					
Mean	49	54	53	47	43
Range	4–99	1–94	11–99	12–89	0–92
Campus ethnicity (mean %)					
White	48	49	43	59	57
Hispanic	44	45	48	30	35
Asian	2	2	2	1	1
African American	6	4	6	9	6

Note: Within each study, no significant differences existed between groups on campus enrollment, free/reduced-price lunch, or campus ethnicity.

Student Characteristics

Variable	Seventh-Grade Year 1 Experiment		Seventh-Grade Quasi-Experiment		Eighth-Grade Experiment	
	Delayed	Immediate	Year 1	Year 2	Control	Treatment
Total count of students	825	796	510	538	303	522
Female (%)	50.6	48.9	52.4	41.9**	45.1	47.9
Individual ethnicity (%)						
White	38.7	48.5	39.6	35.4	65.6	50.0
Hispanic	54.1	44.3	51.7	55.8	22.7	40.7*
Asian	2.0	1.5	2.8	2.6	1.1	1.3
African American	4.7	4.2	5.5	5.3	9.5	6.9
Achievement level (%)						
Low	26.2	22.5	26.9	27.5	23.4	25.1
Medium	42.9	35.9	45.3	41.1	42.4	37.7
High	24.2	28.6	25.9	25.5	25.4	26.6

** $p < .01$; * $p < .05$

Note: Significance tests compared groups within study using a two-level MLM model.

Tables

Table 1

Mathematical Conceptual Frameworks for the Seventh-Grade and Eighth-Grade Curricula and Assessments. M_1 and M_2 refer to the two major dimensions of each framework.

Framework	**M_1 Component** *Foundational concepts typically covered in the grade-level standards, curricula, and assessments*	**M_2 Component** *Building on the foundations of M_1, essentials of concepts of mathematics of change and variation found in algebra, calculus, and the sciences*
Rate and Proportionality for the Seventh-Grade Studies	• Simple $a/b = c/d$ or $y = kx$ problems in which all but one of the values are provided and the last must be calculated • Basic graph and table reading without interpretation (e.g., given a particular value, finding the corresponding value in a graph or table of a relationship)	• Reasoning about a representation (e.g., graph, table, or $y = kx$ formula) in which a multiplicative constant k represents a constant rate, slope, speed, or scaling factor across three or more pairs of values that are given or implied • Reasoning across two or more representations
Linear Function for the Eighth-Grade Study	• Categorizing functions as linear/nonlinear and proportional/nonproportional • Within one representation of one linear function (formula, table, graph, narrative), finding an input or output value • Translating one linear function from one representation to another	• Interpreting two or more functions that represent change over time, including linear functions or segments of piecewise linear functions • Finding the average rate over a single multirate piecewise linear function

Table 2

Summary of Key Contrasts Between the Seventh-Grade and Eighth-Grade Studies

	Seventh-Grade Study	Eighth-Grade Experiment
Start of the experiment	Summer 2005	Summer 2006
Duration	2 years	1 year
Embedded study	Primary experiment and Quasi-experimental study	Primary experiment only
Mathematical content focus	Rate and proportionality	Linear function
Teacher professional development (TPD)		
Delivery model	Researcher led	Train-the-trainers model
Mathematics knowledge for teaching	2-day TEXTEAMS workshop	Not emphasized
Total length of treatment TPD	5 days (2 days TEXTEAMS and 3 days SimCalc)	3 days SimCalc only
Counterfactual		
Classroom implementation	Business as usual	Business as usual
TPD	TEXTEAMS workshop covered content knowledge pertinent to the unit	TMT3 workshop covered orthogonal but important content knowledge

Table 3

Student Test Scores at the Student Level. The seventh-grade assessment had 30 items and the eighth-grade assessment had 36 items

