GETTING TO SCALE WITH INNOVATIONS THAT RESTRUCTURE DEEPLY HOW STUDENTS LEARN MATHEMATICS

Jeremy Roschelle, Deborah Tatar, and Jim Kaput

Most people enter the field of educational research with good intentions of improving education and the lives of children. However, countless good ideas remain in university halls. Curricular and pedagogical innovations rise and fall in schools because there is insufficient understanding of how the innovation can and will be used. At the same time, technological innovations in schools are becoming more and more politically contentious. People, reasonably, want to know that we are giving their children “proven” curricular materials. To Congress, this means that evaluators should engage in “scientifically based research” ("No child left behind act", 2001). Not willing to cede the definition of scientific methodology to lawmakers, educational researchers have begun their own vigorous debate. Fundamentally, they ask (and this book asks): “What should count as a scientific warrant that evidence supports a claim?”

Partisans of two important educational research perspectives have made strong progress toward defining and defending their answers. Taking the perspective of program evaluation and educational psychology, some researchers ask: “What works?” (Shavelson & Towne, 2002). The objects of their inquiry are selected from available materials or programs and the measures are related closely to today’s critical tests. Although researchers in this group acknowledge the utility of multiple research methods, they make their strongest case for the virtues of experimentation, preferably with randomized controlled designs (Cook, 2000; Torgerson, 2001).

Taking the perspective that builds on cognitive science insights and aims for the design of new materials, technologies, or practices, another group (Design Based Research Collaborative, 2002) asks: "What *could* work?" Their object of inquiry is the design of materials for students to learn important yet vexingly difficult mathematics and science concepts. Proposals often argue that existing materials do not meet this challenge, and, therefore, research must be coupled tightly with design. Although researchers in this group acknowledge the utility of multiple methods too, they make the strongest case for design research methods. Research methods that reveal students' deep progress on difficult concepts and that yield theoretical insight into the "active ingredients" of innovations are valued especially (Cobb *et al.*, 2002; Edelson, 2002).

In this chapter, we highlight a third perspective. Like the design researchers, the third perspective attends to *innovation*. And like the program evaluators, this perspective also is concerned fundamentally with *what works*. Linking these two concerns is a drive toward scale. We ask: "*What could scale up?"*

Getting to scale is an increasingly important mandate for educational research (Elmore, 1996; Fullan, 2000), but the history of educational change is largely the history of failure, either failure to be sustained or failure to reach what Elmore (1996) calls the "core"—what happens in classrooms on a daily basis. Innovations, particularly those involving technology, have often been cited as having high potential to address our largest learning challenges (President's Committee of Advisors on Science and Technology, 1997). However, vociferous critics point out the long history of failures of innovations to solve real problems in schools at scale, particular those involving technology (Cuban, 2003).

In mathematics education, technology has achieved an impressive scale already in the form of the graphing calculator. Evidence from the National Assessment of Educational Progress (NAEP) shows that “eighth-graders whose teachers reported that calculators were used almost every day scored highest” (National Center for Education Statistics, 2001, p. 144). This evidence is correlational and thus does not constitute proof that technology causes high scores (perhaps smart students are more likely to use calculators). The link between technology use and high test scores on NAEP does suggest that careful research about the effects of scaling up technology in mathematics education could address questions of great interest to parents and policy-makers.

The authors of this chapter affiliate with a group of researchers who believe that this failure or success of innovations in education is not the result of an intrinsic match or mismatch between technological innovations and educational problems but, rather, a failure of past research to address the problems of scaling up innovations (Blumenfeld *et al.*, 2000). Research on scaling up innovation, we argue, is different from research on “What works?” or “What could work?” Like “what-could-work” research, scaling-up research is concerned with the potential of new innovations that may achieve ends not measured in established tests. Like “what-works” research, scaling-up research is concerned with the presence of a systematic effect despite environmental variability. Thus, scaling-up research can be concerned with showing that an outcome is reproducible and measurable at scale despite the fact that the outcome (a) does not have an established benchmark already and (b) may occur infrequently in current conditions.

Research on scaling up innovation is not merely a matter of extending interventions to a larger number of subjects (Cobern, 2002). In addition to extending the number of subjects (n), research on scaling up involves an intellectual agenda with new foci, methods, and warrants coming into play over time. The research seeks explicitly to move from a concept to a scalable intervention. This research starts before an innovation is well specified (and thus long before conclusive evaluation research could begin). Researchers must resolve a large set of design details before research at scale can begin sensibly, while at the same time they must build evidence that makes a convincing provisional case that the research program is promising and deserving of continued support. This process takes a long time, spanning multiple, two-to-four-year, funding cycles.

By using a case study, we seek to clarify issues at the heart of research programs concerned with scaling-up innovation. We will analyze the case at two levels. First, we use the case to draw out characteristics of research programs that incorporate a drive to scale. The case will reveal how a program of scaling-up research is different from a program of basic research or evaluation research. Second, we highlight three driving questions that lie at the transition of projects from design research to implementation research. We consider what sorts of warrants are appropriate to link evidence to claims in the context of these driving questions.

The case study presents the research trajectory of SimCalc (http://www.simcalc.umassd.edu/software) , an innovation that supports learning the mathematics of change and variation. By presenting this case, we aim to ground our arguments in details about the evolution of a research program over time. We describe SimCalc research synoptically as occurring in six phases, each moving successively closer toward scale. We group the six phases roughly into two larger parts containing three phases each. The first part of the case is about research in the context of design. The second part of the case is about research in the context of implementation. We begin each section with a brief presentation of a theoretical framework that helps to organize the case.

PART 1: RESEARCH IN THE CONTEXT OF DESIGN

International comparisons (Schmidt et al., 2001) as well as national tests (National Center for Education Statistics, 2001) have highlighted the gaps between (a) the Americans' aspiration to provide world-class mathematics and science education and (b) the middling inequitable outcomes of the current for kindergarten through grade 12 educational system. To close these gaps, leaders and policy-makers call for an increased role for research in determining how to scale up new approaches to improving mathematics and science education (National Science Foundation, 2004; President's Committee of Advisors on Science and Technology, 1997).

Technology is often featured in discussions of opportunities to improve mathematics and science education (e.g., Bransford *et al.*, 2000). Prior literature particularly draws out the importance of leveraging the unique representational affordances of technology (Roschelle *et al.*, 2000b). By 1990, basic cognitive research on the use of representation in learning difficult mathematics and science concepts had concluded largely that the application of technology must be domain-specific (J. Kaput, 1992). This leads to a central design problem—the problem of designing *representations* particular to subject matter that unlock cognitive potential for learning.

This design problem is at the heart of the first part of the SimCalc case that we present here. As we shall see, the SimCalc team was deeply concerned with using cognitive (and later social cognitive) perspectives to improve mathematics education. Team members sought to employ the innovative potential of technology and they sought warrants at the level of mathematical cognition to relate evidence to claims.

In the first few years of the SimCalc project(J. Kaput, 1987a; J. Kaput, 1987b) (J. Kaput, 1994, 1997, 2001; J. Kaput *et al.*, 2001; J. Kaput & Roschelle, 1998; Nemirovsky *et al.*, 1998; Nickerson *et al.*, 2001; Roschelle *et al.*, 2000a), the researchers worked within a tacit framework, which we retrospectively call "Restructuring Knowing" (RK), to express the coherence across varied design challenges. This framework may be visualized as a Venn Diagram of three interlocking circles. The circles signify overlapping perspectives. The challenge of designing for restructuring knowing is to find overlaps that unite all three perspectives (Figure 1).

<Insert Figure 1 about here>

The three intertwined perspectives of the Restructuring Knowing design framework are:

1. Learners' Strengths. RK designs build on a developmental analysis of well-established cognitive, kinesthetic, and linguistic strengths and interests (as well as misconceptions) that learners can bring to mathematics, rather than merely their level of mastery of the current curriculum's prerequisite structure.

2. Representational Technology. RK designs emphasize the novel capability of technology to represent important domain concepts dynamically, in ways that

generate insights among learners.

3. Reorganized Curriculum. RK designs reorganize the curricular sequence to be more learnable and efficient, drawing on deep analysis of the historical precedents and the conceptual structure of mathematical ideas.

Each of these perspectives has a strong base in educational and cognitive theory. The perspective of learners' strengths is intrinsic to constructivism: if students are to construct knowledge, they must construct it from prior knowledge. The alternative (but less sound) view is that we could eliminate bad knowledge (misconceptions) and start with a blank slate (See J. P. Smith *et al.*, 1993 for a critique of this view). The perspective that technology provides novel representational capabilities also has a long-standing and distinguished theoretical pedigree (Roschelle et al., 2000b). Finally, investigators have noted that our present curricular structure is in many ways an artifact of the media available for representing concepts (J. Kaput *et al.*, 2002). We are not making the problematic argument that media alone affect the quality of learning (Clarke, 1983); rather, we argue that media enable and constrain curricular possibilities, and new media can give rise to radically reorganized curricula (J. Kaput, 1994, 1997).

Thus, the cardinal principle of RK design is to:

> Identify design elements that draw elegantly upon and unite the perspectives of learners' strengths, representational technology, and reorganization of the curriculum and its underlying epistemology.

