Educ Stud Math (2008) 68:149–170
DOI 10.1007/s10649-008-9119-3

The role of scaling up research in designing for and evaluating robustness

J. Roschelle · D. Tatar · N. Shechtman · J. Knudsen

Published online: 27 March 2008

Abstract One of the great strengths of Jim Kaput's research program was his relentless drive towards scaling up his innovative approach to teaching the mathematics of change and variation. The SimCalc mission, "democratizing access to the mathematics of change," was enacted by deliberate efforts to reach an increasing number of teachers and students each year. Further, Kaput asked: What can we learn from research at the next level of scale (e.g., beyond a few classrooms at a time) that we cannot learn from other sources? In this article, we develop an argument that scaling up research can contribute important new knowledge by focusing researchers' attention on the robustness of an innovation when used by varied students, teachers, classrooms, schools, and regions. The concept of robustness requires additional discipline both in the design process and in the conduct of valid research. By examining a progression of three studies in the Scaling Up SimCalc program, we articulate how scaling up research can contribute to designing for and evaluating robustness.

Keywords Democratization of access to mathematics · Educational technology · Mathematics education · Randomized experiments · Scaling up

1 Introduction

Mathematics education research has an uncertain relationship to the issue of scaling up innovations to widespread use. Researchers design innovations to improve mathematics teaching and learning; these innovations can include new teaching practices, new

J. Roschelle (✉) · N. Shechtman · J. Knudsen
SRI International, Center for Technology in Learning, 333 Ravenswood Ave.,
Menlo Park, CA 94025, USA
e-mail: jeremy.roschelle@sri.com

N. Shechtman
e-mail: nicole.shechtman@sri.com

D. Tatar
Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

curriculum materials and new applications of technology. The design of innovations can be in the service of developing or refining theory (Cobb et al. 2003). We suspect that many researchers want to go beyond innovation as theory-building; researchers want to see their innovations in use at a scale beyond the few settings in which they conducted their initial research (Fullan and Earl 2002). Yet, few mathematics education researchers participate in research about the challenges, demands, and outcomes of implementing innovations at scale (e.g., Elmore 1996), and thus are not focused on designing explicitly with these factors in mind. For example, an international review by Lagrange et al. (2003) found that only 5% of the papers on the selected topic of "computer algebra systems" addressed the issue of integrating Information and Communication Technologies (ICT) into everyday school conditions.

Kaput's scaling up challenge James J. Kaput was an exemplary researcher in this regard. His vision in the SimCalc research program, described in more detail below, encompassed both theory-building (Kaput and Roschelle 1998; Kaput and Shaffer 2002) and large-scale impact (Roschelle and Kaput 1996). Most researchers move on to the next problem but stay at the same level of analysis. In contrast, Kaput stayed with the central problem, deepening his understanding of its ramifications. He moved from analysis of a few students to analysis of a few teachers to analysis in multiple schools in multiple states. He progressed from microanalyses (Roschelle et al. 2000) to large classroom-based studies (Tatar et al. 2008). At the same time, he strategically added colleagues to the SimCalc circle who embodied and integrated the multidisciplinary perspective required for each new challenge. He pursued engagement with school districts, policy makers, and companies who could influence mathematics education. Finally, he worked diligently to move the core technology out of computer labs and into student hands by porting the software from expensive desktop computers to inexpensive and widely available graphing calculators (Kaput 2000) and to structure the software with an open architecture so that others might add their ideas to it (Roschelle and Kaput 1996). He did not treat scaling up as mere "dissemination" (usually a code word for publications and presentations in the researcher's preferred venues) but as a mode of action in the world involving moving an innovation out of the mathematics education research community and into large scale use (Roschelle and Jackiw 2000). Scaling up was a core component of the research program (Roschelle et al. 1998; Roschelle et al. 2008).

Despite his strong drive towards scaling up, Kaput never provided a definition of "Scaling Up Research." He did, however, engage in lengthy discussions with one of the authors of this paper (Roschelle). A theme of those discussions can be summarized in a question about research with more than a few classrooms at a time: What can we learn from research at the next level of scale that we cannot learn from other sources?

This question has been at the heart of our research program for at least 5 years. It has not been an easy question to answer. The question does not allow one to justify scaling up research merely in terms of its most obvious feature, a larger population. If we go to a larger population but learn nothing new, the question is unanswered. Indeed, other researchers have argued against defining scaling up merely in terms of n (Coburn 2003). It is also tempting to define scaling up research in terms of methodological features of the research design, e.g., random assignment of teachers to condition (Reichardt 2007). Kaput was quick to point to his prior experience with psychologists in which they could design an experiment to reproduce an effect he already knew about; if nothing new was learned, he felt this was a trivial and wasteful exercise. Another possible definition of scaling up research is in terms of audience, that is, the claim that scaling up research supplies the

information that policy-makers want (Schneider and McDonald 2007). Kaput also had ample experience with policy makers—locally, at the state level, and in the United States Senate. He'd say, "And do you know how many times they asked me for a randomized controlled experiment? None!" Kaput's question focused our attention away from simple features of the design or the audience and towards the view that scaling up research must measure its worth in terms of the uniqueness of its contributions to knowledge.

Instrumental vs. institutional relationships view of scaling up A specifically mathematical analysis of questions of scaling up has distinguished two broad views in the research literature of ICT integration in mathematics education (Lagrange et al. 2003). An "instrumental" view focuses on a technology itself and what students learn from use of that technology. In a complex process called "instrumental genesis," student learning is mediated by how the student cognitively generates and uses the affordances of the tool (Assude and Gelis 2003). In contrast, an "institutional" view focuses on the influences of institutional context on how students use and learn with a technology. We see both views as necessary and complementary. For example, in the "integrating research teams" approach being undertaken in Europe, research teams in different countries cross-experiment with each others' technologies (Cerrulli et al. 2007). Early results from this work suggest that a team's theoretical frame neither specifies the process and intended outcomes of learning with a given tool adequately (a challenge under the instrumental view) or make suitably clear and explicit descriptions of classroom experiments in different institutional settings (a challenge under the institutional view). By addressing these challenges via an expanded research methodology, researchers can make progress on issues that are core to the eventual integration of new innovations into large-scale, everyday use.

In this article, we focus on a potentially unique contribution with a new emphasis on experimental research, a style of scaling up research that is gaining attention in the United States. Our experimental research involves large scale experiments in which many teachers are assigned at random to either use a new combination of ICT and curriculum or to continue with their business-as-usual approach. In this research, we answer Kaput's question with respect to both instrumental and institutional views. We examine the instrumental genesis that occurs for various different types of students as they take up the new technology and curriculum, employing the tools for deep conceptual learning. We also examine technology use across different settings, such as classrooms, schools, districts, and regions.

Modeling variation and the concept of robustness This research investigates institutional processes by engaging in statistical modeling of the variation in the data. Attributions of variation can help us understand the prospects of the innovation. In particular, innovations that show a lot of variation in implementation when implemented in different settings may require different strategies for scaling. Even within a single Scaling Up study, we can begin to characterize how an innovation holds up as it spreads beyond the original trial classrooms (Baker 2007). In general, by starting where more traditional research ends—with accounts of the impact of the new approach in a small number of trial classrooms—scaling up research allows us to distinguish more from less *robust* innovations and examine sources of variation between implementations of an innovation in different settings (Hedges 2007). By robustness, we mean the consistency of the innovation's benefits for student learning when deployed consistently to a wide variety of students, teachers, and settings. This focus is the answer to Kaput's challenge. The assessment of robustness is one unique key contribution of scaling up research.