		Total Score						Effect Size of Gain Score Difference
		Pretest		Posttest		Gain		
	N	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	
Total score								
Seventh-Grade Year 1 Experiment								
Control	825	12.7	5.7	15.0	5.7	2.2	3.8	0.63***
Treatment	796	13.2	5.7	19.0	6.0	5.8	4.0	
Seventh-Grade Quasi-Experiment								
Delayed Year 1	510	12.8	5.2	15.2	5.5	2.4	3.9	0.50***
Delayed Year 2	538	12.6	5.4	17.7	6.2	5.1	3.9	
Eighth-Grade Experiment								
Control	303	12.5	7.6	15.4	8.4	2.9	5.2	0.56***
Treatment	522	11.9	7.3	18.9	8.7	7.0	5.0	
M_1 subscale								
Seventh-Grade Year 1 Experiment								
Control	825	7.2	2.7	8.0	2.5	0.8	2.2	0.10
Treatment	796	7.5	2.6	8.6	2.0	1.1	2.1	
Seventh-Grade Quasi-Experiment								
Delayed Year 1	510	7.3	2.5	8.2	2.4	0.8	2.3	0.13*
Delayed Year 2	538	7.3	2.6	8.5	2.2	1.2	2.1	
Eighth-Grade Experiment								
Control	303	7.2	3.8	8.7	4.0	1.5	2.9	0.19
Treatment	522	7.2	3.6	9.4	4.2	2.2	2.7	
M_2 subscale								
Seventh-Grade Year 1 Experiment								
Control	825	5.5	3.6	7.0	4.0	1.4	2.7	.89***
Treatment	796	5.7	3.8	10.5	4.5	4.7	3.3	
Seventh-Grade Quasi-Experiment								
Delayed Year 1	510	5.4	3.4	7.0	3.8	1.6	2.8	.69***
Delayed Year 2	538	5.3	3.5	9.2	4.5	3.9	3.2	
Eighth-Grade Experiment								
Control	303	5.3	4.4	6.6	4.9	1.4	3.5	.81***
Treatment	522	4.7	4.2	9.5	4.9	4.8	3.3	

*** $p < .0001$; * $p < .05$

Table 4

Two-level MLM Models Run in Each Study for Each Factor Predicting M_2 Gains

	Seventh-Grade Year 1 Experiment (Main effect is experimental condition) N = 1,444		**Seventh-Grade Quasi-Experiment** (Main effect is year 1 vs. year 2) N = 997		**Eighth-Grade Experiment** (Main effect is experimental condition) N = 657	
Model	Value	SE	Value	Value	SE	Value
Unconditional						
Intercept	3.03***	0.275	2.82***	0.252	3.26***	0.377
Level 2 Variance	4.72		1.31		4.88	
Residual Variance	8.07		9.16		8.82	
$\bar{\chi}^{2\ \dagger}_{01}$	463.03***		76.85***		174.64***	
Main Effect Only						
Main Effect	3.55***	0.343	2.46***	0.179	3.26***	0.544
Intercept	1.34***	0.236	1.63***	0.280	1.44***	0.416
Level 2 Variance	1.55		1.53		2.11	
Residual Variance	8.07		7.65		8.84	
$\bar{\chi}^{2}_{01}$	134.21***		109.66***		51.55***	
School is in Region 1						
Main Effect	3.74***	0.377	2.83***	0.204	3.26***	0.544
Region 1	0.35	0.611	0.62	0.760	0.00	
Region 1 Interaction	-1.10	0.909	-1.48***	0.415	0.00	
Intercept	1.28***	0.262	1.47***	0.314	1.44***	0.416
Level 2 Variance	1.55		1.62		2.11	
Residual Variance	8.07		7.56		8.84	
$\bar{\chi}^{2}_{01}$	120.24***		110.34***		51.55***	
Free/reduced-price lunch (%)						
Main Effect	3.57***	0.344	2.55***	0.177	3.27***	0.553
SES	0.53	0.877	0.52	0.975	1.03	1.847
SES Interaction	-1.25	1.247	-3.53***	0.636	-2.35	2.500
Intercept	1.35***	0.236	1.59***	0.263	1.41***	0.423
Level 2 Variance	1.54		1.31		2.17	
Residual Variance	8.07		7.45		8.84	
$\bar{\chi}^{2}_{01}$	120.34***		81.42***		50.09***	
Student is Hispanic						
Main Effect	3.73***	0.383	3.49***	0.259	3.63***	0.570
Hispanic	-0.25	0.282	0.10	0.292	0.40	0.506
Hispanic Interaction	-0.41	0.391	-1.84***	0.351	-1.04	0.603
Intercept	1.47***	0.269	1.56***	0.293	1.33**	0.433
Level 2 Variance	1.49		1.22		2.05	
Residual Variance	8.05		7.41		8.82	
$\bar{\chi}^{2}_{01}$	123.96***		75.04***		49.43***	

