6

SCALING UP DESIGN OF INQUIRY ENVIRONMENTS

Jeremy Roschelle, Claudia Mazziotti, and Barbara Means

Introduction

Researchers, developers, and educators often describe their motivation for designing inquiry learning environments as a response to societal challenges. In today's global economy, many routine tasks may be performed by machines, and people may be called upon to address complex problems requiring innovative, insightful solutions. All students need to be able to think critically and solve problems given the accelerating economic, technical, political, and cultural changes. Thus, we need inquiry learning opportunities that will make an impact at scale and thereby broaden participation in inquiry-related learning and careers. How should developers, researchers, innovators, and educators who are advancing inquiry think about scale?

The literature on scaling up educational innovations has intensified over the past two decades, spurred by funding both for bringing innovations to scale and for studying the scaling process itself. For example, in the United States, the Interagency Educational Research Initiative (IERI, n.d.) was launched in 1999 and led to over 70 research projects and over 100 publications documenting the scaling process across projects. Two edited volumes provide an overview (Schneider & McDonald, 2007a, 2007b). This program was followed in the United States by an Investing in Innovation Fund and later by the Education Innovation Research program, cumulatively investing hundreds of millions of dollars in scaling up promising educational programs. Similarly, Singapore invested in scaling up inquiry-based technologies (e.g., Looi & Woon Teh, 2015), and the United Kingdom created the Educational Endowment Fund (https://educationendowmentfoundation.org.uk/). Many other countries undertook similar efforts to scale up promising education programs and approaches. But perhaps even more consequential than this surge in funding for scaling up was the increased availability of the Internet and inexpensive, powerful computing devices. Greater availability of devices and connectivity have enabled technology-based products to reach large numbers of people with unprecedented speed and sometimes at no cost (e.g., open educational resources) or low cost. Another stream of contributions to the intensification of work on scaling up was the engagement of multidisciplinary scholars, moving beyond traditional educational evaluators to also include computer scientists, domain experts, statisticians, sociologists, educational data mining experts, and other social scientists. Additional forms of expertise become necessary when the problem shifts from defining an innovative support for inquiry to understanding how and why use of that inquiry support spreads or fails to spread (Roschelle, Tatar, & Kaput, 2008). For example, sociologists can shed light on the role that context plays in the implementation of a new educational approach (McDonald et al., 2006). Furthermore, as focus

shifts to school district uptake and improvement, data mining, learning analytic, and improvement science can become valuable for monitoring and adjusting improvement (Krumm, Means, & Bienkowski, 2018).

In this chapter, we review definitions of scaling up, causes of failure, strategies linked to success, and remaining challenges with a focus on inquiry learning environments.

Definitions and Key Aspects of Scaling Up

Building on McDonald et al. (2006), we define scaling up as "achieving greater reach with predictable, measurable impact." In contrast to McDonald et al. (2006), we do not assume that scaling up occurs after something is first "proven," because as we discuss later, there are multiple pathways to scale and some begin before an innovation is proven.

"Greater reach" captures the essence of "scaling up" as "expanding" without being prescriptive about how a particular initiative defines growth. As we will discuss, there can be good reasons to conceptualize growth as something more nuanced than counting the number of users. "Predictable, measurable impact" indicates that the purpose of scaling up is to improve learning for a large population of students, and we need to measure learning to know whether it has improved. Furthermore, we add the word "predictable" because understanding variability is important. Variability is intrinsic to scaling up because of the many local factors in education that vary from place to place; the important thing is to be able to make sense of and predict the variability that will occur when an approach is scaled and to identify strategies for adapting to or otherwise addressing this variability. Greater reach and predictable, measurable impact should be complementary objectives for scaling efforts, but we recognize that progress in these dimensions does not always occur at the same time. Now we proceed to discuss some key additional aspects of scaling up.

In education, scale often takes a long time and may be achieved in different ways. Research on scaling educational innovation goes back to the mid-20th century. Mort (1953, as reported in Dearing et al., 2015) observed that educational change often takes 25 years or more. Subsequent studies of how research-based innovations move into practice have articulated two contrasting perspectives (Dearing et al., 2015): (1) a linear knowledge transfer model in which researchers create and test the innovations and then pass them on to others for dissemination to those who will implement them and (2) a nonlinear, participant-centered model in which educational stakeholders play a decisive role in the design, refinement, and spread of an innovation. The linear view has been embodied in the structure of many funding programs, which progress from exploratory studies to development of innovations with efficacy and effectiveness testing and finally to scale (see, for example, the "common guidelines" issued by the Institute of Education Sciences and the National Science Foundation, 2013). A seminal reference for the contrasting, participant-centered view of scaling is provided by Rogers' (1962) description of "Diffusion of Innovation" and usefully elaborated in von Hippel's (2005) discussion of democratizing innovation. In the nonlinear model, researchers and developers may collaborate with practitioners throughout the project, for example, to respond to a newly identified problem of practice, help define generalizable components of the innovation, and to study emerging impacts or challenges. The border between research and practice may seem more fluid in the case of nonlinear scaling but is still important to identify, investigate, and consolidate an emerging innovation.

Scale as an Experiment with a Large Sample and Sound Measurement

In its most conventional form, research on scale can be operationalized as an experimental comparison that demonstrates a statistically significant effect on an appropriate outcome in a suitably large sample population and across settings. The nature of the outcome and its measurement need

to be carefully defined. For example, whereas research intended to generate scientific knowledge might choose a measure for its bearing on a particular scientific theory, scale-up research usually focuses on measures that are relevant to educational policy. A challenge for research on scaling up inquiry environments is that policy-relevant measures (such as an end-of-year assessment mandated by a state) may not capture the outcomes of inquiry learning well; yet an expectation for scale-up research is often that it will focus on policy-relevant measures. Hence, scale-up research may challenge the existing assessment regime. An example is research on the nature of science learning that informed the Next Generation Science Standards in the United States with the standards then exerting pressure for new kinds of assessments capturing the "three-dimensional learning" embodied in the standards (e.g., DeBarger et al., 2016).

Scale-up research is sometimes differentiated from effectiveness and efficacy research (Flay et al., 2005). Efficacy research must be rigorous and sound in experimental design and statistical analysis but can be conducted with "best case" levels of support to practitioners. Effectiveness research aims to evaluate the approach in realistic (not "best case") conditions (and with realistic variation across settings) and thus to better establish the practical significance of the approach. Scale-up research may add information about program costs and tools to monitor and improve the quality of implementation. Summarizing with regard to inquiry environments, scale-up research in education should examine a novel approach "under circumstances that would typically prevail in the target context" (IES & NSF, 2013, p. 9) without substantial developer involvement in implementation or evaluation and should include practical information about program costs and how practitioners can monitor and improve implementations.