Ultimately, we will engage in hierarchical linear modeling (HLM). HLM (Raudenbush and Bryk 2002) is a method of statistical modeling that improves on regression analysis by analyzing and attributing variance to measures in the data at different levels of nesting. In particular, this approach can take into account the variation of student level variables, such as learning gains, nested within classrooms nested within schools nested within districts, and so forth. Although subject to the same (and additional) threats to validity as any other attribution of variance, including vulnerability to the reliability of the underlying measures, HLM allows us to make more detailed and accurate attributions than previously available.

Implications outside the United States Although we readily admit that this style of research carries with it many idiosyncrasies that may be seen as valuable only in United States' political context, we believe it is valuable for a wider audience to consider what this style of scaling up research may be able to contribute to mathematics education internationally. On the one hand, we note that scaling up directly relates to the "institutional" view cited in international mathematics education research; robustness is a property of an innovation across institutional settings. On the other hand, we argue that the "instrumental" view need not be lost in this style of study (although it may be difficult to carry out to the same level of refinement as in other study designs). Under the instrumental view, different integrations of a technology into teaching and learning may produce different learning results for students. For example, findings of the TIMSS video studies (Hiebert et al. 2003) show that teachers in different countries enact the same curricular tasks in very different ways, often significantly modulating the aspects of mathematics (e.g., more procedural or conceptual) that students learn. Given an integration of curriculum and technology that aims for a targeted learning outcome, it becomes interesting to investigate the robustness of instrumental genesis towards that outcome. Do teachers (as they do in the TIMSS video studies) significantly modulate this process, changing what their students learn?

The importance of robustness While noting the relationship of this emerging style of research to previous mathematics education instrumental and institutional views, we also wish to avoid merely reducing experimental research to existing views. "Robustness" that can be designed for and evaluated through experimental research is a valuable concept in its own right. We need to be able to answer questions like these: How do we design integrated ICT systems that can have impact across various subgroups of students, teachers, or schools? Does the innovation actually have impact across various subgroups of students, teachers, or schools? Of the many ways in which teachers vary (e.g., their philosophy, expectations, experience with technology, practices, etc.), which have the most consequence for student learning with our integrated approach? Are there populations of students (e.g., rich or poor, girls or boys, Hispanic or white) for which our integration has undesirable or less desirable effects? Are there school conditions (e.g., availability of computers in the classroom or in a lab) that mediate student learning outcomes, and if so, in what direction?

These benefits can be characterized not only in terms of what is learned but also in terms of the sources of variability in how much is learned. Thinking about sources of variability for differences in ICT integration is important because teachers, students and schools are highly variable and this variability can dilute or distort the mechanism (and hence the benefits) of innovations. As Dede (2006) points out, one approach to this variability has been to wrap each innovation in increasingly complex systemic reforms. These reforms try to control more and more of the context of teaching and learning so as to make the environment suitable for the core innovation. (There is an unfortunate irony in

constructivist educators deciding that they need to control many complex aspects of an educational system so that their situated and local innovation might work.) The alternative is to design innovations to be more robust, which can best be done if we understand why results vary.

The overall purpose of this article is to make a case that experimental scaling up research can contribute to mathematics education by providing evidence of the robustness of the innovations developed by researchers in mathematics education. After a short introduction to the SimCalc approach, we make our argument that experimental scaling up research can contribute to mathematics education. We do this by examining what we learned in three phases of research with SimCalc. The first phase involved design experiments and case studies in different locations. Although useful things about scaling up can be learned through comparative case studies, we argue that ultimately each of these "design experiments" has too much freedom to change "the innovation"—complicating the process of making a valid claim about robustness. In the second phase, we carried out an experiment with over 20 teachers. We argue that this level of experiment was sufficient to reveal the degree to which, what, and how much students learn is modulated, but not to understand the sources of variability. Without being able to quantify sources of variability, it is unclear how to improve the student learning experience. For example, if most of the variability is at the student-level, we might improve the learning experience by customizing it for different student populations (e.g., ethnic groups, ability levels). On the other hand, if most of the variability is at the teacher-level, we might improve the learning experience through teacher professional development. If so, what should the professional development focus on? More comfort with technology? Deeper mathematical understanding? A shift to a constructivist pedagogy? We argue that a well-designed program of experimental scaling up research can allow researchers to make valid claims about the answers to such questions. Such a program structures a design process that emphasizes robustness at all levels of variation and enables measurement of the sensitivity of the innovation to variation in implementation conditions. The measurement process in scaling up research seeks (a) to measure the overall effectiveness of an approach in varied settings, (b) to measure similar or differential impacts across subpopulations, and (c) to model the relationships between variability in implementation conditions and student outcomes. We complete our argument that experimental scaling up research can help mathematics educators understand the important concept of "robustness" by examining the kinds of robustness claims we were able to make in our large-scale experiment with 95 teachers.

2 The heart of the scaling up SimCalc program

It is always tempting to begin describing the heart of the Scaling Up SimCalc program by describing software. We have become concerned that this rhetorical approach leads to a pervasive misconception. We use "misconception" in the technical sense of a belief that is hard to dislodge (Smith et al. 1993)—readers tend to see the term "SimCalc" and think "software." In contrast, our research examines the *integration* of three elements—professional development, curriculum and software. There is an obvious reason to look at *integration* in scaling up research: research has repeatedly found that merely injecting software into classrooms does nothing of particular value; researchers repeatedly call for alignment and integration of a combination of program elements, usually including curriculum and teacher professional development (National Research Council 2001). Thus in the remainder of this article, we will use "SimCalc" to refer to an integrated approach to

teacher training, paper curriculum and software. When we wish to refer to software alone, we will refer to the software by its name, "SimCalc MathWorlds®."

We are, of course, not working with an arbitrary integration but rather one that specifically follows the mission and vision that Kaput laid out. The SimCalc mission is to democratize access to the mathematics of change and variation starting in middle school. The vision is to utilize the representational qualities of technology to do this (Kaput 1992). As other researchers have noted, one explanation for the disappointing status of technology use in schools is the overall lack of vision and clarity of goals with regard to technology's role (O'Neil 1995). SimCalc, in contrast, started from a clear mission and vision of technology's role. That mission and vision imply scale and require robustness, especially to include disadvantaged populations (Kaput 1994, 1997).

Throughout the history of SimCalc, Kaput articulated his approach in a set of slides that slowly evolved but never strayed from a few key messages. In reading the points below, notice how Kaput focused on teaching important mathematics with software in a infrastructural role. Hallmarks of the SimCalc approach to the mathematics of change and variation are:

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

We see Kaput's work as embodying three messages through this approach. First, Kaput *demystified* the mathematics of change and variation (Kaput 1994). In the United States, this content is usually taught in an end-of-high school Calculus course to only an elite population. Consequently, many citizens find Calculus to be mysterious and unapproachable. Kaput argued forcefully that this view is both destructive and unnecessary (Kaput 1994). It is destructive because of the many everyday situations in which citizens needed to reason more carefully about change as well as the general loss to society when people fail to understand deep mathematics. It is unnecessary because it is possible to have a curricular strand across many grade levels that results in deep learning; he conceptualized a learning progression for the mathematics of change and variation that would begin in the primary grades and continue through university education. He invoked historical analogies to other literacies, showing that in other ages knowledge that had been held to be elite and mysterious was eventually democratized and made available to all citizens. And he highlighted evidence showing that ordinary children could achieve extraordinary depth of understanding of the mathematics of change and variation with the right resources.