One reason that design research is time-consuming is the difficulty of finding design elements that work from all three perspectives. Table 1 describes the findings of SimCalc (J. Kaput, 2001) (J. Kaput et al., 2001; J. Kaput & Roschelle, 1998; Nemirovsky et al., 1998; Nickerson et al., 2001; Roschelle et al., 2000a) research using the tripartite RK principle. We will refer to this table throughout the case study. We do not provide detailed descriptions of SimCalc's technology and curriculum here, but they are available in related publications (J. Kaput & Roschelle, 1998; Roschelle & Kaput, 1996; Roschelle et al., 2000a).

<Insert Table 1 about here>

Phase 1: Early Planning (Pre–1994)

The path to large-scale adoption of an innovation begins with the definition of a problem and an approach. The problem should be enduring because it likely will take a long time to arrive at a scalable implementation. The approach should draw upon at least one mechanism that plausibly has the potential to affect change at scale.

In the case of SimCalc, Kaput (1994) captured the core goal in the title as "Democratizing access to calculus" a concept which was later clarified as the "mathematics of change and variation." The essence of the SimCalc approach proposed by Kaput was to integrate new curricula with new dynamic representational capabilities made possible by computer technology. The argument for this problem and approach was made using historical, curricular, and literature synthesis methods. By using historical analysis, Kaput argued that changing representations play a fundamental role in determining what and how we think mathematically (J. Kaput, in press). Further, he recognized that reasoning about motion situations was central in the emergence of calculus (Kaput, 1994), but that calculus is taught today with only weak reference to motion phenomena. He noted that the dynamic representational features of technology (visual animation and simulation, in particular, linked to more conventional mathematical representations of functions and graphs) could bring motion phenomena back to the center of calculus education.

But Kaput (1994) did not want merely to improve the standard calculus course (indeed, a parallel but different calculus reform movement was occurring simultaneously). He argued that we are in the midst of a long-term trend requiring more students to gain access to more complex mathematical ideas. For example, whereas only a small percentage of students were expected to learn algebra a century ago, today, we expect all students in the eighth or ninth grades to learn algebra. Thus, he anticipated a need for a bigger and more diverse population of students to gain access to calculus earlier (Kaput, 1997). A curricular analysis revealed that the target concepts had been sequestered in a particular layer but could be unpacked into a strand that would begin in middle school (Kaput & Nemirovsky, 1995).

Finally, his literature review and synthesis (J. Kaput, 1992) suggested that dynamic representation was emerging as a powerful force for educational change. Further, it was clear that the technology to support dynamic representation in schools was becoming cheaper and more readily available and could become ubiquitous eventually. Indeed, in the period from 1990 through 1997, graphing calculators—a dynamic representational appliance costing $100 or less—have gone to scale. Likewise, over a similar time span, another dynamic representational tool, The Geometer's Sketchpad (Jackiw, 1988-97), has progressed from research prototype to the most used computer software in mathematics classrooms in the United States (Becker, 1998). Hence, using technology for its dynamic representation capabilities has turned out to be popular among mathematics teachers.

In the planning phase of the SimCalc project, Kaput formulated the above argument, created a movie to depict his concept, and brought together a team of consultants to help him plan a first research effort. At this time, the concept was limited largely to linking motion phenomena to mathematical formalisms (see the first row in Table 1). This idea, present in the video, is still foundational for the SimCalc program 12 years later. However, no other significant design features depicted in the video remain; later research phases discovered stronger alternatives.

Phase 2: Designing Representations for Difficult Mathematics (1994–1997)

In the second phase (1994-1997), the National Science Foundation funded a three-year SimCalc project. The team's objectives were of the "What could work?" variety—how to design representations for key concepts in the mathematics of change and variation that could help students learn. Thus, the research methods focused on microanalyses of very small numbers of students as they tried to learn these topics with preliminary materials and software. The team consisted of a mathematician, a cognitive/computer scientist, and a developmental psychologist–educational researcher. The technology used cost at least $10,000 per student to build because the team used it with only tens of students.

In the first year, the team explored a number of ideas to realize the vision of the original movie. These were presented as video games, a design element that has largely dropped out of the SimCalc program because none of these games proved compelling in small-scale trials with students. Indeed, small-scale explorations suggested that extensive game narrative and reliance on gaming goals stood in tension with the deep learning that the team sought. Neither large-scale research nor a sophisticated methodology was needed to rule out innovations with little potential.

A defining transition came when the team realized the powerful synergies represented in the third row of Table 1, which occurred sometime in our second year of research. Through microanalytic developmental studies, Nemirovsky and colleagues (Monk & Nemirovsky, 1994; Nemirovsky, 1996) found that students reason naturally about motion sequences by breaking the complex motion into intervals. Interval-based analysis is more natural to them than continuous-function-based analysis. In technological explorations influenced by his cognitive science training, Roschelle found that computers made “direct manipulation” (Shneiderman, 1982; D. C. Smith *et al.*, 1982) of piecewise-defined functions easy to represent; potentially, students could “drag” the shape of a piecewise velocity or position graph to a desired shape and see the consequences as a simulation. The project lead, Kaput, saw the opportunity to create a more learnable curricular sequence by exploring inversions of the normal prerequisites. For example, in the normal sequence, position graphs are taught before velocity graphs, symbolic algebra is taught before integration or differentiation, and continuous functions are taught before piecewise functions. Through the use of directly manipulable piecewise functions, all these normal prerequisites could be eliminated, along with the traditional stumbling block at the onset of a calculus course—the definitions of continuity and limit processes.

At the end of this phase, the team was able to build on this insight (and the related ones in Table 1) to develop a fairly complete software tool, called SimCalc MathWorlds, in about a year. Kaput's slogan at the time was "software without new curriculum is not worth the silicon it's written in." Much parallel exploration sought to identify curricular structures that could build off these newly recognized learners' strengths, representational capabilities, and curricular starting points. The SimCalc team carried out these explorations in locally available classrooms with Kaput and other project staff teaching the courses because the team did not have time yet to define suitable professional development resources for teachers. The researchers gathered much video and field-note data, which provided a compelling demonstration that students were learning. Pretest and posttest data showed strong gains. But a rigorous, randomized, experimental trial would have been very hard to plan and execute: What conditions would the team try? What would make an appropriate control? Too many variables were still in flux, and the research group was not sure yet of what the most important ideas of the innovation were. Importantly, we had not defined yet a stable innovation package (materials, teachers' professional development, curriculum, assessment) that could be tested in a normal classroom setting. Phase Two was very much a design experiment deliberately probing new territory.

Phase Three: Designing Technology-Rich Curricula (1997–2000)

In a third phase (1997–2000), with a second round of funding, the group began to tackle additional issues of teachers' preparation, curricula integration with core curricular sequences and newly important state and national standards, an assessment framework, and more modular software. The research began to focus on replicated classroom design experiments, each with a defined but varied curriculum and professional development for the teachers. The group was not seeking to show that one package worked across all settings but to understand the variance in settings and the range of curricular approaches and teachers' training that might work. Consequently, we expanded the team to involve researchers at Rutgers University–Newark (New Jersey), Syracuse University (New York State), and San Diego State University (California) and to engage teachers and schools in Boston and the southeastern Massachusetts region. Under the theme of democratization of access to important ideas, the target schools and populations were those where students were least likely to have access to calculus in their academic futures, which in turn meant that students at the (largely urban) sites typically had low or very low socioeconomic status. The explicit assumption was that "if we can make it work here, it can be made to work anywhere."

The research team now included teacher–educators, master teachers, and more frequent interaction with commercial publishers of similar innovations. We redesigned our software to be more flexible, to run on as wide a hardware base as possible, and to the extent it was technologically possible, to continue to allow low-cost, quick prototyping of an ever-widening range of uses. In our design experiments, we deliberately included studies with varied technology, including motion detectors, devices that translated graphs into physical motion, and machines that enabled the exploration of chaotic motion. Versions were built eventually for desktop computers, both Mac OS and Windows, as well as handheld computers, the Palm OS, and the world's most popular platform for school mathematics, the TI-83 Plus graphing calculator. These technological investigations embodied the adaptability of SimCalc's representational features to varied, inexpensive, and readily available platforms, a prerequisite for scale.

One key accomplishment of this phase was independent replication of the value of SimCalc; independent researchers tried it in varied settings and reported their video, field-note, and pretest and posttest data. Importantly, however, these were not "replications" in the traditional sense of replicating an experiment. The core SimCalc curricular insights and technologies were used in each case, but in localized packages of software-based lessons, curricula, teachers' professional development, and assessment. The implementation "package" was not standardized or replicated yet. Further, because the content could be addressed at levels from middle school to university calculus, parallel implementations were less important than implementations that explored the wider curricular range of the innovations. In this way, the team was learning about the variability of settings, implementers' preferences, opportunities and constraints, and alternative ways of packaging the materials, in preparation for tailoring its possibilities in the most advantageous manner for current purposes.

Another important accomplishment was detailed analysis and experimentation with how SimCalc materials could connect into, enrich, and render more learnable the core content in the standard curriculum. This is a critical issue for scale, because an innovation that is too distant from curricular reality will not be adopted widely. The team defined and elaborated a strategy for both improving the learning of important but conceptually difficult topics in the curriculum already and simultaneously adding new opportunities to learn the mathematics of change.