The stipulation of typical circumstances is important because if one scales only to those participants who are most willing, or have the most support, the program may scale to "early adopters" but never "cross the chasm" to broader populations (e.g., Moore, 2014). If the sample does not reflect the variation in contexts and capacities in the eventual target population, using statistics to support the generalization of inferences based on the sample to the target population is not possible (Tipton, 2014). Cartwright (2012, 2013) adds concerns for the degree to which an experimental trial is sufficient to answer practical questions about whether and how an inquiry approach will scale. For example, there are differences between finding that an approach works in an initial collection of settings, that it works in a wide variety of settings, and that it is likely to "work here" (in a practitioner's particular setting). At the heart of Cartwright's argument, external validity of research requires attention to both capacities of an approach (e.g., how a particular inquiry approach drives learning) and the capacities of settings (e.g., the people, policies, and practices in schools); the subtle interrelations between the two is often not captured in reporting on randomized controlled trials. Furthermore, measuring outcomes for inquiry environments is challenging. Measured impact is almost always stronger with proximal measures that are closely aligned to the new learning environment than to more distal outcomes such as end-of-year mandated tests (Ruiz-Primo, et al., 2002). For example, in the study of scaling up the SimCalc learning environment (Roschelle, Shechtman, et al., 2010), researchers found significant positive impacts for assessment items related to the conceptual skills emphasized in the SimCalc mathematical inquiry environment but no significant difference in students' performance on items chosen from the relevant state test.

Researchers can also conceptualize scaling research as applying the method of meta-analysis to many related studies. When many studies have been conducted on an approach, each in different settings, a meta-analysis can combine the findings through statistical methods, resulting both in a more precise estimate of the impact of the approach and also in identification of variables that moderate or mediate the effect. For example, Furtak et al. (2012) identified 37 studies of inquiry-based instruction and found an overall positive effect. In addition, they were able to identify some key variables that mediate the size of the effect, such as the degree of focus on epistemic activities

and whether instruction was led by a teacher. The availability of many studies conducted by different teams in different locations is prima facie evidence that an approach is scaling up. Furthermore, by pooling data from unrelated experiments, researchers can increase their confidence in their estimation of the average size of the effect and ascertain the degree to which it is dependent on factors unique to one setting or one implementation. The combination of prima facie evidence of use in many different settings along with evidence of consistent effects across settings can be used to argue that the approach effectively scales up and can also address the issue of replicability (e.g., Makel, Plucker, & Hegarty, 2012).

Scale Is Not Just a Bigger "n"

When we think about scale, it is important to consider not only the number of participants but also the qualitative nature of their participation, which can include changes in depth, sustainability, ownership, and the evolution of an approach.

Coburn (2003) describes how two innovations may reach a similar number of participants but still vary in what she describes as the *depth* of scaling. For example, an inquiry approach may be superficially employed by asking students to conduct a fixed lab experiment that is related to instruction or more deeply employed by having students design their own investigation of a driving question. Likewise, two approaches may each reach 200 teachers but vary in the density of penetration. One approach may penetrate a school district thoroughly, reaching every science teacher in the district. Another may choose friendly teachers in 200 different school districts, which has less depth from a district's perspective. Another depth factor may be the degree to which the outcome measure aligns to a deep conception of inquiry; a performance task or scenario-based task is generally regarded as a deeper assessment of inquiry than a set of multiple-choice items (Scalise & Gifford, 2006). Likewise, an innovation that is used for a very short amount of curricular time would be considered to have less depth than an inquiry approach used for an entire block of instruction or school year.

Another measure of scaling depth is *sustainability:* how easy or difficult it is to keep the approach going after an initial usage in a new setting or after research-based support is withdrawn. A related indicator of scaling depth is *shift in ownership*, with educators coming to view the innovation as "their" approach rather than something coming from an external entity. For scaling to occur, educators must come to feel ownership of the innovation.

To Coburn's list of characteristics, Clarke and Dede (2009) added *evolution*, by which they mean the degree to which the adopters, in collaboration with the developers of the approach, are learning and revising as scale occurs and improving the fitness of the approach for further scaling. Realistically, few things scale without adaptation to local settings, and an evolution dimension of scaling can reflect a process of making an approach adaptable without sacrificing its integrity.

For Whom and under What Conditions?

Building on the brief discussion of moderator variables above, a further important set of considerations regarding scale has to do with *for whom and under what conditions* an innovative approach delivers improvements. Typical moderator variables can include the age of the students, their gender, ethnicity, or language-learner status, socioeconomic status, and prior achievement scores. Thus, it is important to find out whether an inquiry learning environment is scaling only to boys or only to students for whom English is their first language. Furthermore, researchers are often concerned with the "Matthew Effect" (Kerckhoff & Glennie, 1999), whereby students who already have an advantage benefit more from a novel learning approach relative to students who have less incoming advantage. When the Matthew Effect is present, an innovative approach may increase achievement gaps between advantaged and less advantaged students, which is not desirable. Furthermore,

higher-income settings may have more capacity to sustain novel learning approaches. If the approach is more effective and is better sustained in high-income settings, then scaling the approach could increase achievement gaps. The level of school capacity is one example of a "condition" that might moderate the effectiveness of an approach; other typical influential conditions include alignment to standards and accountability, degree of support from administrative leadership, the degree and nature of support for teacher learning, alignment to other materials in use with the same students, and availability and quality of necessary technology or other infrastructure.

One powerful way to conceptualize research on "for whom and under what conditions" an innovation is effective is in terms of the generalizability or external validity of a program of research (Hedges, 2018; Tipton, 2014). Generalizability is also sometimes called "external validity." Analyzing generalizability in a research study requires two things. First, the study must capture a set of variables that describe the study participants ("for whom") and contexts ("what conditions") and that could plausibly moderate the effectiveness or impact of the inquiry learning approach. Second, one needs a data set that describes the prevalence and distribution of those variables in the broader population beyond the study. If these conditions are met, then one can estimate the range of situations to which the approach's measured impact may be reasonably expected to generalize. Imagine that a study finds inquiry learning is working well for both low- and high-SES students but that the study's sample did not include many students who are English Language Learners (ELL). Furthermore, imagine that the study's sample found the approach was more effective in districts that use performance tasks as a district-wide assessment of science learning. If a database with these variables is available for all the schools in a state or country, one could color a map green (the results are likely to generalize), yellow (the results may generalize, but effects may be weaker), or gray (no match between the place and places in the existing data and thus not enough information) to show the approach's demonstrated potential for scalability (see Roschelle et al., 2018, for an example of such maps). Cartwright (2012, 2013) further argues that determining which variables are relevant requires a lot of specific knowledge about how new approaches and existing conditions/practices may or may not fit together. We should not be content to analyze generalizability merely in terms of well-known policy variables, such as ELL status, reduced price and free lunch status, or student race, ethnicity, and gender.

Alternative Research Designs

As mentioned above, the conventional scaling method has been to stage experiments with larger and larger groups of participants, while also measuring impact by contrasting outcomes of a treatment to an untreated condition. For example, the U.S. Institute of Education Sciences has programs that provide funding for larger-scale and rigorously controlled research on implementations of a program. Budgets increase as an investigator goes from exploratory studies to development projects then efficacy projects and finally scaling projects, a progression that also involves increasing the number of students and classrooms experiencing the intervention. Furthermore, the standards of evidence strengthen from correlational and quasi-experimental methods to randomized controlled trials. In some cases, there is also attention to cost-effectiveness, which involves both measuring program costs (which in the case of inquiry-based learning might include instructional materials, experimental apparatus, teacher professional development, and other costs) and program impacts (Levin & Belfield, 2015).