Second, Kaput articulated not just the vague notion that representations are important, but an entire *representational view* of how to transform teaching and learning of the mathematics of change and variation (Kaput 1992; Kaput and Roschelle 1998). Along with many in the ESM community, he saw deep links between representation and epistemology; inclusion of new representational media could enable new avenues to the foundational roots of the mathematics

of change and variation, for example, the Fundamental Theorem of Calculus. When thinking about representational infrastructure, Kaput would often first start with what he called the "Big Three"—algebra, tables, and graphs—and emphasize the advantages of making graphs more prominent. An overly simplified version of his argument for graphs is "children are more insightful about big ideas in the mathematics of change and variation when reasoning with graphs." Next Kaput would insist that these three representations are insufficient and argue for putting motion phenomena at the center. He would argue that the representations need to be *about* something and that, historically, the mathematics of change and variation is about motion. When conceptualizing motion representations, Kaput included both kinesthetic and visual experiences—students should re-experience both the movement of their own bodies and the motion of things they can see. Finally, Kaput strongly pushed for considering narrative stories to be a primary representation, where the stories developed to precisely describe motions in terms that enable mathematical analysis.

Third, Kaput *integrated* curriculum and technology. He was always simultaneously working to improve both. A favorite slogan was "new technology without new curriculum is not worth the silicon it's written in." When talking about scaling beyond his own design experiments, Kaput was quick to include teacher professional development into his integration and, in fact, offered workshops to teachers throughout his region and nationally. Per the paragraph above, Kaput viewed technology as useful for its representational features (and later, its connectivity features, but these have not yet been part of our scaling up research program). SimCalc MathWorlds® software capitalizes on new technological capabilities to enable students to learn complex mathematical concepts through multiple, dynamic representations (Kaput 1992; Roschelle et al. 2000). In particular, SimCalc MathWorlds® software enables students to visualize the concepts behind change and motion by connecting and contrasting narrative, symbolic, graphical, and table-based representations and by linking graphs to simulated motions.

To summarize, the *innovation* in Kaput's approach was the *demystification* of the mathematics of change and variation, the elaboration of *representational* resources that could strengthen teaching and learning, and the *integration* of curriculum, technology, and teacher professional development in the service of transforming school learning.

Preparing SimCalc for rigorous scale-up research was a multi-phase process spanning nearly a decade (Roschelle et al. 2008). Because the effectiveness of any classroom intervention is mediated by teaching practices and contextual conditions, it was important to determine that the complex underlying concept was strong and adaptable to a wide variety of teachers and settings. Thus, early stages of research focused on replicated classroom design experiments, each with varied curriculum and professional development components. The design research was conducted primarily in disadvantaged urban schools, in order to support the idea that the innovation could be successful anywhere, even without vast resources and "boutique" settings. Knowing that innovations too distant from the reality of existing curricula and teacher practices will not advance to widespread adoption, the research also examined the compatibility of SimCalc materials with the existing curriculum. The design research yielded positive pre-test/post-test results for students in many different settings, thus providing evidence that the core concept of SimCalc could potentially succeed at scale.

3 Scaling up research and the diffusion of innovation tradition

In this section, we review a powerful research tradition, diffusion of innovation, that can inform the choices we make in envisioning successful scaling up. Theory about diffusion of

innovation helps us understand how SimCalc is positioned, where its strengths are, and where its weaknesses are. Additionally, by thinking in these terms, we have arrived the notion of robustness, introduced above as a central contribution of scaling up educational research.

In summarizing the diffusion of innovation tradition, Rogers (2003) proposes that five factors influence rate of adoption: observability, trialability, compatibility, complexity, and relative advantage. Rogers further highlights the importance of *re-invention*, a capability that increases adoption by allowing users to make an innovation fit their local needs. He also stresses the importance of forging an alliance between *change agents* and *opinion leaders* in order to facilitate the transition of an innovation into the realm of practice.

SimCalc MathWorlds® software has always been strong in *observability*: after a short demonstration of the software, most teachers express the feeling that their students could benefit from the graphic and animated representations that SimCalc provides. (We note however, that when using SimCalc MathWorlds® teachers do not always observe how their students are learning by interacting with software; some teachers subsequently decide to use the software only in demonstration mode, which is a "lethal mutation" of Kaput's intent.) Furthermore, SimCalc was designed to be highly adaptable by users, a quality that aligns with Rogers' notion of *re-invention* and bodes well for scalability.

SimCalc research has been organized throughout with *relative advantage* in mind. Throughout the SimCalc development trajectory, the team accumulated evidence about the relative advantage of the innovation. First, the team was able to show ordinary students learning more complex mathematics; the team was also able to articulate the potential advantage of using new representational capabilities to draw upon learner's strengths and to re-organize the curricular content to be more learnable. Shifting to controlled design experiments with carefully defined outcome measures, the team was able to show a causal relationship between SimCalc and enhanced student learning.

As we approached the Scaling Up study, *trialability* posed a challenge. However, by defining the initial unit of adoption in the Scaling Up study as a replacement unit with a very clear scope, situated at a specific place within the curriculum, we substantially increased the trialability of the intervention and facilitated the recruitment of teachers to experiment with the software compared to more diffuse presentations of its possibilities.

Additionally, Kaput, an external *change agent* who has sought to influence the use of innovations in local schools, built an alliance with Bill Hopkins at the University of Texas at Austin, and the Texas Educational Service Centers (ESCs)[1]. Hopkins, the University of Texas and the ESCs have been *opinion leaders* in the adoption of innovations in mathematics education in Texas. This alliance set the stage for a study of a large-scale implementation of the SimCalc innovation.

The last of Rogers' factors, compatibility and complexity, became major foci for the Scaling Up research. Because SimCalc focuses on a topic, the mathematics of change and variation, that is important but not focal within current US mathematics standards, SimCalc is weak in compatibility. Additionally, SimCalc has traditionally been somewhat difficult for teachers in terms of complexity because it requires extensive use of technology and a new approach to complex mathematics concepts. The diffusion of innovation tradition drew our attention to the need to connect SimCalc more closely with the mathematics topics, standards, and curriculum considered important by most mathematics educators and to make the curricular materials and training workshops as clear and simple as possible.

[1] Educational Service Centers are public regional organizations that offer educational support programs, ranging from financial or personnel support to innovative professional and curriculum, to districts throughout the state of Texas.

Although Rogers' work is influential in our thinking, in the end, we characterize our questions about the adoptability and impact of SimCalc in terms of a factor not investigated by Rogers: robustness. The broad diffusion literature is based upon settings that are considerably less variable than American schools, especially American agriculture. In agriculture, the most important variations in implementation conditions can be easily predicted and reproduced under the control of university researchers, without the participation of ordinary farmers. Thus the robustness of the benefits of a new fertilizer can be established without involving everyday farmers and then the new fertilizer can be diffused to everyday farmers. In education, we cannot establish robustness without involving ordinary teachers and students—we do not know enough about which variations in implementation conditions are most important or how to produce them without involving ordinary teachers and students. Hence, scaling up research in education cannot simply be diffusion of innovation research; it must also allow us to understand and quantify robustness of an innovation through research with teachers and students. In the diffusion of innovation research tradition, robustness of the candidate innovation is a given; in educational scaling up research it is not.

We will now turn to our sequence of three research phases in the scaling up SimCalc program to show how valid research on robustness of innovations in mathematics education can take the form of a large-scale field experiment. In doing so, we have two purposes. The first is of course to design a study that can investigate robustness. However, the second is that, in so doing, we have the opportunity to make the innovation more robust.