	Seventh-Grade Year 1 Experiment (Main effect is experimental condition) N = 1,444		**Seventh-Grade Quasi-Experiment** (Main effect is year 1 vs. year 2) N = 997		**Eighth-Grade Experiment** (Main effect is experimental condition) N = 657	
Model	Value	SE	Value	Value	SE	Value
Student is Female						
Main Effect	3.49***	0.374	2.60***	0.246	3.35***	0.594
Female	-0.01	0.211	0.20	0.253	-0.51	0.411
Female Interaction	0.14	0.307	-0.27	0.359	-0.17	0.505
Intercept	1.35***	0.259	1.53***	0.308	1.69***	0.459
Level 2 Variance	1.54		1.52		2.06	
Residual Variance	8.08		7.66		8.78	
$\bar{\chi}^2_{01}$	133.05***		107.46***		49.89***	
Teacher-rated prior achievement						
Main Effect	3.45***	0.387	2.56***	0.261	3.02***	0.609
High Group	0.47	0.273	0.36	0.315	0.48	0.490
Low Group	-0.35	0.256	-0.30	0.303	-1.11*	0.505
High Group Interaction	0.77	0.399	0.69	0.435	0.11	0.614
Low Group Interaction	-0.81*	0.379	-0.91*	0.424	0.41	0.626
Intercept	1.33***	0.264	1.61***	0.303	1.70***	0.470
Level 2 Variance	1.55		1.41		2.05	
Residual Variance	7.76		7.36		8.67	
$\bar{\chi}^2_{01}$	141.98***		97.42***		45.97***	

† $\bar{\chi}^2_{01}$ statistic is an adjusted chi-square statistic from a likelihood ratio test of the given model against a model without random intercepts. See Gutierrez, Carter, and Drukker (2001) for details.

Note: Full model is $Y_{ij} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}X_{ij} + \gamma_{03}T_j * X_{ij} + r_{ij} + u_j$, where X_{ij} may be a level 1 or level 2 covariate. All models within an experiment fit are estimated on identical sets of cases.

Figure Captions

Figure 1. MathWorlds Software Animation of Motion

Figure 2. Student Mean Difference Scores (± SE of total using MLM) at the Student Level

Figure 3. Mean Student Learning Gains on M_2 by Subpopulation Group. Data are presented at the student level.

Figure 4. In the Seventh Grade Experiment, Days Spent Teaching Various Basic Operations, M_1, and M_2 Topics

Figure 5. In the Eighth-Grade Experiment, Days Spent Teaching Various M_1 and M_2 Topics

Figure 1. MathWorlds software animation of motion.