A set of complementary methods focus on the design dimension of scaling up. These methods recognize that scale involves not only testing something with more people but also designing it to be more adaptable and robust in varied implementation environments. Researcher-Practitioner Partnerships (e.g., Coburn & Penuel, 2016; Coburn, Penuel, & Geil, 2013) emphasize the importance of identifying authentic problems of practice and having researchers and educators work

together to address them. Design-based Implementation Research (e.g., Penuel et al., 2011) emphasizes the layers of design needed to support uptake and high-quality implementation of an approach at scale. Networked Improvement Communities in education (see LeMahieu et al., 2017) combine improvement science approaches and a focus on measuring variability and finding ways to retain program integrity and impacts while allowing for adaptations (e.g., Lewis, 2015). Improvement science is one of many continuous improvement approaches (with roots going back to Total Quality Management, e.g., Sallis, 2002) that involve a series of successive advances on a defined metric and thus emphasize iterating toward the future state rather than a single definitive experiment to estimate the impact of an intervention. Like scaling up research, improvement science has an interest in examining local conditions and in analyzing changes caused by an intervention on many different levels (rather than just on student learning outcome measures). In a Networked Improvement Community, multiple entities in the network are employing improvement science practices in designing, implementing, and refining approaches to address a common aim. Variation in the conditions in which the different entities operate becomes a source of information about what's necessary and sufficient for the approach to work. The six core principles for running Networked Improvement Communities involve focusing on a problem of practice, attuning to variation, taking a systems perspective, using measurement to drive improvement, anchoring specific improvements in collaborative investigations, and accelerating overall improvement by sharing in networked communities (LeMahieu et al., 2017).

When Does Scaling up Happen?

Finally, we observe that the phrase "scaling up," like the linear knowledge transfer model, misleadingly implies a discrete phase that happens some time later, after initial research and development is complete. Yet in reality the path to scale is rarely linear or stage-like (Prewitt, Schwandt, & Straf, 2012). Some educational programs with inquiry potential, such as the Scratch computing environment (https://scratch.mit.edu/) and the FIRST Lego League robotics competition (https://www.first-lego-league.org/en/general/what-is-fll.html) have scaled very rapidly, often before early stage R&D focusing on efficacy was available or published. As initial versions of these environments scaled, the researchers and developers working on them continued to define and develop multiple components—technological, human, and organizational—to maximize the learning value of students' engagement (e.g., Melchior et al., 2016). Thus, it is possible for either scaling or program refinement to happen first, and these processes can also occur simultaneously.

A disadvantage of the traditional stepwise approach to getting education innovations to scale is that years of R&D may be devoted to developing and refining an approach, only to find later that it doesn't scale well. This realization can make it worthwhile to seek alternatives to the traditional strategy of first getting an inquiry learning environment working in one or two classrooms, then expanding to 6–10 classrooms, then 100 classrooms and so on—the conventional, "step-by-step" scaling up process. For example, one can scale a digital infrastructure for inquiry learning first and gather data from it to drive improvement research. For example, the Scratch programming environment scaled quickly; this allowed researchers to later look for programming constructs which students are learning or not learning to use (e.g., Aivaloglou & Hermans, 2016). The technology sector often espouses this approach—releasing a "minimum viable product" intended to attract large numbers of users and then leveraging user feedback and data to inform cycles of product refinement (e.g., Münch et al., 2013); increasingly, technology and publishing companies also care about conducting high-quality research (Newman, Jaciw, & Lazarev, 2018).

By making scale an intentional focus early in an R&D program, teams may become aware of pitfalls and address these earlier. Nonetheless, in the field of education as in medicine, a first principle should be "do no harm." If the inquiry learning system will supplant a significant part of

existing instruction in areas for which there are serious stakes for students and teachers, launching an ineffective product at scale may be unacceptably risky. Regardless of the path chosen by a particular inquiry environment team, the lessons about scaling up in this chapter should be considered early in the process, as scaling any ambitious learning activity system (and all inquiry learning environments) is sure to require the disciplined effort of a dedicated team over a long period of time.

Summary

Scaling up is a complex process involving not only reaching more participants but also strengthening measurement and prediction of impacts in varied environments. As one scales, there are changes both to design and in the types of research. Within the notion of "reach," it is important to consider metrics other than the number of participants served and the estimated treatment effect. Additional metrics include shift of ownership, sustainability, and evolution. Furthermore, it is highly important to better understand for whom and under what conditions an approach to inquiry-based learning works. Experimental methods are valuable, as they enable measuring impacts under variable conditions, and much can be learned by doing them. However, complementary design and improvement methods are also valuable and important. Starting small and gradually increasing reach is not necessarily the only or best scaling plan. Especially when technology or favorable policies are available and risks to participants are low, it may make sense to begin implementation at scale. In any event, scaling should be a design consideration early in any significant program of education research and development.

Why Is Scaling Hard?

Scaling inquiry learning environments would seem to be obviously desirable, because opportunities to learn through inquiry respond to pressing needs to educate future citizens for the realities of the future society, culture, and workforce. And yet it is hard to scale inquiry learning environments. Why?

We conjecture that inquiry environments are challenging to scale because they are "Ambitious Learning Activity Systems" (building on the phrase in Roschelle, Knudsen, & Hegedus, 2010). They are "ambitious" because introducing inquiry is typically a big change from existing educational practice (Anderson & Helms, 2001; Roehrig & Luft, 2004). They involve new roles and responsibilities for both teachers and students (van der Valk & de Jong, 2009). Teachers may also be reluctant to implement inquiry-based learning systems because they emphasize depth rather than breadth of content, whereas testing regimes often stress the latter (Penuel et al., 2009). With regard to "learning," because inquiry environments stress the active, mindful engagement of the learner, they cannot be scaled simply by distributing new teaching resources. Teachers often find that they have not experienced inquiry learning themselves, and they cannot be assumed to have the content background and instructional strategies needed to produce supports for this kind of learning successfully (Donelly, Linn, & Ludvigsen, 2014). With regard to "activity," inquiry learning requires changing how students and teachers interact with each other and with resources. Thus, what is to be scaled is not just a new or better piece of content but rather a different form of participation in cognitive and social interactions with resources, peers, and teachers. And finally, conceptualizing a "system" is necessary, because changing learning at scale requires changing many factors at once in a coherent way. A systems perspective can organize the different inputs (like curriculum materials, technologies, teacher professional development) and processes (like new uses of classroom discussions, small-group work, and assessments) into a well-organized and coherent approach to change. Consequently, this chapter focuses on what can be learned from efforts to scale ambitious learning activity systems, with an eye to applying those lessons to inquiry environments.

A good place to contextualize our intuitions about the difficulty of scaling ambitious learning activity systems is a recent broad historical review of efforts to change education by Cohen and Mehta (2017). Their review focuses on adoption of education "reforms" in general rather than inquiry learning environments per se, but the lessons they draw are highly relevant to instructional reforms such as inquiry-based learning. They examine reasons why reform is not easy, particularly in countries like the United States, Canada, Germany, and India, with decentralized control of education.

A first challenge is ***local adoption***. In the United States, for example, education is primarily a function of the states and local education agencies (e.g., school districts). The U.S. Department of Education does not have the authority to impose a curriculum or a teaching approach on the nation's schools. As a consequence, any effort to introduce an educational innovation cannot succeed by winning over a single centralized, national education authority. Instead, reformers must win over each of the 50 states and often tens of thousands of school districts one by one. Furthermore, when adoption is local, one must convince not only educational professionals but also parents and school board officials, and thus a corollary to local adoption is the wide span of stakeholders who must be engaged and convinced. Different communities may have different sensibilities with regard to inquiry; people with different political, religious, or cultural perspectives may want to emphasize (or de-emphasize) different aspects of inquiry in their local schools.