4 Research phase 1: distributed field trials

Roschelle et al. (2008) provide the primary discussion of our experience with distributed field trials. Here we recap the highlights of that phase of research with an eye to what we can learn about robustness from them.

Within the overarching SimCalc Project, our first distributed design experiments served as precursor work to designing for robustness. The very earliest work with SimCalc's integration of professional development, curriculum, and software occurred near Kaput's University of Massachusetts, Dartmouth location. In this phase, investigators in Newark, NJ (Roberta Schorr); Syracuse, NY (Helen Doerr) and San Diego, CA (Janet Bowers and Susan Nickerson) engaged in their own design experiments using the SimCalc MathWorlds® software, aligned with the SimCalc vision and mission. Each investigator produced their own form of professional development and their own curriculum to accompany the software. Each targeted different grade levels and student populations. Some worked after-class, and some worked during class. Some focused on pre-service teachers; others focused on in-service teachers. One important thing we learned from these design experiments was the breadth of the curricular goals for which SimCalc MathWorlds® could be useful. For example, a potentially powerful approach to periodic functions emerged in which the properties of periodic functions under integration and differentiation are explored by examining simpler piecewise linear approximations to a continuously varying periodic function. Likewise, we learned that the software could be applied to enable students to make sense of simultaneous linear equations. Understanding the scope of applicability can be a useful goal in scaling up representational software across settings.

The strategy of distributed design experiments has a precedent in the work of Hawkins (1997). As Hawkins argued, the strategy has the advantage of rapidly probing a fairly wide sample of the possibility space for a visionary approach. Much mathematics innovation is

created in the context of design experimentation (Cohen et al. 2003). In a design experiment, the researcher can characterize the exciting new content from evidence that emerges in the classroom or in a post hoc analysis. For example, the researcher can focus on the critical moment as manifested in the behavior of a small number of students—without previously anticipating the critical moment or the mathematics within it. Thus, even though the innovation may ostensibly be addressed to teaching the mathematics of change and variation, an incident which reveals that a given student had an insight about fractions can become part of the mix of potential seen in the innovation. The potential for appropriation for different learner needs is arguably an important aspect of an innovation and may lead to a change of emphasis for the innovation or the associated pedagogy (Brown 1991). Indeed, this was the case. This work established the plausibility of Kaput's vision of a set of curricular modules that together make up a learning progression for the mathematics of change that stretch from middle school through university education.

While we accumulated curricular innovations and empirical evidence of the effectiveness of these innovations across various groups, more work would be required to reach valid conclusions about the robustness of the SimCalc integration of teacher professional development, curriculum, and software. Every investigator chose different learning outcomes and produced their own unique integration. This makes it very difficult to make claims about the integration. Interpretations could be made to an overly general level (that any integration is robust) or an overly specific level (that use of the software is what matters). We did not believe either of these. Kaput also did not think just any curricular integration or any professional development would do. Nor did he think the only necessary common element was the SimCalc MathWorlds® software. More generally, although such comparative case studies are crucial in informing our view of instrumental genesis and institutional integration, by themselves, they are too weak to lead to valid conclusions about the robustness of research-based approaches to mathematics education.

5 Research phase 2: pilot experiment

Tatar et al. (2008) provide the primary discussion of our first scaling up experiment. Here we recap the highlights of that study with an eye to what we can learn about robustness from it. At the time we began this work, very strong arguments were being made in the United States policy context about validity. Cook (1999) and others (e.g., Torgerson 2001) were arguing that only randomized experiments were valid for causal inferences about the effect of an intervention. We had not yet articulated our focus on robustness, but did have the above-mentioned concerns about validity of design experiments. We decided that our intervention was mature enough to test its causal influence on student learning and decided to embark on the journey of conducting a randomized experiment.

In this second phase, we began with an experiment involving a small sample of teachers. While this experiment served a strong purpose in its own right, we also conducted it with an eye toward developing the materials, instruments, implementation infrastructure, and initial findings necessary to eventually take the study to large scale. Thus, it was a pilot study.

This phase was an important turning point, because shifting to a question of scale forced the research design process to focus specifically on creating a unitary robust intervention. In order to do an experiment, we had to have a reasonably uniform integration of teacher professional development, curriculum, and software. Our testable integration, further, had to adequately represent Kaput's vision with enough specificity to demonstrate appeal to local decision makers and to be easily usable by a wide variety of teachers. Along with

Cerulli (Cerulli et al. 2007), we found the drive to scale to be a valuable forcing function for making the theory and practical details of our program more explicit.

Interestingly, this shift to an experimental perspective drove specifically *mathematical* aspects of our research. The goal of many mathematics education innovations is to enable teachers to provide opportunities for students to learn more mathematics. But if the teachers are unclear about additional concepts or skills they are to cover, it is unlikely they will be effective in helping students learn those additional concepts or skills. Again, consider SimCalc. Our mission is to "democratize access to mathematics of change and variation." What exactly might this mean to the average 7th grade teacher? What is the chance that if we give 7th grade teachers our software and curriculum and arm them with this slogan that they will understand what new mathematics we seek to provide to students? We must describe the intended instrumental genesis, if teachers and classrooms are to achieve it. In this way, our approach follows upon and aligns to earlier work with Cabri-Géomètre, which found that integrating ICT into the everyday life of the classroom requires finding a balance between the old and the new (Assude and Gelis 2003).

Achieving the appropriate integration of old and new turned out to be a complex and subtle problem in our 7th grade experiment. Our preferred target grade level was 8th grade; we ended up working in 7th grade because our Texas partners felt that 8th grade was too sensitive due to high stakes accountability measures in that grade level—few schools would allow experimentation in the 8th grade. Further, we planned for a two to three week long replacement unit, because again, this is all we could expect schools to sign up for as a "trial" of a new innovation. Given the grade level and the short time available, we could not target the signature "opportunity to learn" of the SimCalc MathWorlds® software environment—the relationship between velocity and position graphs. What would be a reasonable goal for SimCalc curriculum and technology in the 7th grade? What could we expect to measure under these constraints?

Answering these questions required extended and concerted effort from a team of mathematicians, curriculum and teacher professional development experts, and measurement experts. We quickly identified "proportionality" as the Texas curriculum standard most related to our innovation. Further analysis revealed that the existing curricula treated proportionality only with the formula "$a/b=c/d$." In a SimCalc approach, we would introduce the idea of a constant of proportionality, e.g., "k" in $y=kx$. This constant represents rate of change, a key SimCalc goal. However, as we worked on the issue, we came to the realization that the goal of SimCalc was not simply to shift teachers from a formula with four symbols to a formula with three symbols. Indeed, we realized that the existing Texas curriculum only presented students with proportionality tasks in which they were given three numbers and asked to find a fourth. If we simply taught students to use a formula with three slots instead of a formula with four slots, we would have accomplished nothing.

Instead we came to focus on two differences. First, we wanted to create an opportunity for students to learn a *function* interpretation of proportionality, in contrast to a *formula* interpretation. Synergistic with this goal, we wanted students to reason with proportional functions across representations, including narratives, equations, graphs, and tables. A key distinction was reasoning with a small set of numbers (e.g., given three numbers, find the fourth) vs. reasoning about the linear, multiplicative mapping from a domain to range. Graphs are useful for presenting tasks that require the latter, function-oriented reasoning. For example, in a graph it is possible to ask (without any numbers or indeed without a precise scale defined on the Cartesian axes), which of these two graphed functions has a larger value for "k" in the function $f(x)=kx$? With this distinction in hand, we could both

articulate to teachers what students already learn and what new learning opportunities we were trying to create. Further, we could define a measurement that would contain two scales, one sensitive to the formula aspects of proportionality and another sensitive to the function and multiple representation aspects.