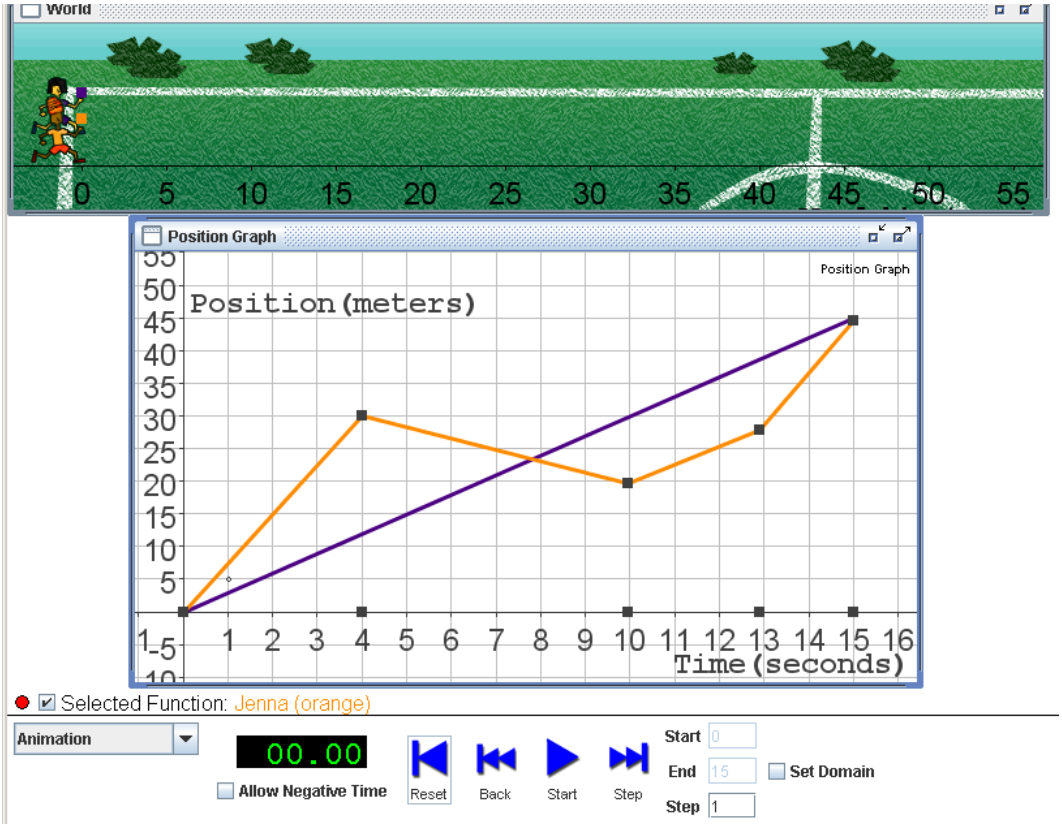


Figure 2. Student mean difference scores (± SE of total using MLM) at the student level.