A second challenge for educational innovations is their requirement for extensive ***professional learning*** on the part of teachers, school leaders, and district administrators. Many innovators have failed to consider the amount of learning time and support educators will need if they are to implement new ways of doing things successfully (Cobb et al., 2013). Not only does a reform effort need to design and offer the tools and professional learning experiences needed to implement the reform well, it also needs to solve the problem of finding times when that learning can take place. Time for teacher learning, for example, is very limited in the United States (Darling-Hammond et al., 2017). For working in inquiry learning environments, professional learning is very important, because teachers need to learn not just how to use a specific inquiry tool but also how to change their role in the classroom from that of authority to that of facilitator. This change in roles can be ambiguous in practice (Russ & Berland, 2019).

Finally, at every level of the education system (federal, state, local), there is what Cohen and Mehta characterize as "***remarkable vulnerability to public opinion and political pressure***" (p. 5). Plans in the 1980s to encourage more consistency in what is taught at each grade level in the United States by developing and administering a national test were quickly scuttled in the face of deep-seated political opposition. Similarly, activities encouraging adoption of the Common Core State Standards during the Obama Administration were perceived by many as overstepping the appropriate federal role in K-12 education, and citizens in many states rejected the new standards out of hand. More generally, public dissatisfaction with the functioning of their local school district has resulted in an average tenure for the superintendent in large U.S. school districts of fewer than three years. Elected school boards may not resonate with reform goals or approaches and can fire leaders who introduce them. Battles in the specification of approaches to teaching reading or mathematics (Nicholson & Tunmer, 2010; Schoenfeld & Pearson, 2012) may be instructive to proponents of inquiry learning; one dimension of recurrent policy tension in reading and mathematics is between (a) direct, prescriptive and (b) meaning-making approaches; inquiry learning environments may incur similar debates.

Despite these challenges, some educational changes have scaled successfully (examples follow). Cohen and Medha (2017) report that education reforms that have scaled successfully provided a solution to something that was a problem in the minds of educators or addressed a broader issue perceived by the general public or government (e.g., the need to provide a safe and supportive environment in which five-year-olds could acquire the social and behavioral competencies needed to benefit from academic instruction in first grade). Successful reforms offered the guidance, tools,

and resources educators would need to implement them, and they were consistent with the values of most educators, parents, and students (Cohen & Mehta, 2017).

Going beyond Cohen and Mehta's historical review, researchers and innovators have identified specific factors that can make instructional innovations, such as inquiry learning environments, hard to scale (see Cohen, Raudenbush, & Ball, 2003 for an overview). First, there are three qualities that affect adoption and implementation:

- **Degree of ambition**. The bigger the change from standard instruction and the more it requires revamping the basic organization of education, the more challenges an innovation will encounter, including resistance from those who have a vested interest in the current system or who are simply risk-averse. Thus, an inquiry learning environment that is used as a supplemental or enrichment activity for a topic that is broadly taught is easier to scale than one that would require an entirely new approach to mathematics learning with most of the instruction conducted online. Inquiry environments are often quite ambitious.
- **Complexity**. Complexity often goes hand-in-hand with ambition but is conceptually distinct from it. The more complex the target educational practice, the harder it is likely to be for educators to learn how to do it, and the more pieces of the reform will need to be developed and aligned so that educators can achieve the desired practice. Inquiry environments are often complex, with long-term, multistage activities that a teacher must orchestrate smoothly.
- **Resource-intensiveness**. The more resources an innovation requires, the fewer classrooms, schools, and systems will be willing and able to assemble them in order to implement the innovation. If teachers require 40 hours of training and additional coaching to learn to implement an inquiry learning environment as intended, many decision-makers will judge the intervention as too costly. In addition, inquiry environments sometimes require unusual and specific technology.

Accompanying these three adoption and implementation dimensions, there are two additional factors that influence the likelihood that an effective intervention will retain its efficacy when implemented on a broader scale:

- **Degree of specification**. The clearer and more detailed the description of a desired new practice, the better educators and education systems know what they're aiming for and the easier it is to measure the presence or absence of the target practice. If educators do not have a clear understanding of what constitutes inquiry or of the practices teachers need to implement, their chances of really implementing the innovation are small. When specification is weak, "lethal mutations" can emerge (Brown & Campione, 1996), where the adaption no longer honors the original vision.
- **Adaptability to fit local capacity, conditions, and practices**. At the same time, as an innovation is tried in more contexts, unforeseen difficulties and tensions with some portions of the ambitious learning activity system are likely to arise. If implementers do not adapt to fit their circumstances, the "replica trap" (Dede, 2005) can occur, because identical materials and teacher moves may be understood and perceived quite differently in different contexts and settings. Most importantly, a clear goal of inquiry learning is for students to experience ownership of an authentic driving question; doing so often involves developing culturally relevant pedagogies (Ladson-Billings, 1995) and this is not always conceptualized as part of the inquiry environment.

There are tensions among these principles that can only be addressed while considering the specifics of a particular inquiry learning approach. One tension is between ambition and complexity.

Inquiry learning is ambitious and yet designers may need to reduce ambition in order not to become overly complex for educators to adopt. Likewise, there is a tension between adaptability and specification. One way it can be partially resolved is for the design team to become more specific about what is adaptable in their approach and what should be changed only with great caution. This is hard to do a priori, and thus design teams typically capture information about variability and then respond. In one example, a research team found some teachers adapted to the pace of a mathematics curriculum by skipping part of every lesson so they could initiate the next lesson on the prescribed pace (Dunn, 2009. This was maladaptive and led to guidance to skip some optional lessons entirely rather than skipping parts of the most important lessons. Another step is often to be clear with those who will make adaptations about what principles are to be honored. In another example from the SimCalc research, when some teachers did not have access to a computer lab on the right day, they were able to keep the student-centered intention of a curriculum by having students plan investigations as a class, and then have one student perform the investigation on a projected computer. This honored the student-centered intent. Another adaptation, where the teacher demonstrated on the computer instead of allowing students to drive the work, did not honor the intent.

Finally, many scaling efforts, such as the Building Blocks (Sarama & Clements, 2013, discussed in more detail below), are framed in terms of equity challenges. Inquiry learning opportunities need to become more common overall, and it is especially important that they become as common in the classrooms of underserved students as they are in classrooms serving predominantly white and higher-income students. Indeed, investigators have found that inquiry activities can be especially beneficial for students in low-income schools (Ben-David & Zohar, 2009). Yet some inquiry learning environments require resources that are less available in schools that serve low-income students. Some of them require teachers with knowledge and skills that are in short supply in the teacher labor force (e.g., computer science skills for computational thinking initiatives) and are inequitably allocated across schools. Or they may call on skills that take extended practice for teachers to acquire, and higher rates of teacher mobility in low-income areas may impede progress. Thus, the issue of **equity and access** adds another layer to the challenge of scaling up an inquiry environment.

Strategies for Scaling Inquiry Environments

In this section, we focus on successful examples of scaling inquiry for approaches that are implemented in schools. We selected six well-known inquiry environments for which there is published evidence with regards to efficacy as well as scholarly reflections on the scaling up process. Our intention in choosing these cases was to illustrate the variety of inquiry environments that have scaled up as well as common issues that were addressed within the scaling efforts. In the scope of this chapter, we did not have space for a comprehensive review of every case.