It is beyond the scope of this paper to describe our complete assessment development process. A more limited point is that mathematicians were required at multiple stages in that process: to define the assessment blueprint precisely, to generate candidate test items, to align the items to standards, and to analyze student protocols (for the purpose of seeing if students were in fact invoking the target concept or solving the problem by some "trick"). The bottom line is that the process of describing target mathematics with enough specificity and building an assessment to measure it is a challenging and important problem for any innovation that aims to go to scale.

There was also much specifically mathematical work in defining our pilot SimCalc intervention. This was an integration that consisted of a curriculum unit and fifteen SimCalc MathWorlds files, along with 5 days of teacher training. As discussed above, we chose a topic—proportionality as represented by linear functions of the form $y=kx$—at the heart of Texas' 7th grade curriculum. The *Managing the Soccer Team* unit developed the proportionality concept through ten lessons, keyed to specific software files. Lessons were built around problems related to soccer team management tasks such as timing runners and purchasing new uniforms. Guiding questions for each problem led teachers and students through common SimCalc instructional routines such as predicting the outcome of an animation based on a graph and using the software to check and modify their predictions. A 5-day training workshop had two parts. The first 2 days used a high-quality Texas professional development workshop, called TEXTEAMS, that gave teachers adult-level foundational mathematical knowledge of the $y=kx$ approach to proportionality. The last 3 days focused on teaching the SimCalc unit. Teachers experienced the unit as learners, with additional focus on practical implementation issues. Teachers wrote day-by-day lesson plans for using the materials in their particular classroom contexts. Details of unit formatting were designed to facilitate ease of use: We provided individual student workbooks with color illustrations and space in which to write answers and show work. In addition to SimCalc-specific instructional routines, we included approaches familiar to Texas teachers. In these ways, we balanced the competing design constraints of fidelity to the innovative characteristics of SimCalc while maintaining familiarity and ease of use for teachers who were used to using textbooks with a less interactive approach and with little or no technology integration.

In this 2-year randomized experiment, we sought to answer the primary research question: Can a wide variety of teachers use innovative technology to create new opportunities for students to learn complex and conceptually difficult mathematics? We used a pre-test/post-test control group design in Year 1, and a delayed treatment for control teachers in Year 2. Teachers were randomly assigned to either the Treatment or Control group. In Year 1, teachers in the Treatment group received the SimCalc intervention as outlined above, and were asked to teach the SimCalc replacement unit in place of their usual rate and proportionality unit. In designing an appropriate control condition, important considerations were that teachers across both groups were comparable in their experience of the usefulness of the intervention, belief that they were part of a new and special project, amount of work required for participation, compensation for work, opportunity for teachers to interact with colleagues around intervention topics, and support from the research team. Thus we offered Control teachers the same high-quality 2-day TEXTEAMS training as we did in the SimCalc intervention and promised Control teachers they would receive the

SimCalc intervention in Year 2. During the first school year, we asked Control teachers to teach rate and proportionality as usual, with the option of supplementing with TEXTEAMS materials.

Our main outcome measure was student gains on an assessment of knowledge of rate and proportionality. Students took the pre-test just before the proportionality unit was taught and the post-test just after. We compared the growth in student knowledge between the Treatment group (using the SimCalc replacement unit) and the Control group (teaching rate and proportionality as usual).

Complete discussion of the pilot results with our sample of 21 teachers will soon be published (Tatar et al. 2008). Within the scope of this article, we focus on what we learned about robustness. The pilot demonstrated that we could measure the impact of SimCalc's integrated approach versus the best available alternative. In particular, we found that students in the Treatment group learned considerably more than students in the Control group—particularly mathematical concepts that went beyond the Texas standards emphasis on simple proportional relationships (e.g., that can be described with the formula $a/b=c/d$) to an understanding of proportionality as a function over a domain across multiple representations (graphs, tables, formulas).

Measuring the impact is important, because if there is no measurable impact then one cannot possibly investigate its robustness. However, measuring overall impact is not enough to assert robustness; we need to measure impact across different subpopulations, as well as model variability. By measuring impact across different subpopulations, we can gather empirical evidence to support whether or not our intervention is truly robust. By modeling how different sources of variation contribute to student outcomes, we can learn about ways that the intervention may vary in robustness or may require modification to strengthen its robustness.

The pilot revealed variability across contexts. As we worked with the results, one particular representation became the focus of our attention. Figure 1 shows the mean gain on the student assessment by classroom. It makes clear that there was a lot of variation, with some classrooms realizing student learning gains that were twice as large as the others. Further, we came to understand how little we knew about the key factors underlying that variability. Despite a substantial literature, we realized that we knew too little to isolate just a few variables, and we had too few participating teachers to reach any valid conclusions about any variables.

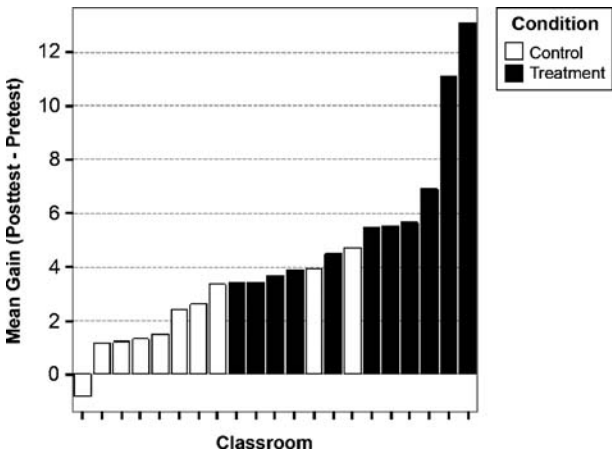

Fig. 1 Graph shows the mean gain on the student assessment by classroom, in the pilot. The assessment had 30 items total

As we observed growth in all of our Treatment classrooms, we had some strong evidence to suggest that the intervention was robust across varying contexts. However, with this limited sample size, we were still limited in the generalizations that we could make about robustness across the multitude of subpopulations within the state. We would need to collect more information to be able to make stronger claims of generalization and external validity.

6 Research phase 3: full scale experiment

Given the success of our pilot study, we decided to use the same SimCalc intervention and basic experimental design. Building on the development and findings, we also improved three things in the full-scale experiment. First, we improved the measures. We revised the assessment of knowledge and increased precision and number of measures of variability in the context of implementation (e.g., teacher daily logs of implementation). Since we were not sure that any single factor would dominate robustness, we decided to collect data on many possible factors. Second, we moved to a more naturalistic recruitment method; one aligned to one way in which new materials ordinarily go to scale in Texas. Texas is divided into 20 educational service regions, each of which has an Educational Service Center (ESC) that provides professional development to teachers. Working with ESCs gave us a region model, which in the end allowed us to contrast robustness by region. Regions in Texas differ in important ways: urbanicity, ethnic distribution, and socioeconomic status (for example, we recruited in both metropolitan Austin and the very poor region along the Mexican border). Third, we increased the number of participating teachers to attain enough power to test complicated models.

With the increase in number of teachers, we determined also to measure relevant characteristics associated with the teachers. This in turn led us to the work of Deborah Ball and colleagues (Ball et al. 2005) on Mathematical Knowledge for Teaching (MKT). We hypothesized that how much mathematics a teacher knows would relate to students learning (and instrumental genesis). A team that included mathematicians and mathematics educators set out to create an assessment for teachers that reflects the kinds of mathematical knowledge needed in classroom teaching with SimCalc. For example, we have observed that students in SimCalc classrooms often generate unusual conjectures (e.g., "a shorter line means a faster motion"). We wanted to know if teachers could evaluate these conjectures. Again, we see that going to scale was an important forcing function for becoming more specific about our mathematics. In this case, we had to become more specific about what mathematics teachers should know in order to best support student learning.