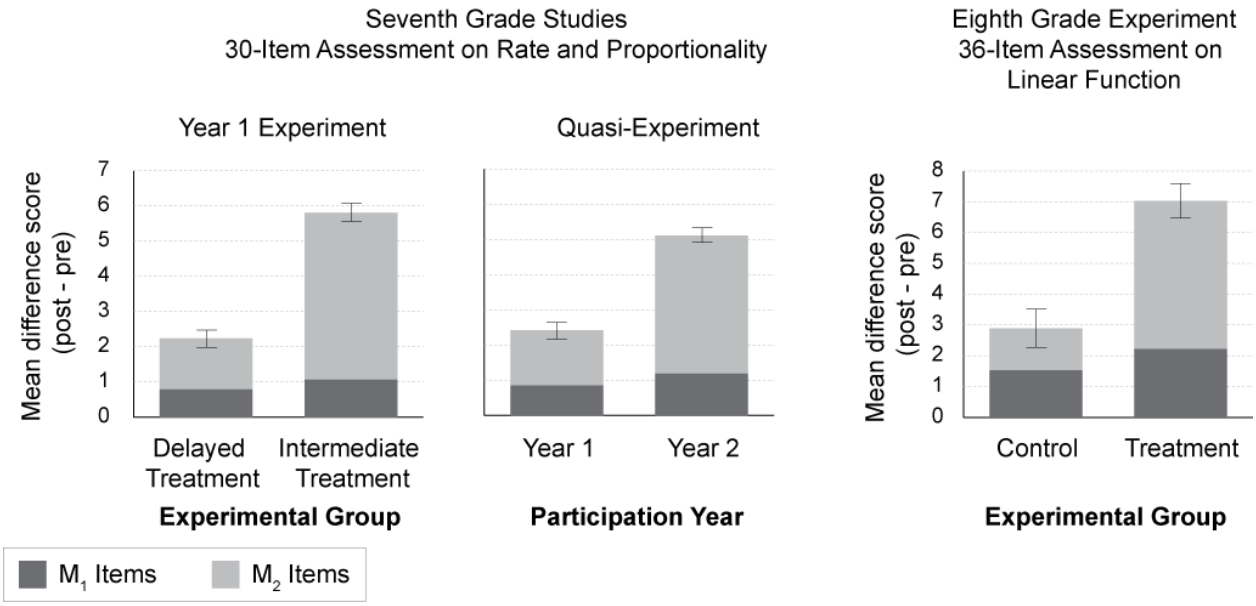
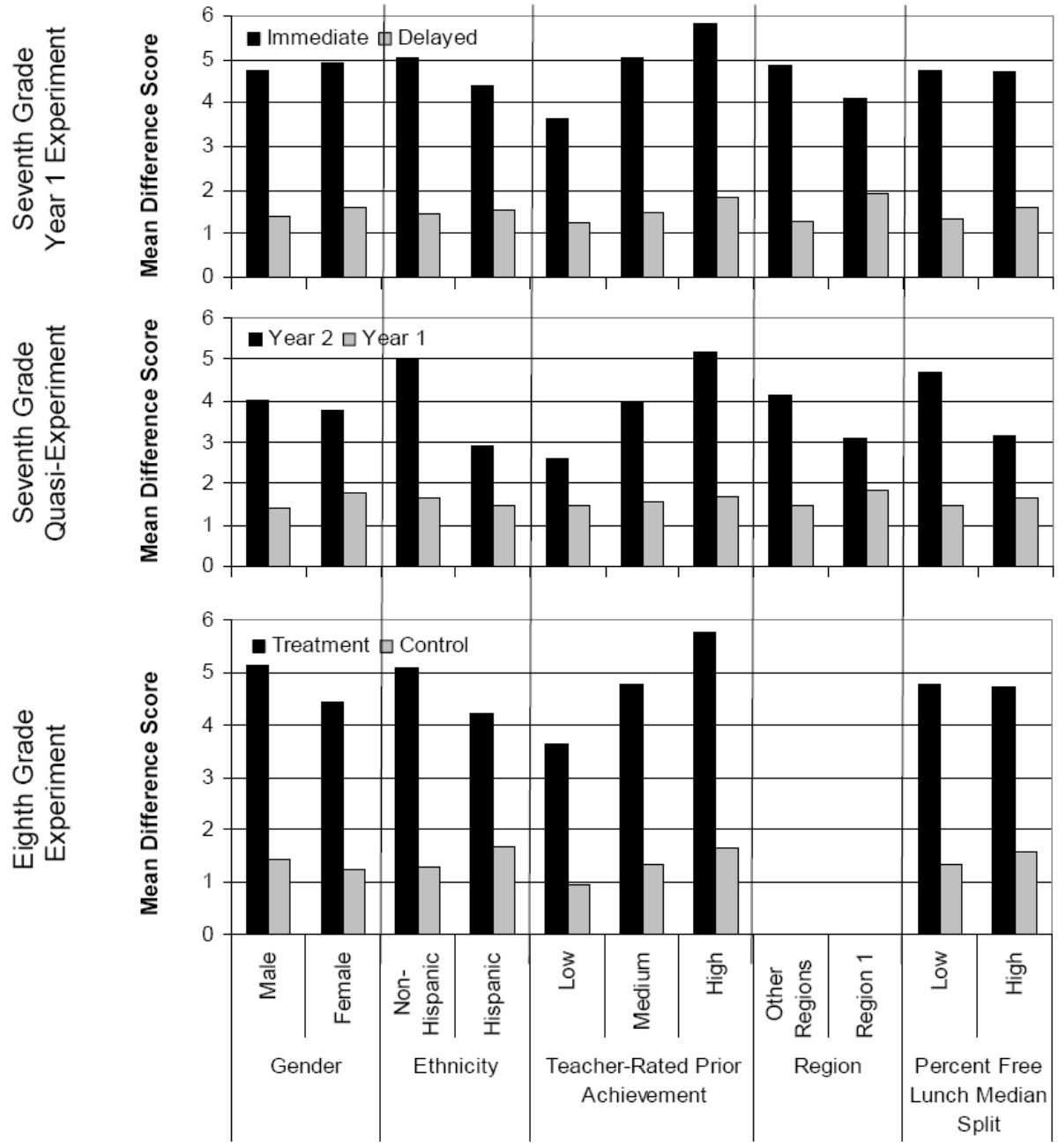
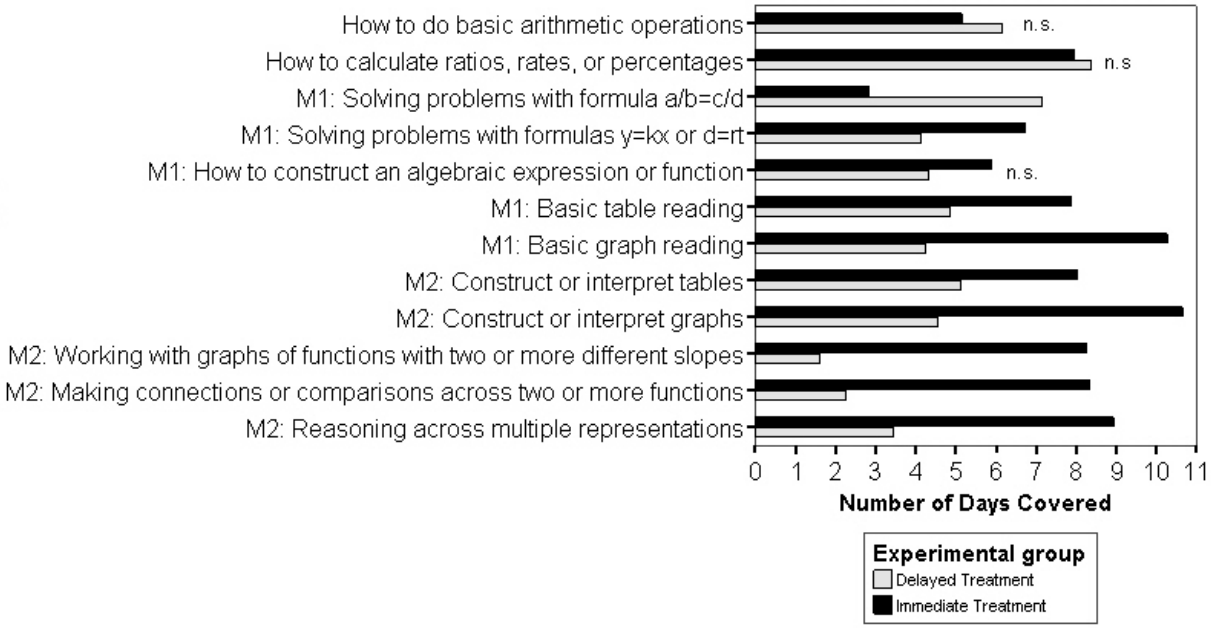