In each of these cases, the R&D team reflected on its scaling process after they had made significant progress and overcome some major obstacles. Each team explicitly took on the challenge of scaling up themselves, rather than expecting it to happen spontaneously or planning to hand off the scaling process to someone else (e.g., a publisher). The project teams put in "multiple coordinated efforts" to not only "let it happen" but to actually "make it happen" (Sarama & Clements, 2013, p. 176). Each team made scaling up their inquiry learning environment a programmatic feature of their work and organized their leadership to manage this aspect of the work. Each team found the process challenging and sought to learn from their initial experiences and make improvements to improve their ability to scale further.

For each of the six cases, Table 6.1 provides a brief description of the inquiry learning environment as well as evidence for its efficacy and scale.

Table 6.1 Six examples for scaled and efficient inquiry learning environments entailing common inquiry learning features

Project	*Inquiry learning features*	*Content and target group*	*Example efficacy studies*	*Scale metrics*
GLOBE	1 exploration and 3 collaboration	Environmental science (e.g., atmospheric science) Grades K-12	Quasi-experiment (N = 123): EG GLOBE (n = 60 students) vs. CG Non-GLOBE (n = 63 students) EG outperformed CG on hydrology assessment scores with an effect size of 0.10 (Penuel et al., 2005)	In 2020, the project scaled to approximately 37,000 schools, 40,000 teachers, 7000 teacher trainees, and 809,000 students worldwide (see GLOBE Homepage, 2020)
Building Blocks/ TRIAD	1 exploration and 2 visualization	Mathematics Pre K to grade 2	Experiment (N = 25 classrooms, 209 students): EG Building Blocks (N = 13 classrooms) vs. CG Non-Building Block (N = 12 classrooms) EG outperformed CG on Research-based Early Mathematics Assessment scores with an effect size of 0.62 Follow-up experiment with 42 schools, 106 classrooms, and 1375 preschoolers replicated the beneficial effect (g = 0.72) of learning with Building Blocks (Sarama, Clements, Starkey, Klein, & Wakeley, 2008; Clements et al., 2011)	In 2018, the project scaled to approximately 180 Pre-K teachers, 2160 children from MA, Buffalo, NY and Nashville, TN often coming from HeadStart and low-resources schools (see TRIAD Homepage)
LASER	1 exploration and 3 collaboration	STEM disciplines Mainly grades 1–5 but also grades 6–8 and kindergarten	Matched-paired RCT (N = 2601 students): EG LASER (n = 1429 students) vs. CG Non-LASER (1172 students) EG outperformed CG on Partnership for the Assessment of Standards-Based Science performance assessment scores (Smithsonian Science Education Center, 2015)	In 2015, the project scaled to 60,000 students mainly from different U.S. school districts but is also used in other countries such as Mexico, Sweden, and Chile (Smithsonian Science Education Center, 2015; Devés & Lopez, 2007)

SimCalc	1 exploration and 2 visualization	Mathematics (rate and proportionality) Grades 7–8 (originally)	RCT (N = 1621 students): EG SimCalc (n = 796 students) vs. CG Non-SimCalc (n = 825 students) EG outperformed CG rate and proportionality understanding scores with an effect size of 0.63 Follow-up experiments with another 1048 seventh graders and 825 eighth graders replicated the beneficial effect of learning with SimCalc with effect sizes of 0.50 and 0.56 (Roschelle, Shechtman et al., 2010)	The SimCalc approach scaled to 4 regions of Texas, 95 teachers (in 7th grade) and 56 teachers (in 8th grade) and thousands of students with diverse backgrounds (SES levels, ethnicity, region) Later, the SunBay environment scaled to over 25,000 students per year in Florida. The Cornerstone work in the United Kingdom scaled to over 100 schools and 203 teachers (Roschelle, Shechtman et al., 2010; Vahey et al., 2013; Clark-Wilson et al., 2015)
River City	1 exploration and 4 metacognitive learning	Science (infectious disease) K-12	Quasi-experiment (N = 1000 students; 11 teachers): EGs with 2 variants of River City vs. Non-River City CG Both EGs outperformed CG on biology posttest (Clarke et al., 2006)	In 2009, the project scaled to 250 teachers, 15,000 students from the United States and Canada (Clarke & Dede, 2009)
WISE	1 exploration, 2 visualization, 3 collaboration, and 4 metacognitive learning	Science (physics, chemistry, life science, earth science) K-12	Two-time delayed experimental groups (N = 4328 students; 26 teachers) EG WISE vs. CG Non-WISE EG outperformed CG on explanation-based knowledge integration scores with an effect size of 0.32 (Linn et al., 2006)	In 2018, 73 different WISE-projects were listed in the project library; 5 projects are in Dutch and 4 projects are in Spanish (WISE Homepage, 2020)

⋆ in Table 6.1 with examples for successful scale-up projects

Note: EG = experimental group; CG = control group; RCT = randomized controlled trial. When available, sample sizes and effect sizes were reported. References listed in Table 6.1 are marked with a star in the reference list. Inquiry learning features are adopted from Donelly, D. F., Linn, M. C., & Ludvigsen, S. (2014). Impacts and Characteristics of Computer-Based Science Inquiry Learning Environments for Precollege Students, *Review of Educational Research, 84*(4), 572–608.

To illustrate how these environments support inquiry learning, we draw on the four inquiry learning features identified in a review of inquiry learning environments (Donelly et al., 2014). These features are elements that allow students to (1) explore meaningful and authentic scientific contexts, (2) use powerful visualizations, (3) collaborate with others, and (4) develop autonomous, metacognitive learning practices (Donnelly et al., 2014, p. 4). Like most inquiry learning environments, all of the examples in Table 6.1 engage students in exploration and investigation. LASER and GLOBE, in particular, are noteworthy for involving students in working with scientists in authentic scientific investigations. Second, many inquiry environments support students in visualizing concepts or empirical evidence. SimCalc and Building Blocks, our two mathematics examples, introduce technology-supported visualizations. Third, some inquiry environments focus on student collaboration and argumentation; WISE is our example that best illustrates this characteristic. Finally, River City and WISE are examples of inquiry learning environments that emphasize metacognitive learning as well as learning in the field of study.

In all six cases, R&D teams were concerned with establishing efficacy, that is, providing evidence for a causal argument that implementing their approach would improve student learning outcomes. For two of the cases (Building Blocks and SimCalc), researchers conducted randomized controlled trials to establish efficacy. In other cases, an efficacy case was built through a series of design studies, case studies, and quasi-experimental evaluations. See Table 6.1 for more details.

In terms of scale, each of our example cases sought to test inquiry learning across diverse settings beyond the setting in which the approach was first designed (most often, inquiry learning environments are tested in a single context or several similar contexts). Each reached thousands of students. GLOBE and LASER in particular had ambitious scaling goals right from the beginning, even as they were still under development. Both of these inquiry learning environments met their objective of scaling internationally and involved hundreds of thousands of students.

Through our review of each team's reflections on these cases, we identified five useful scaling strategies. We see these as complementary strategies, not alternatives.