In this phase, we worked on the conceptual challenges presented in the latter three rows of Table 1. For example, we came to understand that a key aspect of SimCalc across all of its implementation was how the materials emphasized, advanced as a problem, and worked through the relationships among rate and accumulation representations. From the developmental theory and learners' strengths perspectives, Walter Stroup joined the team and brought a Piagetian focus to the question of how children come to understand the difference between "how much" and "how fast"—the core developmental dilemma that SimCalc addresses (Stroup, in press). Technologically, we related this to one of the most cited benefits of technology in the cognitive science literature: the use of multiple representations (Goldenberg, 1995; Kozma *et al*., 1996). But whereas "multiple representation" is quite a generic term in that literature, for SimCalc it came to emphasize the ability of students to think fluidly with and between both position and velocity graph representations. During this time, we explored many other representations but rejected them as more problematic, less powerful, or too difficult to fit into school mathematics within existing constraints, for example, phase–space descriptions of motion or related quantities. On the curricular side, the SimCalc group came to understand that a key powerful functionality of SimCalc was "snap to grid"—a technological ability that could be used to limit graphic adjustments of functions to whole number values. This has the curricular benefit of producing a calculus in which computations can be guided geometrically and executed within the arithmetic of whole numbers and simple fractions, instead of algebraically.

Thus, by the end of this phase, the team was increasing its warrant for the belief that the underlying concept was strong and adaptable to a wide variety of settings. We had a good understanding of which things among the many that varied across settings were essential to success in all the settings, and which of the many ways to align the learners', technological, and curricular possibilities brought the most benefit within existing school constraints (e.g., technological, curricular, and, to a lesser extent, teachers' capacity). Such understanding is critical to the issue of what to try to scale up among the many variations of materials that are invented and tried.

Part Two: Research in the Context of Implementation

Toward the end of Phase Three, the concept of scaling up was in the air. The Internet was beginning its exponential explosion in availability and use. Commercial learning tools specific to mathematics, such as The Geometer's Sketchpad and the TI-83 graphing calculator, were showing market success. Further, researchers and policy makers in the community funded by the National Science Foundation were beginning to ask questions about how innovative research materials and practices could go to scale.

The SimCalc team had always been driven by a vision of "democratizing access to the mathematics of change and variation"—a vision that implies scale, especially to include disadvantaged populations. Two strategies were developed to move toward scale. One, encouraged strongly by the funders and led by Kaput, was to make SimCalc commercially available—a move toward sustainability of the effort required to keep materials available and up-to-date. The other strategy, led by Roschelle, sought to engage in research relating to scale. Early on, the team recognized that it was not ready to engage directly in large-scale research; thus, it sought and won a planning grant, followed by a

Phase One grant and a Phase Two grant, all under the Interagency Educational Research Initiative.

Although the restructuring knowing framework had served well to date, scaling up exposed new design challenges, and a new framework was needed. The team turned to the work of Cohen and Ball (1999), who had formulated a simple elegant theory of scaling up classroom innovations (D. K. Cohen & Ball, 1999; D. K. Cohen *et al.*, 2003). Their theory called attention to the trade-off between the degree of ambition in an innovation and its degree of elaboration (or careful specification). Innovations that are not ambitious are close to current classroom practice and thus need little specification. On the other hand, ambitious innovations like SimCalc that deal with deep changes in representational infrastructure require a great deal of specification. Coupled with the growing need to articulate our innovations to varied audiences, particularly teachers, this work helped us realize that our innovation was not specified sufficiently yet.

Further, the Cohen and Ball (1999) framework is classroom-centric; it emphasizes features of classroom learning in a school environment. In essence, Cohen and Ball call attention to education as an interaction among teachers, students, and resources (i.e., textbooks, software) in an environment (Figure 2). Relative to this triangle, Phases One and Two had dwelt mostly on the student-to-resource relationship and had bracketed the environment completely. Phase Three began to explore the teacher-to-resource relationship through design experiments in preservice and in-service professional development for teachers. Elements of the environment such as curricula, assessments, and teachers’ professional development were beginning to come into focus weakly. Arguably, the environmental features we understood best at the end of Phase Three were the nature of available technical platforms that could support widespread dissemination and use, and, to a slightly lesser extent, the curricular constraints at work across the country, constraints that increased in salience during Phase Three due to the rapid rise in accountability systems across many states.

<Insert Figure 2 about here>

We now continue the case study through three somewhat shorter phases, using the Cohen and Ball (1999) framework to organize the details.

Phase Four: Planning for Scaling Up (2000–2002)

In a fourth phase, starting in 2000, the SimCalc team began to work intensively on the issue of scale. We began our planning effort with a substantially expanded team, supplemented by expert advisors. The two new additions to the team were an experimental psychologist and an assessment development expert. The experimental psychologist brought discipline to the process of developing a research question and experimental design. The assessment development expert helped us formulate an assessment blueprint. Through discussion with our expanded team, reading the literature, and meeting with our advisors, we became aware of the many dimensions of the scaling-up challenge.

This phase consisted entirely of planning and preparation; no new empirical research was performed. In the course of planning, the team sought to address:

- broadening the team and grounding in scale-up literature
- specifying the innovation

- refining the experimental design
- choosing an implementation partner.

We discussed each point in turn. A key message of this phase is the huge number of decisions required for a transition from design research to experimental implementation research. It takes a substantial planning effort to work through these decisions.

Curriculum

The curriculum is a dominant factor in what teachers teach and what students learn. At scale, we have the choice of either integrating with the variety of textbooks in use or writing one of our own. Textbook authors and publishers, we found, were not interested in working with us to integrate our materials deeply. They perceived their existing constraints as barely manageable and were reluctant to take on new ones, especially ones that involved technology and that might add yet another barrier to adoption. Writing a new textbook would be very expensive and time-consuming.

Teachers' Professional Development

Teachers require significant support to learn to use ambitious materials like SimCalc. The literature suggested that a major trade-off exists between short-term, relatively isolated support for teachers and long-term support that is highly integrated with school structures and communities. The latter is clearly better but would introduce major expense and place the measurement of results far into the future.

Systemic Reform

The National Science Foundation had made a major investment in systemic reform as well as standards-based curriculum development. Many researchers held the belief that it made sense to test an innovation only in the context of systemic reform.

However, this clearly would limit the generality of the findings because only a small fraction of schools engage in systemic reform and subject our relatively simple experiment to the uncertain results of a larger, more complex experiment.

Standards

Also during this time, decision-making at the school district and school level was influenced increasingly by national and state standards. SimCalc had not been developed initially to address a specific standard but, rather, to address a conceptual strand that spanned many years of development—the mathematics of change and variation. We could adapt our innovation more narrowly to particular state standards but potentially move the innovation away from its core strengths balancing mathematical fidelity, attention to student learning processes, and pragmatism about the classroom , or we could connect only weakly to standards but potentially face major problems in recruiting school districts and schools.

Assessment

The most reputable assessments are those used for the most important testing by states, comparisons among states, and comparisons among countries. Yet, assessments tend to be very conservative in the content they cover. We could find very few items on standardized tests that cover SimCalc's topic directly—the mathematics of change and variation. Thus, we could take the risk of using either an established test with poor alignment to our innovation or a less established test (that we would have to construct ourselves) that was well aligned.

Grade Level

SimCalc materials had been developed and tested primarily for use in mathematics classes in grades 7 through 11, although they had been used with students as early as the elementary grades and as late as university calculus. Optimally, the creators envisioned the program as longitudinal, with students revisiting and deepening their understanding of an important strand of mathematics across many grade levels. But a multiyear longitudinal experiment seemed too risky because we would have to commit to so many details without feedback at the beginning. Furthermore, even one round of testing would exceed the likely available budget. Thus, we had to choose a particular grade level.

An experiment consists of many features that are not central to the hypothesis but need to be implemented satisfactorily nonetheless. A methods professor might call the process of getting these details right the “art of experimentation.” In part because of these factors, one wants to aim for simplicity, supplemented with replication and extension of the crucial points. Because of the desire for simplicity, specifying the innovation was harder than it might seem. After six years of research, the team knew a lot about the adaptability of SimCalc to many different student populations, teacher styles, textbooks, state standards, etc. We needed to winnow the innovation down to a core testable intervention. The team asked itself: “What is the potentially implementable essence of SimCalc, and what is worth subjecting to a rigorous test?”

The essence of SimCalc was a serious question because, by this time, we had software running on the Mac OS, in Java, on graphing calculators, and on the Palm OS, each with subsets of different feature. We had many versions of curricular and teachers’ professional development materials covering topics ranging from the idea of rate in middle school to the fundamental theorem of calculus in university calculus. We had a very large collection of assessment items that we had tried over time. If we defined SimCalc too simplistically, as a particular software version and curriculum materials version, then our results might not generalize to the most important version of software, materials, teachers' professional development, and assessment items. In the end, we did not start by specifying the innovation; rather, we addressed all the other points first. With these constraints in place, specifying the details of the innovation became much easier. (The case study will return to the issue of specifying the innovation in Phase Five.)

Our planning came to focus on: "What was worth subjecting to a rigorous test?" Three possibilities came to mind:

1. A test parallel to the force concept inventory (Hestenes *et al*., 1992).
2. A full-year curriculum in the eighth grade.
3. A replacement unit strategy in multiple grades.