Roschelle et al. (2007) provide the primary discussion of our full scale experiment. Here we focus on a few outcomes relevant to the discussion of robustness. The results of the full experiment revealed important characteristics of the robustness of SimCalc's integrated approach. As in the pilot experiment, we found a main effect of student learning (e.g., students learned more about the $y=kx$ approach to proportionality in the Treatment condition). This established the validity of our claim that the SimCalc integration is responsible for the outcomes we observed. Also as with the pilot, a similar picture emerged as in Fig. 1—the mean gain varied by classroom. Now we have enough data to begin modeling which dimensions matter.

Previously, we discussed the TIMSS finding that teachers in different countries modulate the use of curricular materials, for example by emphasizing routine procedures or conceptual understanding. One way we measured this in the full-scale experiment was by

asking teachers to complete a daily log. On the daily log, teachers responded to the question "To what extent did you focus on the following performance goals for your students?" Five goals were listed. After the fact, we grouped these goals into two categories. We considered "memorizing facts definitions and formulas," as well as "perform procedures/solve routine problems" to be "simple" goals. We considered "communicate understanding of concepts," "solve non-routine problems/make connections," and "conjecture, generalize, or prove" to be "complex" goals. We condensed our data for analysis by aggregating the frequency and intensity with which each teacher emphasized complex goals over all days of teaching the rate and proportionality unit.

Recall also that we sought to measure student learning of both the "formula" and "function" views of proportionality. In the formula view, students find a missing value. In the function view, students must consider co-variation between two variables (e.g., position and time). We determined to look at the relationship between teacher report of "complex" teaching goals and student learning of the "function" view. Figure 2 shows the results.

Our main effect can be seen in this figure by comparing the data points for the treatment condition (SimCalc) vs. the data points for the control condition (the existing curriculum). Most of the treatment data points are higher than most of the control data points, corresponding to our main finding that students who used SimCalc learned more. One can also see considerable variability in the complexity of teachers' reported goals. Given the very short (but realistic) amount of time available for teacher professional development, we did not try to get all teachers to have the same teaching goals, and clearly they didn't choose the same goals. Yet teachers who used SimCalc were more likely to report more complex teaching goals [$p<0.05$]; this can be seen in Fig. 2 in the clustering of treatment teachers' data points further to the right. Most interestingly, the two measures correlate strongly [$r(95)=0.40$; $p<0.01$]. In classrooms where teachers reported more complex teaching goals, students learned more with respect to the function view of proportionality.

In further analysis, we plan to investigate this variable along with others in a hierarchical linear model (HLM). A hierarchical model is needed because we will be looking at the impact of variables at the teacher level (in this case, teacher goals) upon outcomes at the student level. HLM modeling will be able to help us understand the sources of variability. How much of the variability in outcomes relates to what students bring to the classroom vs. how teachers teach? We may be able to rank different factors, such as the importance of the level of experience of the teacher, their comfort with technology, the depth of their mathematics knowledge, the amount of time they allowed students to use the technology,

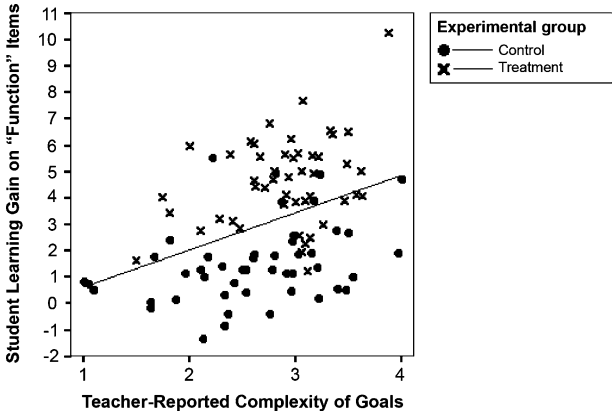

Fig. 2 In the full scale experiment, student learning gains on the more advanced "function" view of proportionality were higher when teachers reported more "complex" teaching goals. The *line* represents the regression of gains onto complexity

etc. Knowing which variables have the most impact would allow us to target efforts to further improve SimCalc to the most powerful factors.

It is worth noting that while individual teachers may have modulated the impact of SimCalc, as a group, students learned more with SimCalc regardless of what their teachers did. While the findings related to this claim will be reported in more detail elsewhere, here we give a brief look at major subpopulations at different levels of our research design.

In terms of students, we disaggregated the data and looked at girls and boys. Both have equivalent gains. We also looked at Hispanic versus white students. Both have similar gains, although the Hispanic students, on average, start with a lower incoming score. As Hispanic ethnicity is highly correlated with poverty in the study's regions of Texas, the same results apply to students in schools that are of lower versus higher socioeconomic status.

Across classrooms, we also saw robustness. One important dimension of variation was the average pre-test score of the classroom. We found that the SimCalc integration produced gains both for classrooms that start out with a low average pre-test score and for classrooms in which students know more about rate and proportionality at the onset. This is, of course, very important, because it is hard to target educational materials only to students with particular levels of incoming knowledge. Conversely, it is good to know that the materials produce benefits in classrooms throughout the range of possible incoming knowledge.

Finally across regions, we also found robustness. The Rio Grande Valley region (near the Mexican border) is a very different place than suburban Austin (the state capital) or suburban Fort Worth (a major metropolis). Likewise, the very rural region of western Texas is also quite distinctive. Yet, the pattern of results looks remarkably similar in all these places.

7 Discussion: the challenge of designing robust interventions

The overall purpose of this article is to make a case that experimental scaling up research can contribute to mathematics education by providing evidence of the robustness of the innovations developed by researchers in mathematics education. We have made that case by examining three phases in the SimCalc research program. In the first phase, we performed design experiments in different locations. While much was learned, incommensurate analyses made it difficult to address robustness. In the second phase, we undertook a pilot experiment with two dozen teachers. Although this allowed us to demonstrate a main effect, we also found considerable variation in the classroom learning gains attained by different teachers. Given the small sample, we were unable to understand the sources for this variability. In the third phase, we executed a large randomized experiment with 95 teachers. Not only were we able to directly establish robustness across varied institutional settings (classrooms, schools, regions), we have been and will continue to be able to identify sources of variability in the results. For example, we found that the complexity of the goals reported by a teacher correlated to student gains in learning the function view of proportionality. In forthcoming work, we plan to engage in hierarchical linear modeling to more fully explicate how variations in what teachers think, know, and do relate to variability in the learning of their students. We note that only the full scale experimental research was able to satisfactorily address the question of robustness.

Although we do not think that experimental scaling up research reduces to previously discussed "instrumental" and "institutional" views of research on ICT integration, it does relate to these views. As in the institutional view, robustness enables us to compare what happens in different institutional settings. But it also goes beyond comparing institutions by

integrating data at the individual, classroom, and contextual levels in one hierarchical linear model. As in the instrumental view, within experimental scaling up research we can look at the degree to which students develop different cognitive outcomes using the same mediating tools and materials. In this paper, we highlighted one factor in varied cognitive outcomes with SimCalc—the degree to which a teacher reports having the goal of going beyond memorizing facts and performing simple procedures. As we examine the detailed case studies that have been embedded in selected classrooms in the full scale experiment, we may be able to root an account of "instrumental genesis" with SimCalc MathWorlds® in a broader picture of varying teacher practice and student outcomes. In addition, experimental scaling up research may cause us to rethink aspects of instrumental genesis. For example, it is clear that less economically advantaged children come into a SimCalc experience with less preparation for all it can offer than do more economically advantaged children, even though both populations learn more with SimCalc than with their existing materials. This draws our attention to the need both to understand instrumental genesis and to work to achieve it with different populations.