Strategy 1: Understand the Context. Teams that succeed in scaling up inquiry environments invest considerable energy in understanding and defining the niche in which they can scale and the needs of the educators who will adopt their approach. Although they may have ambitions for universal adoption, realistically they focus on niches where growth is possible. They define stakeholders in their approach and seek to learn more about what those stakeholders care about, what obstacles they face, and what supports they need. For example, the SimCalc team started with a vision of "simulations for calculus learning" (hence "SimCalc") but later focused on student learning of ratio and proportion in their scaling activities because these topics were a bigger problem for schools than precalculus was. The GLOBE program initially emphasized its data collection protocols and the accuracy of the data students submitted on their local study site. Over time, however, the GLOBE leaders realized that winning time for their environmental inquiry program within the regular school schedule required mapping the curriculum onto the standards for which teachers and schools are held accountable. The program even developed GLOBE books for early readers that teachers in the early elementary grades could use to teach literacy and environmental inquiry concepts at the same time.

Strategy 2: Engage a Breadth of Expertise. Teams that succeed in scaling their innovations incorporate multiple types of talent within their teams specifically to help with problems of scale. This means going beyond the small group that developed the initial design concept. For example, the LASER team included experts in curriculum, assessment, professional development, administrative and community support, and materials delivery. The Building Blocks team (see Sarama et al., 2008) focused on developing strong relationships with stakeholders in their implementation sites and invited input and feedback from stakeholders. The SimCalc program incrementally added experts with additional expertise as the range of concerns to be addressed

expanded (Roschelle et al., 2008). In general, a design-based implementation research approach—which emphasizes inclusion of practitioners in decision-making and focuses specifically on the layers of additional design needed to support implementation—becomes highly relevant in scaling up inquiry approaches (Penuel et al., 2011).

Strategy 3: Develop a Coherent Learning Activity System. Teams that succeed in bringing inquiry learning innovations to scale integrate the many different elements needed to support the desired change in teaching and learning into an ambitious learning activity system (defined above). For example, they take care to pull together and align the curriculum, technology, professional development, and assessment components of their approach. In the Scaling Up SimCalc project (Roschelle et al., 2010), this entailed writing replacement curriculum units that were specific about how and when dynamic representations on a computer were to be used, and interconnecting the technology activities with non-technology activities. Furthermore, teacher professional development was very tightly synchronized to what teachers would need to know to use the curriculum workbooks and technology together. More broadly, we noted that successful scale-up designs carefully consider what educators need at different stages of experience with the innovation—for example, to make the decision to adopt the approach, to learn how to initially use it, to become expert in their use of the approach, and eventually to sustain it on their own and help others learn to use it. Overall, teams that succeed at scaling view their work as building systems for teaching and learning, not just disseminating an isolated tool or material, and they focus on a clear image of the teaching and learning activity that every aspect of the system will work to support (Clements et al., 2011; Sarama et al., 2008).

Strategy 4: Work with Practitioners to Improve Implementation. Teams that scale their innovations successfully use design methods that invite participation of practitioners early on and throughout the design and scaling process. These methods include co-design, design-based implementation research, and researcher-practitioner partnerships. One especially important focus for co-design is on the teacher professional development that will be needed to support scaling up an inquiry learning environment. A recent meta-analysis (Lynch et al., 2019) is a good starting point for considering the nature of effective teacher professional development for scaling STEM inquiry approaches. Reviews and syntheses of best practices in STEM teacher professional development for inquiry also offer guidance (i.e., Capps, Crawford, & Constas, 2012; Gerard et al., 2011; Lederman & Lederman, 2012; Wilson, 2013). All projects acknowledge that designing an inquiry environment for scale requires allowing for its adaptation to fit different contexts and its expandability to additional content and needs (Clark & Dede, 2009). For example, to promote depth of scaling, the River City team "employs design-based research methods in order to understand what conditions are more flexible and adaptable to meet needs of students and teachers in various conditions" (Clark & Dede, 2009, p. 358).

Strategy 5: Measure and Iterate. To meet the scaling criterion of "predictable, measurable impact at scale," teams develop measures they can use to monitor their progress. This often includes designing student learning measures that fit the intention of their inquiry environment, as most large-scale assessments fail to measure what students learn from inquiry environments. Effort is often made to show the relationships between the newly designed measures and aspects of curricular standards or frameworks that are important. In some cases, studies seek to measure both more aligned and accountability-oriented measures. For example, the Scaling Up SimCalc program (Roschelle, Shechtman et al., 2010) included two subscales, one of which was better aligned to the curriculum's inquiry goals and the other which used relevant items from the state accountability assessment. Teams also develop indicators of the ease of use of their materials, ways of monitoring how frequently and in what ways their various tools are used, and likelihood of continued usage. Although most developers of inquiry learning environments have leaned toward the idea of adaptation of their system to local conditions as opposed to strict "fidelity of implementation"

(e.g., Dede, 2005), successful scaling efforts nonetheless develop ways to detect inappropriate or weak uses of their system and to help implementers improve. All the projects in Table 6.1 describe the scaling up process as highly iterative, with many cycles covering a wide range of issues that arise during implementations. The teams built systems and practices to collect examples of implementation issues that they used to plan future versions of their learning environments.

Partnerships for Sustaining Inquiry Environments

The six cases in Table 6.1 were selected in part because both evidence of impact and scholarly reflections on the scaling process were available. Application of these criteria ruled out many successful cases of scaling involving partnerships where the scaling work was performed by a company. A reason to also focus attention on partnerships is their relationship to sustainability.

Scale and sustainability are interrelated, but they are not the same. Scale is spreading a practice; sustainability is keeping it in use for a longer period of time. There are economic dimensions of both scale and sustainability. Programs that are very expensive can have a hard time attracting initial adoptions. Economics comes into play with sustainability because it involves recurrent costs as well as initial costs. For an innovation to be sustained, a mechanism for covering recurring costs must be identified. These costs may be easily measured direct costs (such as annual licensing fees) or may be more subtle and even nonmonetary in nature—such as the costs of maintaining a pool of talented teachers who can enact the innovation or the need to refresh and refine the innovation on a regular basis in response to changing circumstances or even just the cost of keeping the focus on continuing to implement the innovation in the face of other shiny, new objects promoted by others.

One way to address sustainability is through partnerships with business. Businesses are structured to sustain their offerings in a market. We offer some examples, but then turn to other means of sustainability. Read180 is one well-known example where the scale and sustainability of an ambitious learning activity system was led by a company. Read180 is a reading program by the company HMH which configures the classroom as a series of stations in which different modes of reading activity occur. Research conducted by Ted Hasselbring and Laura Goin (e.g., Hasselbring & Goin, 1988) at Vanderbilt University provided the underpinnings for the initial design of Read180. Another example of scale and sustainability is Carnegie Learning, a company that scaled up intelligent tutoring technologies developed in partnership with researchers at Carnegie Mellon University (Ritter et al., 2007).

Two additional examples that have achieved remarkable scale and sustainability highlight the parallel and coordinated contributions from researchers and companies. The prominence of graphing calculators arose from separate but related efforts of Texas Instruments (a company) and university professors. Ohio State professors Frank Demana and Bert Waits were integral to the drive for calculator adoption and use in mathematics classroom (e.g., Waits & Demana, 1998). Furthermore, Demana and Waits developed a large teacher professional development network that was independent of, yet closely affiliated with, Texas Instruments. Texas Instruments often shaped their product development roadmap in response to suggestions from this network. Subsequently, independent researchers analyzed the impact of graphing calculator use; for example, a meta-analysis by Ellington (2003) found graphing calculators were effective for developing conceptual understanding (this is likely because teachers use calculators to free students from doing tedious calculations and thus can focus more on concepts). Graphing calculators were first developed in the 1990s and remain prominent in mathematics and science classrooms 20 years later.