Figure 3. Raw mean student learning gains on M_2 by subpopulation group. Data are presented at the student level.



Note: Region 1 only participated in the Seventh Grade Studies, not the Eighth Grade Experiment, due to a shift in local circumstances in the year between recruitment for the two studies.

Figure 4. In the Seventh-Grade Studies, days spent teaching various basic operations, M_1, and M_2 topics.



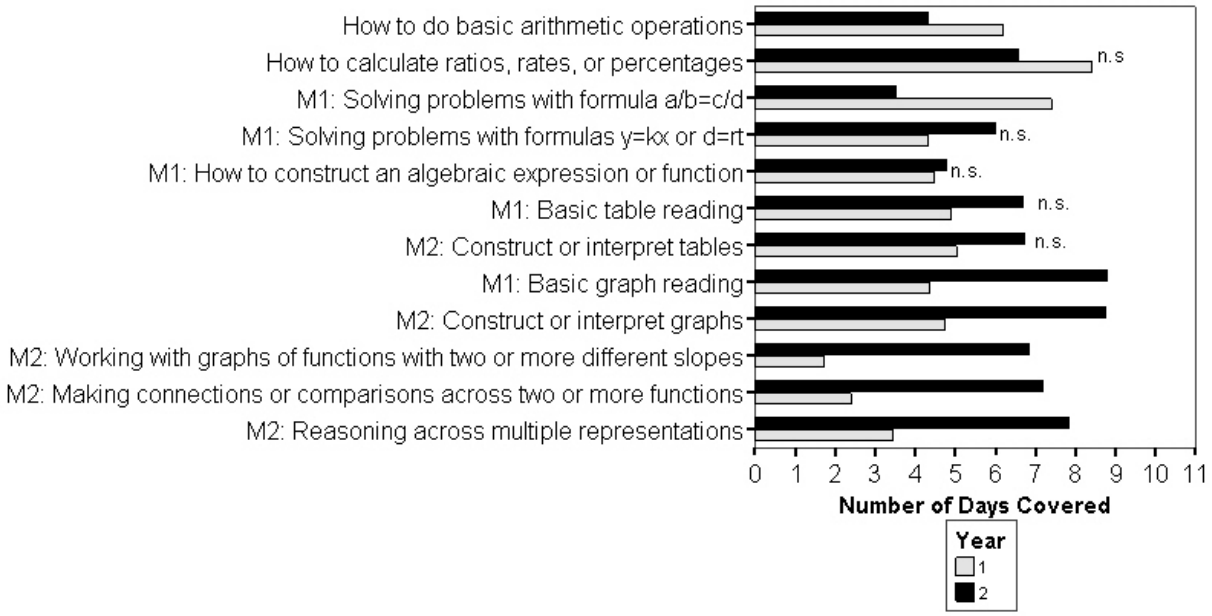


Figure 5. In the Eighth-Grade Experiment, days spent teaching various M_1 and M_2 topics.