Our reading of the work of Cohen and Hill (2001) suggested the last alternative—a replacement unit strategy. Professional development centered on replacement units had been somewhat successful at a large scale in mathematics reform in California. We were particularly attracted to the idea that replacement units offered good opportunities for both teachers' learning and students' learning. Further, SimCalc had developed some materials at each grade level but not enough materials to cover an entire grade level course. A replacement unit strategy could allow us to test a carefully targeted subset of our materials at each grade level, consistent with SimCalc's multigrade strand orientation.

Several additional considerations brought us to the conclusion that a replacement unit strategy was our best choice. First, although a longer intervention promised more significant benefits than a shorter one, we doubted that we could persuade a wide variety of teachers to try something unknown, which potentially departed significantly from existing standards and which required substantial use of scarce computing resources. Second, a long intervention meant that we would have to put tremendous initial emphasis on materials development. Third, a long intervention would mean more difficult implementation metrics; that is, it would be harder to monitor whether and how teachers complied with the condition. Fourth, a longer implementation would mean that teachers would need more ongoing support, exceeding our likely budget. Our feeling was that if SimCalc was successful under experimental conditions, then we would be positioned later to ask questions about how much support was required for successful, long-term adoption.

In this context, it was finally possible to pose a specific research question. We had shown already positive pretest and posttest results for students in many different settings. Most of these settings were economically disadvantaged, but one was quite affluent. As we considered how to pursue the question of students' learning across settings, we began to focus on teachers. In most of the prior work, we worked with "convenience samples" of teachers—teachers who were, if not enthusiastic, certainly interested in our approaches, located near a major research institution, and supported throughout the year by researchers or their staff (who were, on occasion, more experienced SimCalc teachers). As the Cohen, Raudenbush, and Ball (2003) triangle highlighted (Figure 2), a focus on student-to-materials interactions was only one leg of the classroom learning process. We knew the materials were ambitious and at more risk of failure than materials closer to existing teaching practice. We thus identified an important uncertainty about "What could scale up" relating to teacher-level variables. A critique we often heard also influenced our increasing emphasis on teachers, a critique we paraphrase as: "Sure it works with your boutique teachers, but it won't work with the teachers in my school." We concluded it would be an important step toward scale to show that SimCalc could work with a wide variety of teachers.

The focus on teachers necessitated further refinement of our expectations. Although our major hypothesis was to predict a main effect for SimCalc compared to no SimCalc across a wide variety of teachers, we had to ask whether we expected all teachers truly to benefit. Two factors conditioned our expectations: one was the belief that students with the "strongest" teachers and sites would be able to make the best use of the intervention, the other was the belief that students with the "weakest" teachers and sites would benefit the most from the clarity of the interface and the depth of their experience. "Strong" here means high socioeconim status, well-prepared students, a low percentage of students with limited proficiency in English, teachers with experience and a positive attitude, a functional school, and so forth. "Weak," at the extreme, means the opposite of all these criteria.

Three alternative models deriving from these factors were that: (a) the strong gain more, (b) the weak gain more, and (c) the very strong and the very weak would gain more than the middle. Our democratization prediction, and our hope, was that students of all kinds of teachers would gain with SimCalc but that students of weak teachers would narrow the performance gap.

Following from the selection of a replacement unit strategy and a wide variety of teachers, it became clear that we would need to build our own assessment. Nonetheless, to increase validity, the team vowed to use items from validated tests to the greatest extent possible. An assessment blueprint was developed that specified how we would seek to measure outcomes. It also became clear that we would have to explore a relatively limited professional development strategy for the teachers rather than the longer-term strategy most often recommended; it would not make sense to scale up a three-week replacement unit that required three years of teachers' professional development. We decided we would need an implementation partner with existing outreach to a wide variety of teachers and a track record of success with shorter-term professional development.

At this point, we discovered that the Dana Center at the University of Texas, Austin was leading mathematics teachers' professional development for the state of Texas. Teachers and administrators appeared to trust the Dana Center and Dana Center personnel had long histories of teaching mathematics and working with mathematics teachers. In the 2000–2001 school year, 21,000 mathematics teachers attended professional development workshops designed by the Dana Center. Further, the lead of the Dana Center had a long-standing interest in the strand of mathematics leading to advanced placement calculus and had performed much work to align Texan standards across grades to support the goal of democratizing access to advanced placement calculus. Thus, it would be easier to align SimCalc to Texan standards than those in some other states. The data from this census are made public. We met with Dana Center leadership, and they became interested in joining forces for a scale-up research project.

The only downside was that they wanted us to focus first on the seventh grade because an eighth-grade intervention was too close to the highly charged, eighth-grade state test. We perceived the seventh grade as less promising than the eighth grade for showing SimCalc's impact, but we agreed to start with the seventh grade and expand longitudinally to the eighth grade.

Thus, after nearly two years of planning, we came to a set of compromises and decisions that enabled us to plan our first scaling-up research project. Our deliberations had resulted in a lower risk plan with regard to curriculum (a replacement unit), grade levels (only slightly longitudinal), alignment to standards, and a research question (focusing on an important mediating variable, the teacher), but a high-risk plan elsewhere. Providing only short-term, relatively bounded, professional development for teachers, developing our own assessment, targeting the seventh grade, and not attempting systemic change all seemed to lower the odds of finding a systematic effect in a controlled experiment. With these decisions in place, the team set out to test them and our organization through a pilot experiment.

Phase Five: Preparation and Pilot of an Experiment (2002–2004)

Preparing and piloting an experiment was a complex process requiring another two years. In this process, we had to return to the problem of specifying the innovation and solve it. Further, the details of the experimental design required definition; a particularly thorny problem was defining the control group. Next, the team had to identify and/or create measures for all the variables of interest. As will become clear, we needed either to find or to create instruments to measure each aspect of the Cohen and

Ball (1999) framework. Last, but hardly least, we had to recruit teachers to participate in our pilot. We discuss each aspect of the process below.

Given the decisions of Phase Four, the team knew it needed a replacement unit for the seventh grade with accompanying professional development for the teachers. The selection of the specific topic was made by examining the Texas state standards (http://www.tea.state.tx.us/rules/tac/chapter111/ch111b.html). Rate and proportionality were highlighted in the standards:

> Within a well-balanced mathematics curriculum, the primary focal points at Grade 7 are using proportional relationships…. Students use algebraic thinking to describe how a change in one quantity in a relationship results in a change in the other; and they connect verbal, numeric, graphic, and symbolic representations of relationships. p. B-5

We should point out that the curricular choice, constrained as it was by the state standards, methodological needs, and time and resource limits, left out a large amount of previously developed and tested content and technological capacity. We avoided velocity functions, velocity–position connections, and their generalization across other, nonmotion quantity types, algebraic notations, and, indeed, most of the mathematics of change.

A middle school mathematics curriculum designer with extensive experience in teachers' professional development, Jennifer Knudsen, joined the team. She reviewed the existing SimCalc materials, whose coverage of the target topics was not in the form of a self-contained unit, did not contain the needed supporting materials for the teachers, and were lacking the requisite production quality. Hence, Knudsen wrote a new workbook and Kaput's team adjusted the software to match. The materials were designed to be explicit about both the teachers' and the students' tasks. The materials also were designed to be easy to comprehend in a short amount of time, using text sparingly, for example. We note that although we were still *designing* in this phase, there was no design research. Rather, the team trusted in the accumulated wisdom from the past phases of research to guide the specification of the innovation.

Equally important to specifying the innovation in an experiment is specifying the control condition. Choosing an inappropriate control could render the experiment meaningless. Thus, the distinction between a "What works?" question and a "What could scale up?" question becomes critical. The former tends to lead to control conditions that provide an existing benchmark for students' gains against which the treatment could be shown to lead to stronger gains. The latter, however, may not have an existing established benchmark for students' gains, as reflected in the word "could."

The SimCalc project was interested in showing that students could learn more complex and conceptually difficult mathematics than is measured typically. Moreover, we had narrowed down this extra learning to a particular context: a replacement unit on rate and proportionality. The question of "What could scale up?" had turned into the demonstration that a suitable effect size could be obtained in the presence of significant, randomly distributed, variation in context. The control condition would serve to guarantee that the observed effect was due to the intervention, and not due to, for example, merely taking the same test again three weeks later.

A consequence of this decision was that there would be many "What works?" questions that remained unanswered by the design. For example, the design would not show that technology works better than no technology because we did not insist that teachers in the control group use no technology. We also could not assert on the basis of this experiment that SimCalc is the "best" approach to teaching seventh-grade rate and proportionality or that we have the best approach to teachers' professional development for using technology in seventh-grade mathematics.

The best control for this interpretation of our study, then, might have appeared to be providing no intervention for teachers in the control group and merely measuring any gains by students across two administrations of the same test. However, in designing an experiment, one has to consider the Hawthorne effect: the idea that knowing one is in an experimental condition produces beneficial results. Thus, we had to create a control condition in which the control teachers would feel similarly that they were engaged in meaningful professional development. Then, we could argue, using standard experimental logic, that any differences between the control and the experimental groups were due to our intervention.