In a closely related example, probes and sensors were developed through separate but interrelated efforts of researchers and several small companies. Probes and sensors are used to foster hands-on inquiry instruction (Soloway et al., 1999). Early research on microcomputer-based labs

(e.g., Mokros & Tinker, 1987) and subsequent research to further develop probes and sensors were closely related to long-standing efforts at many companies to develop commercially viable probes and sensors, resulting in wide availability of low-cost technologies. The Concord Consortium nurtured a symbiotic connection between researchers and industry to continue advancing the scalable technologies along with related science inquiry research.

We would also caution that such success stories involve partnerships and collaboration that span decades of back-and-forth dialogue on an educational problem and approach. Despite the involvement of a commercial entity, they are not well described by terms such as "transferring" or "commercializing" a research-based discovery or invention. Furthermore, a partnership like this is not necessarily the only route available to innovators who would like to scale an inquiry learning environment. In some cases, a university-affiliated group itself becomes the long-term engine scaling an innovation, though usually outside the tenure track demands of a university department. Such is the case with FOSS (Powell & Wells, 2002), a hands-on inquiry activity system that has been sustained by the Lawrence Hall of Science, which is affiliated with the University of California, Berkeley. It is also possible that the thrust of an inquiry-based approach may be sustained as a school of thought and thus by a person, team, or institution providing intellectual thought leadership. For example, the Exploratorium, an informal science institution in San Francisco, could be said to have had this effect with its "to do and notice" approach to engaging wonder and investigation through physical activities (Oppenheimer & Cole, 1974). Likewise, sustainability may be created institutionally, such as when inquiry-based learning environments or approaches become the basis of policies that are adopted by school systems. In such cases, the exact tool or environment may not be sustained, but the core features of the inquiry approach may be. One might see the growing adoption of maker spaces by schools in this light; they create a dedicated school space where inquiry practices might be sustained.

With regard to sustainability, there are also limitations to what can be learned from studying existing cases. Leaders who develop inquiry learning environments may find the niche for inquiry learning within schools may be too small, the possible adopters too hard to reach, or the available financing too little. Faced with these challenges, the team may simply move on to a new research topic. We have noted that public-private partnerships can sometimes overcome these barriers, but few scholarly reflections on the nature of educational public-private partnerships are available. The other key feature we identified in this context was the shift of ownership, where institutions different from those who developed an inquiry learning innovation take on the life of an innovation. The thought-leadership approach requires new owners who take on and sustain the thoughts. The institutional space approach requires new owners who independently figure out how to cover the initial and recurrent costs of the new space. As ownership shifts either to partnerships or to new institutions come to the fore, issues of maintaining the quality of the innovation continually come to the fore—the innovation may become less ambitious, mutate fatally to become something different, or become less identifiable or prominent. Sustainability of ambitious learning activity systems remains a "wicked problem," not a well-mapped challenge.

Discussion: The State of the Art and Remaining Challenges

Bringing inquiry learning environments to scale is an important issue for society, especially given the needs for stronger inquiry skills among future citizens, employees, and leaders. Scaling up is a complex challenge for any educational innovation, and we have argued particularly so for ambitious inquiry learning innovations that may not find a good fit with prevailing priorities in many of today's classrooms and communities. Nonetheless, it can be done: We have described six examples of inquiry learning environments that achieved considerable scale and four additional long-term partnerships. To tackle the challenges of scaling, the example projects planned

for scaling from the earliest stages of their work. They invested in scaling up for a long period of time, and their approach evolved to incorporate insights gained through their experience in the field. They also reflected on which principles helped them reach scale on many different dimensions.

The principles presented above are interwoven, and implementing them is labor-intensive. Overcoming the many obstacles to realizing a vision is a long-term commitment best undertaken with a highly dedicated team. Although there is still much to be learned about how to work effectively on scaling up an inquiry learning environment, due to an expansion of research on scaling education innovations in the past two decades, there are now proof points that it is possible and there is much less mystery about how to do it.

There are also limitations to what can be learned from studying existing cases. Sustainability after R&D funding is exhausted and in the face of changing education priorities and staffing remains a major challenge. Inquiry projects can achieve scale, and yet the learning environment may not continue to spread or keep going after the funding ends. The niche for inquiry learning within schools may be too small, the available financing too little, or the team may simply move on to a new research topic. We have noted that public-private partnerships can sometimes overcome these barriers, but few scholarly reflections on the nature of educational public-private partnerships are available. Other models are available as well, as noted above, but the paths to sustainability can appear to be idiosyncratic to the personalities of the individuals involved. Further insights on how inquiry teams could achieve sustainability are very much needed.

A second enduring challenge is addressing equity. We found that few scale-up examples were as specific as we would like about the degree to which they overcame the pervasive equity issues. Absent intentional strategies to counteract preconceptions about who can and should engage in inquiry learning, inquiry learning environments may scale primarily to classrooms that are already doing student-centered instruction. For the most part, scale has been achieved by what Cohen and Mehta characterize as "niche reforms." This type of reform fits a place within the educational system but does not challenge the system as a whole. For those aspiring to make inquiry learning the centerpiece of systemic reform—to design entire school systems for the purpose of fostering inquiry among all students (see, for example, Collins, 2017)—more research is needed.

A third challenge regards incentivizing researchers to focus on scaling. Given how slow and hard scaling work is, working on this issue may detract from building the kind of publication track record prized by universities. Scaling work tends to force one to become a generalist, because of the range of problems one encounters—and this too runs counter to academic career rewards for specialization. Scaling efforts may not fit the mission of a department, lab, or institution. And scaling requires patience, because the costs and problems arrive early, and the benefits and successes materialize more slowly.

Future research and funding initiatives related to scaling inquiry learning environments should pay attention to the limitations noted above to what has been achieved thus far. The field has much to learn about how to achieve sustainability, equity, and aligned incentive systems for the implementation of inquiry learning systems at scale.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grants 2021159 (CIRCLS) and 1837463 (CIRCL). 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

Aivaloglou, E., & Hermans, F. (2016). How kids code and how we know: An exploratory study on the Scratch repository. In *ICER '16 Proceedings of the 2016 ACM Conference on International Computing Education Research* (pp. 53–61). New York: ACM. http://doi.org/10.1145/2960310.2960325

Anderson, R., & Helms, J. (2001). The ideal of standards and the reality of schools: Needed research. *Journal of Research in Science Teaching*, *38*, 3–16. http://doi.org/10.1002/1098–2736(200101)38:1<3::AID-TEA2>3.0.CO;2-V

Ben-David. A., & Zohar, A. (2009). Contribution of meta-strategic knowledge to scientific inquiry learning. *International Journal of Science Education*, *31*(12), 1657–1682. http://doi.org/10.1080/09500690802162762

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah, NJ: Erlbaum.