A secondary theme throughout the three phases addresses increasing the robustness of the innovation by more fully specifying the theory and intended implementation of an innovation. As Cerulli et al. (2007) reported, when a mathematics education approach goes from its site of origin to additional sites, one often finds what once seemed to be a clearly specified approach now seems in dire need of further clarification. In this drive to more systematically define the approach, much research of a particularly mathematical character occurs. For example, across our phases, we noted the need to re-engage mathematicians with SimCalc to achieve greater clarity in specifying the mathematics we wanted students to learn and the mathematics that we thought their teachers would have to know. Scaling up brings benefits back to mathematics education by creating a pressure for clarity.

Additionally, the process of doing scaling up research foregrounds the warrants for making valid claims about robustness. In our case, it focused our attention on the challenge of recruiting a wide variety of teachers and also on learning how previous curricular innovations had achieved scale in Texas. These two considerations led us to engage Texas Educational Service Centers as recruitment and teacher professional development partners. Assessment is another issue that comes to the fore in scaling up research as validity is considered. Valid claims about robustness require careful specification of the target content and the means for measuring students' achievements. Finally, a concern for the warrants for making valid claims about robustness draws attention to experimental design.

The drive towards robustness leads to secondary consequences that may be as important as the primary ones. Cohen et al. (2003) have made a compelling case that researchers should think of the clarity of their innovations in terms of the levels of ambition and specificity. An innovation that is ambitious is harder for schools to adopt and positive results are less likely. An innovation that is nonspecific is easy to implement but prone to undesirable mutations. Some of these mutant variations may be acceptable to the innovators and have a small impact on the usefulness of the innovation. Others may be "lethal mutations" which undermine or negate the intended impact of the innovation (Romberg and Kaput 1999). For example, when the first author first entered this field, he observed a classroom using Logo as a tool for teacher-led typing instruction. This is obviously a lethal mutation from which one would expect none of Logo's purported benefits.

As our team considered SimCalc initially, we saw an ambitious, nonspecific innovation. The ambition is clear in Kaput's slogans, such as "democratizing access to the mathematics of change." That slogan carries the spirit of Kaput's vision, but, as we specified, we had to consider what elements were of primary importance and what elements could be deferred.

Examples include the scope of the curriculum, the nature of associated professional development, and the meaning of the enacted curriculum.

In reality, it would take an inordinately large commitment from schools to test Kaput's vision completely. Pragmatically, schools were unlikely to allow us to occupy more than a few weeks of the curriculum. This raised the dilemma: what could we expect students to learn in 2 to 3 weeks?

Furthermore, while SimCalc's ambition is to provide a new representational infrastructure that provides teachers with open-ended (and thus non-specific) opportunities for students to learn a wide range of more and better mathematics (Fishman et al. 2003), the need for more specification focused us on details that were actually quite important. Specification is required for teacher professional development (Elmore 1996). Our ambitious, non-specific mission was to "democratize access to mathematics of change and variation." Scaling brought us face to face with the realization that this was not something we could say in a meaningful way to a roomful of 7th grade teachers, at least not without substantial elaboration. Even when we become more concrete, specifying "Teach $y=kx$, not just $a/b=c/d$", a substantial question remained about interpreting the meaning of that new mathematics. We argue that, even in this more concrete case, the odds of teachers understanding the goals and purpose of the intervention from this statement alone are low.

The huge range of possible interpretations made it likely that the "enacted curriculum" would differ significantly from the "intended curriculum" (Porter 2002). Not only would this would make it hard to measure the impact of the innovation at scale but it would arguably make it ineffective as an innovation. An early and important argument about scaling up is that innovations require intentional processes that reproduce their success in new settings and allow for incremental growth as teachers become more familiar with the innovation (Kaput and Roschelle 2000). Our theory of change calls for teachers to have a supported pathway for innovation.

In our scale up research, we tackled the issue of ambition by more precisely defining the mathematics to be covered, as discussed in the prior section, and confining ourselves to a two to three week intervention. Specificity, however, turned out to be a difficult challenge. One approach would have been to transform the SimCalc MathWorlds® software from an open-ended tool to a very tightly scripted series of applets. We decided against this because it would undermine the core character of the innovation as a "representational infrastructure." Instead, the team decided to carefully contextualize the software with a well-specified paper curriculum.

8 Conclusion: towards robustness in design and evaluation

One of the great strengths of the Kaput's research program was his relentless drive towards scale. The slogan "democratizing access to the mathematics of change" was enacted by deliberate efforts to reach an increasing number of teachers and students each year. Further, Kaput asked:

> What can we learn from research at the next level of scale (e.g. beyond a few classrooms at a time) that we cannot learn from other sources?

The answer developed in this article is:

> We can learn about the robustness of our innovations for teaching and learning mathematics.

Additionally, research at scale promotes robustness through two mechanisms: providing evidence about variance in implementation and performance and by driving the innovators to think systematically about and develop pragmatic programs of action that reflect the nature of their hopes and expectations.

The intention to go to scale is a good forcing function for surfacing implicit curricular structure and demanding clear choices about how specific an innovation must be for a wide variety of teachers to have a reasonable chance at implementing it well. The designers of each innovation, of course, must make their own choices about how ambitious a change they will try to make in classrooms and how elaborate their materials and services that support that change will be. We think the overall strategy, however, of being more specific and less ambitious in first implementations but relaxing these constraints as teachers gain experience makes sense. Further, we found that the process forced consideration of potential lethal mutations and what to do about them. In design experiment research, lethal mutations are often not considered because careful choice of participants and close monitoring of implementation rules them out. Yet at scale, designers have much less control over what will happen. It is important to the potential success of an innovation to make the decisions about ambition and specificity more explicit.

While we do not wish to argue that scaling research or our path in it is the only valid route to this knowledge, we note some particular advantages to engaging in a large-scale randomized, controlled experiment with hierarchical linear modeling when the goal is to evaluate robustness. Carrying out a comparison with a control group ensures that we are evaluating the robustness of our intervention and not simply gains due to some other component of students' experience. Assigning teachers randomly increases our confidence that the groups were equal at the onset of the experiment; again it strengthens our claim that we are measuring the robustness of the innovation and not a preexisting difference in two different sets of classrooms. Conducting a large-scale experiment allows us to consider the performance of important subgroups, not just an overall average effect. Further, hierarchical linear modeling allows us to account for the fact that variability in teaching and learning occurs at multiple levels, for example, a student-level, a classroom- or teacher-level, a school-level, and a regional-level. By modeling, we can begin to learn which factors matter most as an innovation goes to scale. Ultimately, this can allow us to target designs and resources to mitigate important disadvantages found in certain settings, while avoiding wasting resources to address differences with less important impact on outcomes.