Happily, the Dana Center had developed a highly regarded workshop for middle school teachers on teaching rate and proportionality. The force of the workshop was to encourage the teachers to take a $y = kx$, or rate-based, approach to proportionality. This approach was complementary to our own. To ensure that we were treating the groups equitably, we decided to teach this workshop to both the experimental and the control teachers. The nature of our control condition was influenced also by the idea that, as in medicine, few people want to be in the control condition. Therefore, we decided to give the experiment a "delayed treatment" structure in which we promised to give all the teachers the SimCalc treatment condition in the second year. Thus, all participating teachers eventually received all the benefits of participating in the experiment, with the delayed teachers waiting a year for the SimCalc materials.

With this design in hand, the team's next step was to identify existing measures wherever possible and to design new measures only where necessary. We sought measures for each aspect of the Cohen and Ball (1999) framework (Figure 2). First consider the vertices of the triangle. We decided to measure students' gains (our major outcome variable) using a paper-and-pencil test. We likewise decided to measure teachers' growth in content understanding, a secondary outcome variable and a possible predictor of students' learning. The quality of the materials was "measured" by vetting them with experts from the SimCalc team, teaching experts in Texas, and an experienced editor of mathematics textbooks.

Next, considering the three sides of the framework (Figure 2), we decided to observe students' interactions with the materials by collecting the complete workbooks at the end of the unit. We observed teachers' interactions with the materials by collecting a daily log in which teachers were asked to record what they did with the materials each day. We captured teachers' interactions with students using an observation protocol based on Schorr's protocol (Schorr *et al*., 2003) and through videotape analysis (Schorr was an advisor to the project). Each of these three connecting lines represents a potentially significant, mediating variable in implementations (i.e., inappropriate teacher–student interaction could explain a failure to produce gains for the student). Finally, we sought to measure many elements of variation in the environment. Given our hypothesis, we first wanted to measure how teachers varied—in their background, their attitudes, and their school setting. We collected background data using an application form, the existing

Teaching, Learning and Computing Survey (Becker & Anderson, 1998), school-setting information from Texas Education Authority data sets, and our own more specific attitude scale.

Each of these measures required a significant effort to select or develop. A discussion of the process for developing some of the intermediate and outcome measures appears elsewhere (Shechtman, Roschelle, Haertel, Knudsen, & Tatar, 2005). Finally, with a well-specified intervention, an experimental design, and the measures all in place, the final step was to recruit participants and run a pilot of the experimental design. The results of our pilot phase have been reported elsewhere (Roschelle et al., 2005); the important point for the present context is that the findings were both positive enough and raised enough interesting questions to warrant a scale-up study.

Before closing this phase of the presentation, we call attention to the issue of building trust within a multi-institutional team. Executing research on scaling up requires a large cohesive team. It would be too easy to see the story of Phase Five only in terms of a rational research design process. It was also a story of identifying and working through tensions and conflicts. Although it is somewhat of an oversimplification to present it this way, each of the three partner institutions brought a different main interest to the study. The University of Massachusetts had the strongest interest in the integrity of its innovation. The Dana Center had the strongest interest in serving teachers in Texas. The SRI International team (including Tatar at Virginia Tech) had the strongest interest in clean research methodology. These interests were often in tension and sometimes in conflict. For example, to obtain sufficient power required treating teachers as individuals across varying school contexts, but the best practices of professional development would work with teachers in teams, and SimCalc had worked most often with teachers in schools undergoing systemic reform. Project members also met with regional education leaders in Texas and developed further mutual understandings about the nature of school districts across Texas and the kinds of constraints and practices in place across the various groups of school districts. A very important component of working through a pilot experiment together is building the relationships needed to form a cohesive team.

Phase Six: First Scale-Up Experiment (2005–2008)

Phase Six will execute SimCalc's first scale-up experiment, starting 12 years after the first research began. The design for the scale-up experiment involves 120–140 seventh-grade teachers and 70–80 eighth-grade teachers. At this scale, any mistake is extremely costly. The agility of the Phase One and Two design processes, in which anything could be tried quickly and cheaply, has been replaced now by very careful, deliberate, slow processes. In Phase Six, SimCalc added two advisory boards to check each step of the process. The first advisory board is reviewing each significant feature of the experimental design. The second advisory board is reviewing the mathematical content.

There are relatively few changes from Phase Five to Phase Six except in recruitment. Whereas the pilot drew teachers randomly from across the state, the full experiment will recruit teachers through Educational Services Centers in particular regions. By concentrating on a few regions, we maintain significant diversity but avoid skimming only the best teachers from across the state.

With regard to experimental design, a statistician has joined the team. Hierarchical linear modeling will be employed more systematically through the design, sampling, and analysis processes. In addition to the experimental contrasts, an education researcher whose area of study is focused on teachers joined the team and will conduct case studies embedded in the main experiment to examine questions that are not immediately amenable to experimental treatment.

WHAT WE ARE *NOT* DOING IN THIS FIRST EXPERIMENT

We call this the "first scale-up experiment for SimCalc" because there is so much left to do. This experiment will occur only in one state at a few grade levels. It will not consider longitudinal impacts on students. It will not consider the implications of extending dynamic representation technology across the curriculum, for example, to data analysis and geometry. It will test only a small subset of SimCalc's curricular scope. It will not consider the relationship between this innovation and systemic reform. It will not consider how gains or losses on proportionality topics interact with gains or losses in other seventh grade topics. It will not test the usefulness of technology compared to no technology but similar curricular materials. It will not test how much professional development for teachers is required for the benefits to accrue. It will not test even whether teachers who resist participation are still able to work with the innovation because participation is voluntary.

We term this the first scale-up experiment for another reason as well. This one experiment cannot determine whether the core innovation—building on students' cognitive strengths using dynamic representation integrated with a restructured curriculum—is "what works" or is a failure. It could be that the experiment will fail, but for reasons of not scaling up correctly rather than the weaknesses of the underlying innovation. For example, a replacement unit strategy may be an incompatible way to scale up the SimCalc approach. On the other hand, if we get results, as in all experimentation, replication and extension will be required for confirmation. After all, perhaps our teachers' professional development or the materials themselves will not scale up. Furthermore, although we have approached what we consider the most tractable, practical, and meaningful questions in this experiment, it may turn out that publishers and policy-makers want other kinds of questions answered before adopting this innovation. Indeed, SimCalc cannot scale up significantly without publishers who are willing to produce and/or integrate the SimCalc curriculum.

Finally, the technology and the larger educational environment (the political and economic dimensions, especially) continue to evolve so that optimizing to current conditions is not likely to serve the longer run without corresponding updates to account for the continuing evolution. For example, wireless classroom networks linking students' handheld computers and a teacher's workstation have become a major technological factor in ongoing SimCalc work, but they are not part of the current experiment because neither the technologies nor the curricula are sufficiently mature to support the level of precision sought in the current experiment.

Case Summary

Many trends can be observed in this case study. First, it is long, covering more than 10 years and five different funded "projects." During this long timeline, the overall drive to scale can be seen in many dimensions of the case study. We note a gradual expansion of the research focus from students' interactions with novel presentations of content to include teachers' interactions with materials and teachers' interactions with students. Over time, more factors in the environment come into focus and under control or measurement: curricular integration, teachers' professional development, assessment, state standards, regional teachers' service centers, etc. Further, the number of students expands by orders of magnitude, from small numbers of students, to small numbers of teachers, to schools to statewide regions. In parallel, efforts were made constantly to reduce costs and increase access to the technological prerequisites. The team expanded over the years to cover more disciplines, eventually including people with over 10 kinds of disciplinary expertise and high-profile advisory boards.

In concert with these changes over time, the intellectual agenda and methods evolved. Work in Phase One had no empirical methodology; the basics of the design relied on historical, curricular, and literature analyses. Phase Two used primarily microanalysis of a small number of students. Phase Three involved design experiments, with pretest and posttest measures. Phase Four was a planning phase and not particularly methodological. Phases Five and Six focused on an experimental design but include embedded case study analyses. At each phase, there are still a huge number of design questions to resolve, more than can be settled using only benchmark methods. Research on scaling-up innovation is a complex enterprise.

A particularly important and difficult transition occurred during Phase Four, when the team finally had to confront the question: "What is the essence of SimCalc and what is worth subjecting to a rigorous test?" We suspect this transition was difficult because of the lack of overlap between design researchers and experimental implementation researchers. Thus, although copious effort in the first three phases of SimCalc had addressed questions of scale, too little effort had been devoted to the eventual demands of scale-up research. In retrospect, we would have put more effort earlier into firming up measures and instruments and documenting better potential recruitment difficulties.

We also wish to emphasize that carrying an innovation through this level of growth and transition requires leadership and vision. The project's mission never varied from "democratizing access to the mathematics of change and variation" (however, the specifics of how the vision was addressed did change: although "rate" was an early focus, "proportionality" itself was not). Throughout, the team sought to reaffirm that it was addressing important mathematics, both in terms of present assessments and in terms of future needs. It also sought consistently evidence that the technology works in traditionally underserved settings. Finally, the role of technology, the basic form of the SimCalc representational system, and the commitment to integrate technology with textual curricula have remained unchanged since 1995.

Generality and the Limits of the Case

We suspect that the long time and the multiproject nature of scaling up educational innovation is typical. Indeed, it is common to think of five-to-seven-year school adoption cycles, and some educational reforms take multiple decades to implement (Schoenfeld, 2002). Similar innovations, like The Geometer's Sketchpad, also took a long time to reach scale, even with a more direct route from research to commercialization (Scher, 2000). Perhaps the most successful educational technology product, the graphing calculator, took about seven years to transition from the initial product to 20% market penetration, which does not count the development of the research base on which it rests (Ferrio, 2004). Long timelines, requiring spanning multiple, two-to-four-year project cycles, are more likely to be the rule than the exception. We note that without the earlier, less structured exploration, the later concern with scaling would have had nothing to bring to scale!

We expect that the two frameworks we used to describe the case are quite general. Cohen and Ball's (1999) framework is fairly simple and has a straightforward ring of truth to it. Our own restructuring knowing framework is purposely specific to the style of the technical innovation that we were pursuing, but we would include in this class the many software products that emphasize the computer as a tool for making visual manipulable representations available in the mathematics classroom.

In addition, many elements of the case fit a third framework. Rogers (2003) has organized and summarized a multidisciplinary body of work on the diffusion of innovations. One can view this case as a series of moves to increase the potential rate of adoption of SimCalc. Rogers argues that five factors influence the rate of adoption. SimCalc has always been strong in *observability* (1); after a short demonstration of the software, most teachers express the feeling that their students could learn more easily with the graphic and animated representations that SimCalc provides. But SimCalc has been weak in *compatibility* (2). Indeed, SimCalc's early aim was at a strand of the curriculum, the mathematics of change and variation, which is not important currently to the average mathematics educator. Starting in Phase Three and continuing through Phase Five, the team worked hard to connect SimCalc more closely with the mathematics topics, standards, and curricula considered important by most mathematics educators. Further, SimCalc has been somewhat difficult for teachers in terms of *complexity* (3), requiring the extensive use of technology that is often unfamiliar to teachers. Some of this complexity is irreducible—the technology is essential—so efforts focused on making the curricular materials and training workshops as clear and simple as possible. A related effort, especially in Phases Four and Five, was to increase the *trialability* (4) of SimCalc by defining a trial-use unit of implementation. We found that a replacement unit with a very clear scope and place in the standard curriculum made it easier to recruit teachers to try the software for our studies. Finally, one can read the trajectory of the case study as a continuous effort to develop research-based messages that communicate strongly the *relative advantage* (5) of SimCalc.

In Part One, we communicated the relative advantage through case studies and the restructuring knowing framework. The team was able to show ordinary students learning more complex mathematics and to articulate the potential advantage of using new representational capabilities to draw upon learners' strengths and reorganize the curricular content to be more learnable. In Part Two, the team wished to make causal claims about SimCalc's relative advantage and to support generalization; hence, the methodological shift to controlled experiments and more carefully defined outcome measures. An additional resonance with the Rogers' framework is the alliance of *change agents* and *opinion leaders* that we formed for the work in Part Two. Our alliance coupled (1) Kaput, a math educator who has worked as an external change agent seeking to influence the use of innovations in local schools with (2) Hopkins at the Texas Educational Service Centers, a math professional development expert, who has been an opinion leader in the adoption of innovations in mathematics education in Texas. Finally, from Phase Two onwards, SimCalc has been designed to be highly adaptable by users to specific settings, a capability that Rogers (2003) terms *reinvention.* Reinvention increases adoption by allowing users to make an innovation fit their local needs. In Phase Four and

Five, we “reinvented” SimCalc to fit the needs of the seventh-grade curriculum in Texas (although still drawing upon the essential innovation). Without being conscious of this framework, it appears that the SimCalc team made a series of moves that fit well-documented patterns in the diffusion of innovation.

Despite the apparent applicability of these frameworks to related innovations, it would be a mistake to overgeneralize from this case; the field needs to consider other cases of scaling-up innovation beyond our own before reaching firm conclusions. Although it is likely that many scale-up projects will require increasingly multidisciplinary teams, projects that start with experimental psychologists might take different paths than projects that start with a mathematician. Further, the funding climate and research policy are always changing. Most likely, if this effort had started today instead of 12 years ago, the team would have taken a shorter path to estimating effect sizes, for instance. Finally, because we did not have the luxury of outside observers, this case study was written by core members of the team. An objective observer might have produced a different account.

Discussion of Research Questions and Warrants

The case of SimCalc makes it clear that multiple kinds of research questions are asked in a scaling-up project over time and that multiple methods are employed to answer them. Yet, this case potentially has a stronger message than merely that multiple methods are acceptable as long as they match the questions asked. Indeed, the major impact of this case may be to clarify what kinds of questions *are asked* in an innovation project. These are not merely “What is happening?” or “Is a systematic effect present?” Nor are the questions asked merely "What could work?" or theory development questions. We suggest that the many specific research questions condense to three categories, described below. In covering each category, we discuss the form of the warrants that might be appropriate in supporting answers.

Question One: "Is the Candidate Innovation Well Specified?"

Researchers studying the scale up of an ambitious innovation must ask: "Is the candidate innovation well specified?" This question is important because, as Cohen and Ball (1999) argue, ambitious innovations must be well specified if they are to be implemented with any degree of fidelity. Further, a strong answer to this question increases the *testability* and reduces the *complexity* factors in Rogers (2003) framework, thus contributing to a higher potential rate of adoption. As a project progresses to scale, different levels of warrants provide reasonable assurance. Initially, SimCalc benefited by showing its materials to panels of experts. Their reaction was sufficient evidence that the prototype curricular materials were not ready. The panel-of-experts approach to obtaining a scientific warrant continues through our use of advisory boards to review our experimental design and expert panels to examine the construct validity of our assessment. Another level of warrant is obtained by design experiments: If a few handpicked teachers cannot implement an innovation, there is no chance that many teachers will implement it at scale. The replication of design experiments in multiple sites with multiple investigators (as occurred in Phase Three) strengthens the case that the innovation is reasonably well specified. The strongest warrant we pursued to date was in the pilot study (Phase Five): A wide range of teachers was given materials and a measured amount of professional development and then asked to implement a SimCalc replacement unit in the context of an experimental design. Their ability to do so, as measured by daily teaching logs, interviews, their own improvement in mathematics content knowledge, and the differential gains of their students (versus a control group) on an assessment, yields compelling evidence that the innovation is well specified. Hence, we suggest that a range of methods is warranted scientifically for addressing this question.

Question Two: "Is the Candidate Innovation Adaptable to a Wide Variety of Circumstances?"

Researchers studying an innovation must ask: "Is the candidate innovation adaptable to a wide variety of circumstances?" Adaptability is important because, given the variety in American schools, some degree of curricular adaptation almost always occurs (Porter, 2002). Further, an innovation is more likely to reach scale if it fits a variety of both circumstances and curricular scope rather than if it is a "point solution" (fitting only a small topic in restricted settings). Rogers (2003) uses the similar concept of reinvention to highlight the need for adaptability.

In the case study we have presented, empirical evidence for adaptability came primarily from design experiments. The team tried using SimCalc software at many grade levels, with many different student populations, and in many different school settings. In most cases, pretest and posttest measures were used to ascertain that the design resulted in students making the expected learning gains. Pretest and posttest design can provide a fairly strong warrant in some cases, particularly when researchers seek to justify that the design provided students with an "opportunity-to-learn" content that very few students at their grade level have the opportunity to learn. For example, students in the eighth grade rarely learn to relate velocity graphs to position graphs (an aspect of the fundamental theorem of calculus). Further, base rates for success at this task are established for advanced placement calculus and the concept is known to be hard for university students; after all, its initial development required the best minds of western civilization. Thus, pretest and posttest results provide a quite convincing warrant when they show that students in the eighth grade make substantial gains on this kind of task. A control group would allow a stronger warrant, but it would be extraordinarily expensive to explore the full range of adaptation of tools like SimCalc or The Geometer's Sketchpad through formal, randomized, control experimentation. We expect the norm to be that investigators will justify adaptability through case studies and will choose one, fairly constrained use of the innovation to establish cause-and-effect relationships rigorously.

Question Three: "How Do the Effects of the Innovation Vary Within a Variable Environment?"

Researchers must ask: "How do the effects of the innovation vary within a variable environment?" This type of question is important because scale implies lack of control over the environment, which varies significantly from teacher to teacher, school to school, school district to school district, and state to state. In Rogers' (2003) framework, research-based claims about effects can produce strong messages about the *relative advantage* of an innovation. Further, because our specific question involved variable teachers and settings, the research also examines the *compatibility* of the innovation with variations in setting. In our case, a question of this type (i.e., "How do students' gains vary with a wide variety of teachers?") will be addressed using a randomized controlled design and analysis through hierarchical linear modeling. This research method produces a strong warrant by allowing the research to establish a cause-and-effect relationship between the treatment and the outcomes and to model the contributions of various, mediating, environmental variables (in our case, variations in teachers and teaching contexts) on the outcomes. We do not see viable alternative methods to answer this sort of research question.

Diagramming the Tensions

The tensions among these three questions, along with the need for a long-term innovation program to have a vision, can be represented in a spider diagram (Figure 3). In this sort of diagram, progress over time is visualized as increasing area. Increases in area occur as the research grows strong along multiple vectors. We place a far-reaching vision and a model of how effects vary with the environment at opposite poles because a more ambitious vision makes measurement more difficult. Measurement is more difficult because current tests are less likely to probe desired outcomes and because it is harder to get a large population to agree to test an innovation that has a long-term focus. Similarly, adaptability and specification are in tension; it is difficult for an innovation to be both adaptable and tightly specified.