Our findings about robustness should have a number of important implications related to bringing educational innovations to scale. Generally speaking, researchers have tended to position more ambitious technologies such as SimCalc, The Geometer's Sketchpad®, Cabri-Géomètre or Fathom as tools that make sense as part of a broad, comprehensive approach to improving instruction—one that includes expansive curricular reform coupled with long-term, integrated teacher professional development. Our pilot results may suggest that with careful attention to curricular integration, educators who wish to introduce students to richer mathematics can proceed to incorporate lessons that use innovative representational technologies into their classrooms and schools without waiting for comprehensive reform. Further, when extensive ongoing support is not possible, a limited training-to-use-materials approach still appears to support teacher learning and enhance students' opportunities to learn in the classroom. Indeed, starting small with a replacement unit and modest teacher training may be an effective strategy for scaling up. A replacement unit gives teachers a well-defined, safe, bounded experience with new curriculum and new technology. Success in the small may transfer to broader and more profound changes later in a teacher's career.

There is much more to do within our program. As Kaput pointed out, no single integration could possibly evaluate his vision and representational infrastructure approach. We are also conducting a parallel experiment in 8th grade, in this case implementing a train-the-trainers model, which is an additional component of scale. Further, we are studying the impact of a second year of teaching with these materials and additional professional development on performance in 7th grade. In the future, we would like to expand to studies that follow students as they make the transition from middle school (6th to 8th grade in the United States) into high school and track implications for both science and further mathematics courses. Eventually, we would like to examine the cumulative, longitudinal benefits for students who experience the SimCalc approach as Kaput intended it, as a portion of the curriculum in each year of instruction from middle school through high school.

At the core of all these directions and all this potential lies Kaput's vision. It was a vision in which the *mathematics* in mathematics education was central. It has been important to us to have the continued support of mathematicians and mathematics educators at the core of our team as the team proceeded through the additional work of scaling up. As we continue to strive for impact and growth, the allegiance to Kaput's profoundly mathematical vision of what students could learn with new representational infrastructure remains at the core of our thoughts.

Acknowledgement Thank you to our colleagues who helped carry out our scaling up research at SRI International, the University of Massachusetts, Dartmouth, Virginia Tech, the University of Texas, Austin, and the Charles A. Dana Center. We also thank all the teachers and educational service center leaders who participated in this research. This material is based upon work supported by the National Science Foundation under Grant No. REC-0437861. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Assude, T., & Gelis, J. M. (2003). La dialectique ancien-nouveau dans l'intégration de Cabri-géomètre à l'école primaire. *Educational Studies in Mathematics*, *50*(3), 259–287.

Baker, E. L. (2007). Principles for scaling up: Choosing, measuring effects, and promoting the widespread use of educational innovation. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education* (pp. 37–54). Lanham, MD: Rowman & Littlefield.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide. *American Educator*, *29*(3), 14–17, 20–22, 43–46.

Brown, A. (1991). Design experiments. Theoretical and methodological challenges in evaluating complex interventions in classroom settings. *Journal of the Learning Sciences*, *2*(2), 141–178.

Cerulli, M., Georget, J. P., Maracci, M., Psycharis, G., & Trgalova, J. (2007). Integrating research teams: The TEMLA approach. Retrieved on September 12, 2007, from http://telearn.noe-kaleidoscope.org/warehouse/TELMA_CERME5_conference_version.pdf.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3–12.

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. (1999). *Considering the major arguments against random assignment: An analysis of the intellectual culture surrounding evaluation in American schools of education*. Evanston, IL: Institute for Policy Research at Northwestern University.

Dede, C. (2006). Scaling up: Evolving innovations beyond ideal settings to challenging contexts of practice. In R. K. Sawyer (Ed.), *Cambridge handbook of learning sciences* (pp. 551–566). Cambridge, UK: Cambridge University Press.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, *66*(1), 1–26.

Fishman, B. J., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and Teacher Education*, *19*, 643–658.

Fullan, M., & Earl, L. (2002). Large scale reform. *Journal of Educational Change*, *3*, 1–5.

Hawkins, J. (1997). *The national design experiments consortium: Final report*. New York: Center for Children and Technology, Educational Development Center.

Hedges, L. V. (2007). Generalizability of treatment effects: Psychometrics and education. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education* (pp. 55–78). Lanham, MD: Rowman & Littlefield.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., et al. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study.* Washington DC: National Center for Educational Statistics.

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: Lawrence Erlbaum Associates.

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

Kaput, J. (2000). Implications of the shift from isolated, expensive technology to connected, inexpensive, diverse and ubiquitous technologies. In M. O. J. Thomas (Ed.), *Proceedings of the TIME 2000: An International Conference on Technology in Mathematics Education* (pp. 1–24). Auckland, New Zealand: The University of Auckland and the Auckland University of Technology Also published in the *New Zealand Mathematics Magazine, 38*(3), December 2001.

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* (pp. 155–170). London, UK: Falmer.

Kaput, J., & Roschelle, J. (2000). *Shifting representational infrastructures and reconstituting content to democratize access to the math of change and variation: Impacts on cognition, curriculum, learning and teaching*. Paper presented at the NSF Workshop to Integrate Computer-based Modeling and Scientific Visualization into K-12 Teacher Education Programs. Reston, VA: National Science Foundation.

Kaput, J., & Shaffer, D. (2002). On the development of human representational competence from an evolutionary point of view: From episodic to virtual culture. In K. Gravemeijer, R. Lehrer, B. van Oers, & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* (pp. 277–293). London: Kluwer Academic.

Lagrange, J. B., Artigue, M., Laborde, C., & Trouche, L. (2003). Technology and mathematics education: A multidimensional study of the evolution of research and innovation. In A. J. Bishop, M.A. Clements, C. Keitel, J. Kilpatrick, & F. S. Leung (Eds.), *Second international handbook of research in mathematics education* (pp. 239–271). Dordrecht, The Netherlands: Kluwer Academic.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

O'Neil, J. (1995). Teacher and technology: Potential pitfalls. *Educational Leadership*, *53*(2), 10–11.

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

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

Reichardt, C. S. (2007). Estimating the effects of educational interventions. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education* (Vol. 1, pp. 79–99). Lanham, MD: Rowman & Littlefield.

Rogers, E. M. (2003). *Diffusion of innovations*. New York: Simon and Schuster.

Romberg, T. A., & Kaput, J. (1999). Mathematics worth teaching, mathematics worth understanding. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum Associates.

Roschelle, J., & Jackiw, N. (2000). Technology design as educational research: Interweaving imagination, inquiry & impact. In A. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 777–797). Mahwah, NJ: Lawrence Erlbaum Associates.

Roschelle, J., & Kaput, J. (1996). Educational software architecture and systemic impact: The promise of component software. *Journal of Educational Computing Research*, *14*(3), 217–228.

Roschelle, J., Kaput, J., & Stroup, W. (2000). 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: Lawrence Erlbaum Associates.

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

Roschelle, J., Tatar, D., & Kaput, J. (2008). Getting to scale with innovations that deeply restructure how students come to know mathematics. In A. E. Kelly, R. Lesh, & J.Y. Baek (Eds.), *Handbook of innovative design research in science, technology, engineering, mathematics education*. Hillsdale, NJ: Lawrence Erlbaum Associates (in press).

Roschelle, J., Tatar, D., Shechtman, N., Hegedus, S., Hopkins, B., & Knudsen, J. (2007). *Can a technology-enhanced curriculum improve student learning of important mathematics? Results from 7th grade, year 1* (No. 1). Menlo Park, CA: SRI International.

Schneider, B., & McDonald, S. K. (2007). Introduction. In B. Schneider & S.-K. McDonald (Eds.), *Scale-up in education* (pp. 1–15). Lanham, MD: Rowman & Littlefield.

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.

Tatar, D., Roschelle, J., Knudsen, J., Shechtman, N., Kaput, J., & Hopkins, B. (2008). Scaling up innovative technology-based math. *Journal of the Learning Sciences* (in press).

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