<Insert Figure 3 about here>

The spider diagram shows that SimCalc began with a strong vision but little by way of specification, adaptability, or documented effects—for good reason: The innovations needed to be designed in the first place. Five years into the project, the team was emphasizing adaptability, with a smaller degree of effort directed at documenting effects and refining the specification. The last two phases of the case study suggest a contraction in adaptability (the scope of the project narrowed to replacement units in two grade levels) but dramatic expansion of efforts to address the degree to which the innovation was well specified and to model how the effects varied with the environment. In general, one might expect efforts that emphasize the vision and adaptability dimensions to use warrants from historical and curricular analyses, literature reviews, and design experiments, whereas efforts that emphasize a well-specified innovation and modeling effects would use warrants from advisory boards and statistical analyses of controlled experiments. Some scaling-up innovation efforts, like SimCalc, might start high on innovative vision. Others might begin with great strength in specification and demonstrating experimental effects but with little evidence of adaptability to varying school contexts. On the path to scaling up an innovation, we conjecture that researchers must expand the area along all four vectors so that a variety of means for obtaining a scientific warrant for claims will be required.

CONCLUSIONS

In the present political climate, innovations in mathematics and science education are both strongly needed (to address national goals) and suspect (because of a history of repeated failures to scale up innovations and a conservative mind-set toward educational innovation). Increasingly, the interested parties in our educational system look to research to overcome this tension by identifying innovations that work at scale. Research can live up to this task only with additional attention to methodology, and, in particular, attention to the question: “What should count as a scientific warrant that evidence supports a claim?” Currently, debate among educational researchers and research policy-makers has begun to shore up answers for two kinds of research: small-scale innovation and large-scale evaluation.

By using the case of SimCalc, we have argued for the need to define and defend warrants for a third kind of research: research concerned with getting a serious foundational innovation to scale. The case suggests that research concerned with getting to scale is different. Design research has the goal of finding out "What could work?" by developing theories of how children or teachers learn, how school-based communities can be supported productively, and so on. Initial evidence that an innovation could work is very valuable to scaling up. The kind of theory development particularly needed in scaling-up research is a theory of the active ingredients in an innovation, a theory that provides warrants that link the claimed active ingredients to the observed effects. Likewise, most evaluation research is concerned with programs, practices, or resources that are well specified already. Given the huge number and the complexity of the design decisions that face an innovation team, research must be more cost-effective and agile to support the formative refinement of an innovation. Design research and evaluation research need a well-articulated middle ground. We term this middle ground "scaling-up research."

We do not believe that scaling-up research requires its own new methods, but it does ask different questions, which, in turn, make new demands on and purposefully integrate existing methods. For example, the question of "What could scale up?" combines elements of an existence proof (like design research) with elements of modeling how randomized variability between settings mediates effect sizes (like large-scale experimental research). Further, the process of scaling up an innovation places particularly strong demands on managing the tensions between (a) program vision and what can be measured and (b) adaptability to many uses and detailed specification for one use.

We hope that future dialogue in the research community firmly establishes warrants for the many questions that must be answered in order to undertake a scientific program of scale-up research, something that a single case study cannot possibly do. In particular, it would be useful to have agreed-upon warrants for using evidence to claim that:

- An innovation, when fully realized, could result in large-scale benefits to the public.
- An innovation is sufficiently well specified to be tested at scale.
- An innovation is adaptable to and attractive for a sufficiently wide range of uses to be worthy of scaling up.
- The experimental model for scaling up the innovation and measuring the results is valid.

We suspect that these warrants will need to relate claims to evidence gathered through a mix of design, expert panel, and experimental methods.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation, most specifically under Grant No. 0228515. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect necessarily the views of the National Science Foundation.

REFERENCES

Becker, H. J. (1998). The influence of computer and internet-use on teachers' pedagogical practices and perceptions.

Becker, H. J., & Anderson, R. E. (1998). *Teacher's survey: Combined version 1-4*: Center for Research on Information Technology and Organizations, University of California, Irvine.

Blumenfeld, P., Fishman, B., Krajcik, J., & Marx, R. W. (2000). Creating useable innovations in systemic reform: Scaling up technology-embedded project-based science in urban schools. *Educational Psychologist, 35*(3), 149-164.

Bransford, J., Brophy, S., & Williams, S. (2000). When computer technologies meet the learning sciences: Issues and opportunities. *Journal of Applied Developmental Psychology, 21*(1), 59-84.

Clarke, R. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445-459.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2002). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

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

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement* (No. CPRE Research Report No. RR-043). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

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.

Cook, T. D. (2000). Reappraising the arguments against randomized experiments in education: An analysis of the culture of evaluation in american schools of education. Retrieved August, 31, 2003, from http://www.sri.com/policy/designkt/cokfinal.doc

Cuban, L. (2003). *Oversold and underused: Computers in the classroom.* Cambridge, MA: Harvard University Press.

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

Edelson, D. C. (2002). What we learn when we engage in design. *Journal of the Learning Sciences, 11*(1), 105-121.

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

Ferrio, T. (2004). What year did the graphing calculator get to scale? (email correspondance).

Fullan, M. (2000). The return of large scale reform.

Goldenberg, P. (1995). Multiple representations: A vehicle for understanding understandings. In D. N. Perkins, J. L. Schwartz, M. M. West & M. S. Wiske

(Eds.), *Software goes to school* (pp. 155-171). New York, NY: Oxford University Press.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Physics Teacher, 30*, 141-158.

Jackiw, N. (1988-97). The geometer's sketchpad (Version various versions). Berkeley, CA: Key Curriculum Press.

Kaput, J. (1987a). Representation and mathematics. In C. Janvier (Ed.), *Problems of representation in the learning of mathematics* (pp. 19–26). Hillsdale, NJ: Erlbaum.

Kaput, J. (1987b). Toward a theory of mathematical symbol use. In C. Janvier (Ed.), *Problems of representation in the learning of mathematics* (pp. 159–196). Hillsdale, NJ: Erlbaum.

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. (2001). Changing representational infrastructures changes most everything: The case of simcalc algebra, and calculus., *NAS Symposium on "Improving Learning with Informational Technology.* Washington, D.C.

Kaput, J. (in press). Technology as a transformative force in math education: Transforming notations, curriculum structures, content and technologies. In E. Galinde (Ed.), *Technology and the nctm standards*. Reston, VA: National Council of Teachers of Mathematics.

Kaput, J., Noss, R., & Hoyles, C. (2001). Developing new notations for a learnable mathematics in the computational era. In L. D. English (Ed.), *The handbook of international research in mathematics.* London: Kluwer.

Kaput, J., Noss, R., & Hoyles, C. (2002). Developing new notations for a learnable mathematics in the computational era. In L. D. English (Ed.), *Handbook of international research on mathematics education* (pp. 51-75). Mahwah: Lawrence Earlbaum 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, UK: Falmer Press.

Kozma, R. B., Russell, J., Jones, T., Marx, N., & Davis, J. (1996). The use of multiple, linked representations to facilitate science understanding. In S. Vosniadou, E. De Corte, R. Glaser & H. Mandl (Eds.), *International perspectives on the design of technology-supported learning environments* (pp. 41-60). Mahwah, New Jersey: Lawrence Erlbaum Associates.

National Center for Education Statistics. (2001). *The nation's report card: Mathematics 2000.* (No. NCES 2001-571). Washington DC: U.S. Department of Education.

National Science Foundation. (2004). *Interagency education research initiative (ieri): Program solicitation* (No. NSF 04-553). Washington DC: National Science Foundation.

Nemirovsky, R., Kaput, J., & Roschelle, J. (1998). *Enlarging mathematical activity from modeling phenomena to generating phenomena*. Paper presented at the Proceedings of the 22nd Psychology of Mathematics Education Conference, Stellenbosch, South Africa.

Nickerson, S. D., Nydam, C., & Bowers, J. S. (2001). Linking algebraic concepts and contexts: Every picture tells a story. *Mathematics Teaching in the Middle School, 6*(2), 92-98.

No child left behind act, United States Congress(2001).

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

President's Committee of Advisors on Science and Technology. (1997). *Report to the president on the use of technology to strengthen k-12 education in the united states*. Washington, DC: President's Committee of Advisors on Science and Technology (PCAST).

Roschelle, J., & Kaput, J. (1996). Simcalc mathworlds for the mathematics of change. *Communications of the ACM, 39*(8), 97-99.

Roschelle, J., Kaput, J., & Stroup, W. (2000a). Simcalc: Accelerating student engagement with the mathematics of change. In M. J. Jacobsen & R. B. Kozma (Eds.), *Learning the sciences of the 21st century: Research, design, and implementing advanced technology learning environments*. (pp. 47-75). Hillsdale, NJ: Erlbaum.

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

Scher, D. (2000). Lifting the curtain: The evolution of the geometer's sketchpad. *The Mathematics Educator, 10*(2), 42-48.

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.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 21*(1), 13-25.

Schorr, R. Y., Firestone, W., & Monfils, L. A. (2003). State testing and mathematics teaching in new jersey: The effects of a test without other supports. *Journal for Research in Mathematics Education, 34*(5), 373-405.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington DC: National Academies Press.

Shneiderman, B. (1982). The future of interactive systems and the emergence of direct manipulation. *Behaviour and Information Technology, 1*(3), 237-256.

Smith, D. C., Irby, C., Kimball, R., & Verplank, B. (1982, April). Designing the star user interface. *Byte,* 242-282.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3*(2), 115-163.

Stroup, W. M. (in press). Understanding qualitative calculus: A structural synthesis of learning research. *International Journal of Computers for Mathematical Learning*.

Torgerson, C. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies, 49*(3), 316-328.

Table 1

SimCalc's Approach To Uniting Three Perspectives

SimCalc's Approach	Three Perspectives		
	Learners' Strengths	**Representational Technology**	**Reorganized Curriculum**
Foregrounding the relation of mathematical representations to phenomenological motion	Perceiving, describing, and reasoning about motion in the concrete.	Linked animation and mathematical notations make tangible the connection between formalism and common sense.	Emphasizes phenomena as a tool for building understanding.
Formalisms are introduced to help consolidate and extend knowledge established previously	Natural progression from case-based, specific learning to more integrative, general understanding.	Unites and reifies mathematical formalism and more intuitive expressive notations in one medium.	Provides substantial informal learning opportunities before introducing formal concepts.
Piecewise functions	Reasoning about intervals can leverage arithmetic and simple geometric skills to compute quantitative aspects.	Can represent visually and allow purposeful manipulation of piecewise-defined motions.	Uses piecewise functions as an essential building block for all calculus concepts.
Emphasizing reasoning across rate (velocity) and accumulation (position) descriptions	Builds on the ability to think about the "same" object in "different" views.	Dynamic links among different representational views of the mathematical object.	Focuses on rate–accumulation relationships expressed qualitatively and arithmetically.
Primary focus on graph-based and linguistic reasoning	Making sense of and guiding action in graphic,	Supports visual presentation and direct editing of graphic	Shifts the emphasis from symbol manipulation to

	visual forms.	forms in newly expressive ways.	more democratically accessible forms of expression.
Inquiry cycle of plan, construct, experience, reflect	Ability to understand a challenge, hypothesize possible solutions, and distinguish success from failure.	Computational apparatus allows for many, quick, iterative, feedback cycles.	A more playful, expressive, microworlds approach to mathematics.

Figure Captions

*Figure 1***.** Three perspectives of the restructuring knowing framework.

Figure 2. Framework for classroom learning. Adapted from Cohen et al., 2003.

Figure 3. Spider diagram of SimCalc's coverage of four concerns.

Figure 1.

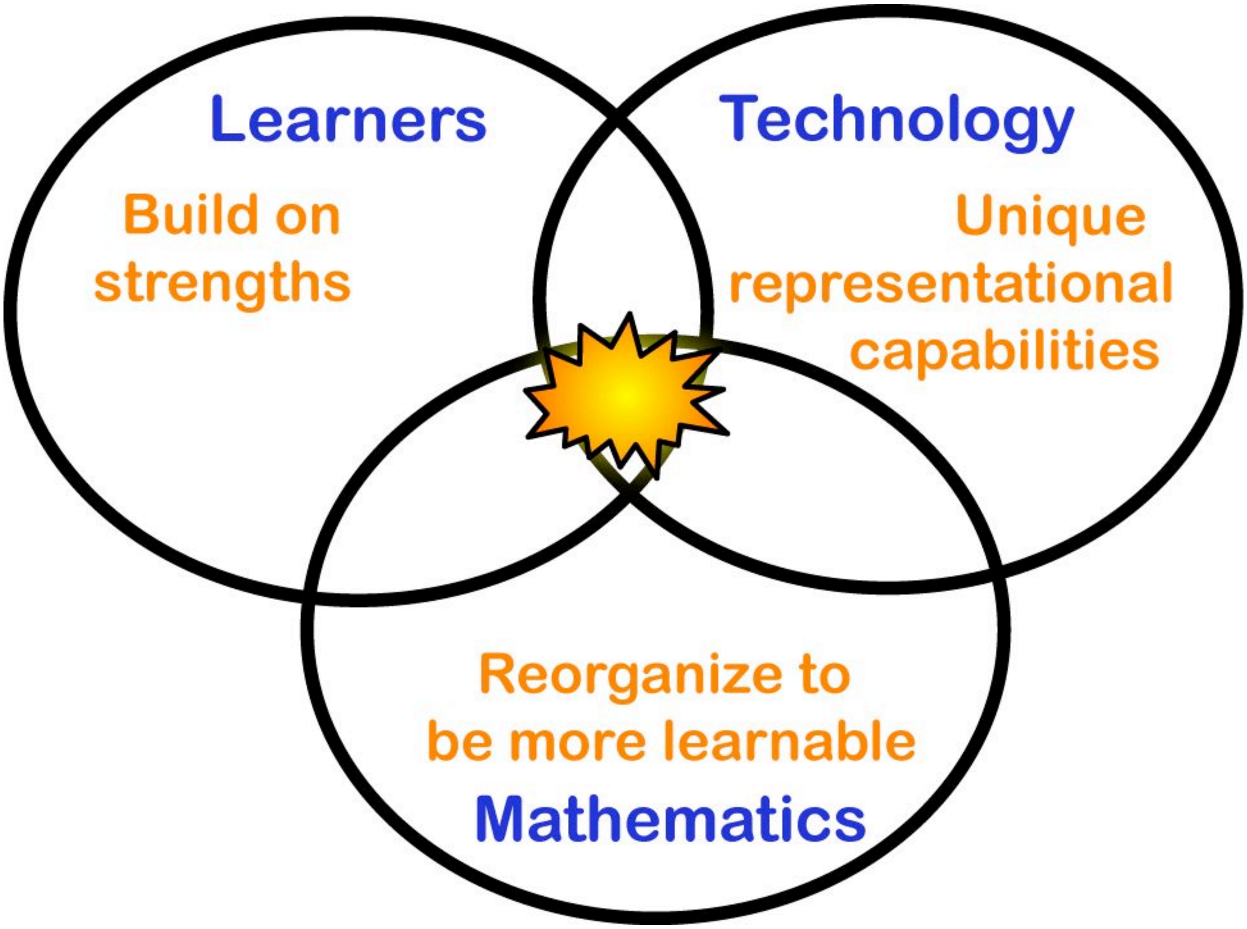

Learners
Build on strengths
Technology
Unique representational capabilities
Reorganize to be more learnable
Mathematics

Figure 2.

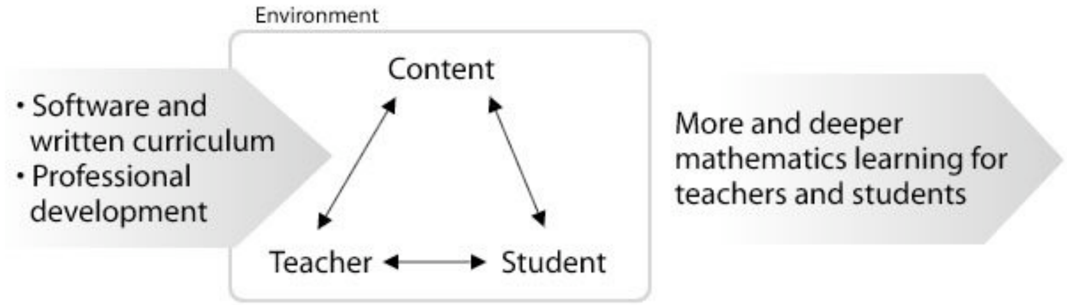
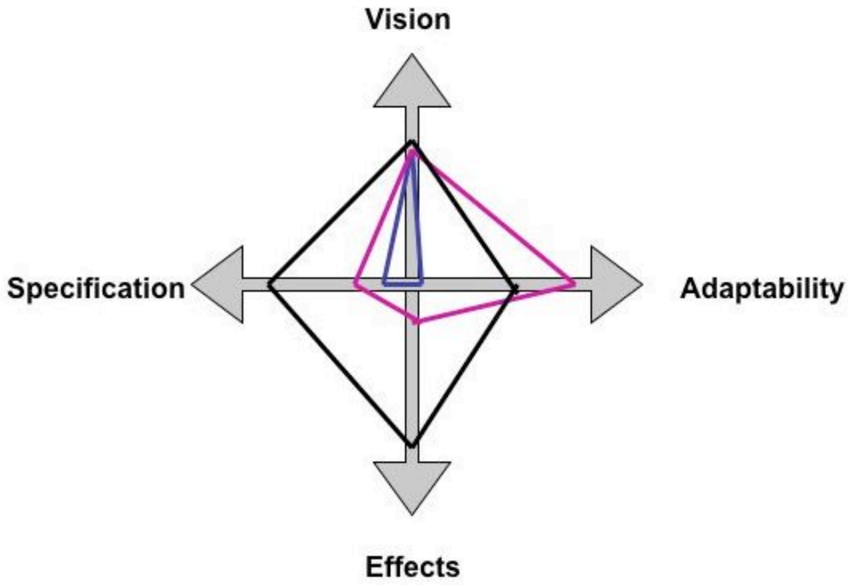

Environment
• Software and written curriculum
• Professional development
Content
Teacher
Student
More and deeper mathematics learning for teachers and students

Figure 3.


Vision
Specification
Adaptability
Effects