Capps, D. K., Crawford, B. A., & Constas, M. A. (2012). A review of empirical literature on inquiry professional development: Alignment with best practices and a critique of the findings. *Journal of Science Teacher Education*, *23*(3), 291–318. https://doi.org/10.1007/s10972-012-9275-2

Cartwright, N. (2012). Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science*, *79*(5), 973–989. https://doi.org/10.1086/668041

Cartwright, N. (2013). Knowing what we are talking about: Why evidence doesn't always travel. *Evidence & Policy: A Journal of Research, Debate and Practice*, *9*(1), 97–112. https://doi.org/10.1332/174426413X662581

★Clarke, J., & Dede, C. (2009). Design for scalability: A case study of the River City curriculum. *Journal of Science Education and Technology*, *18*, 353–365. https://doi.org/10.1007/s10956-009-9156-4

★Clarke, J., Dede, C., Ketelhut, D. J., Nelson, B., & Bowman, C. (2006). A design-based research strategy to promote scalability for educational innovations. *Educational Technology*, *46*(3), 27–36. [Faculty Publication]

★Clark-Wilson, A., Hoyles, C., Noss, R., Vahey, P., & Roschelle, J. (2015). Scaling a technology-based innovation: Windows on the evolution of mathematics teachers' practice. *ZDM Mathematics Education*, *47*, 79–92. https://doi.org/10.1007/s11858-014-0635-6

★Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large scale cluster randomized trial. *Journal of Research in Mathematics Education*, *2*, 127–166. https://doi.org/10.5951/jresematheduc.42.2.0127

Cobb, P., Jackson, K., Smith, T., Sorum, M., & Henrick, E. (2013). Design research with educational systems: Investigating and supporting improvements in the quality of mathematics teaching and learning at scale. In B. J. Fishman, W. R. Penuel, A.-R. Allen & B. Haugan Cheng (Eds.), *Design-based implementation research: Theories, methods and exemplars, Yearbook of the National Society for the Study of Education*, *112*(2), 320–349. New York: Teachers College, Columbia University.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3–12. https://doi.org/10.3102/0013189X032006003

Coburn, C. E., & Penuel, W. R. (2016). Research-practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, *45*(1), 48–54. https://doi.org/10.3102/0013189X16631750

Coburn, C. E., Penuel, W. R., & Geil, K. (2013). *Research-practice partnerships at the district level: A new strategy for leveraging research for educational improvement*. New York: Wiilliam T. Grant Foundation.

Cohen, D. K., & Mehta, J. D. (2017). Why Reform Sometimes Succeeds: Understanding the Conditions That Produce Reforms That Last, *American Educational Research Journal, 54*(4), 644–690, https://doi.org/10.3102/0002831217700078

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142. https://doi.org/10.3102/01623737025002119

Collins, A. (2017). *What's worth teaching? Rethinking curriculum in the age of technology*. New York: Teachers College Press.

Darling-Hammond, L., Hyler, M. E., Gardner, M., & Espinoza, D. (2017). *Effective teacher professional development*. Palo Alto, CA: Learning Policy Institute.

Dearing, J. W., Dede, C., Boisvert, D., Carrese, J., Clement, L., Craft, E. et al. (2015). How educational innovators apply diffusion and scale-up concepts. In C. K. Looi & L. W. Teh (Eds.), *Scaling educational innovations* (pp. 81–104). Singapore: Springer. https://doi.org/10.1007/978-981-287-537-2_5

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2016). Building an assessment argument to design and use next generation science assessments in efficacy studies of curriculum interventions. *American Journal of Evaluation*, *37*(2), 174–192. https://doi.org/10.1177/1098214015581707

Dede, C. (2005). Scaling up: Evolving innovations beyond ideal settings to challenging contexts of practice. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511816833.034

★Devés, R., & López, P. (2007). Inquiry-based science education and its impact on school improvement: The ECBI program in Chile. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement*. Springer International Handbooks of Education (vol. 17). Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-5747-2_48

Donelly, D. F., Linn, M. C., & Ludvigsen, S. (2014). Impacts and characteristics of computer-based science inquiry learning environments for precollege students. *Review of Educational Research*, *84*(4), 572–608. https://doi.org/10.3102/0034654314546954

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

Educational Endowment Foundation. (n.d.). Retrieved from https://educationendowmentfoundation.org.uk/

Ellington, A. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education*, *34*(5), 433–463. https://doi.org/10.2307/30034795

First Lego League. (n.d.). Retrieved from https://www.first-lego-league.org/en/general/what-is-fll.html

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., …, & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention science, 6(3),* 151–175. https://doi.org/10.1007/s11121-005-5553-y

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*(3), 300–329. https://doi.org/10.3102/0034654312457206

Gerard, L. F., Varma, K., Corliss, S. B., & Linn, M. C. (2011). Professional development for technology-enhanced inquiry science. *Review of Educational Research*, *81*(3), 408–448. https://doi.org/10.3102/0034654311415121

★Globe Home Page. (2020, October 13). Retrieved from https://www.globe.gov/de/about/impact-and-metrics

Hasselbring, T. S., & Goin, L. I. (1988). Microcomputer applications to instruction. In E. A. Polloway, J. S. Payne, J. R. Patton, & R. A. Payne (Eds.), *Strategies for teaching retarded students* (4th ed.). Columbus, OH: Charles E. Merrill Publishing Company.

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*(1), 1–21. https://doi.org/10.1080/19345747.2017.1375583

Institute for Educational Sciences and National Science Foundation. (2013). *Common guidelines for educational research and development*. Washington, DC: US Department of Education and National Science Foundation. Retrieved January 30, 2019 from https://ies.ed.gov/pdf/CommonGuidelines.pdf

Interagency Educational Research Initiative (IERI). (n.d.). Retrieved from https://drdc.uchicago.edu/community/main.php

Kerckhoff, A. C., & Glennie, E. (1999). The Matthew effect in American education. *Research in Sociology of Education and Socialization*, *12*(1), 35–66.

Krumm, A., Means, B., & Bienkowski, M. (2018). *Learning analytics goes to school*. New York: Routledge. https://doi.org/10.4324/9781315650722

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465–491. https://doi.org/10.3102/00028312032003465

Lederman, N. G., & Lederman, J. S. (2012). Nature of scientific knowledge and scientific inquiry: Building instructional capacity through professional development. In *Second international handbook of science education* (pp. 335–359). Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-9041-7_24

LeMahieu, P., Grunow, A., Baker, L., Nordstrum, L., & Gomez, L. (2017). Networked improvement communities. *Quality Assurance in Education*, *25*(1), 5–25. https://doi.org/10.1108/QAE-12-2016-0084

Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. *Journal of Research on Educational Effectiveness*, *8*(3), 400–418. https://doi.org/10.1080/19345747.2014.915604

Lewis, S. (2015). Qualitative inquiry and research design: Choosing among five approaches. *Health Promotion Practice*, *16*(4), 473–475. https://doi.org/10.1177/1524839915580941

★Linn, M. C., Lee, H.-S., Tinker, R., Husic, F., & Chiu, J. L. (2006). Teaching and assessing knowledge integration in science. *Science*, *313*(5790), 1049–1050.

Looi, C.-K., & Woon, T. L. (2015). *Scaling educational innovations*. Singapore: Heidelberg.

Lynch, K., Hill, H., Gonzalez, K., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/0162373719849044

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

McDonald, S.-K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher*, *35*(3), 15–24. https://doi.org/10.3102/0013189X035003015

Melchior, A. L., Burack, C., Hoover, M., & Marcus, J. (2016). *FIRST longitudinal study: Participant characteristics, program experience, and impacts at follow-up (year 3 report)*. Waltham, MA: Center for Youth and Communities, Heller School, Brandeis University. Prepared for US FIRST.

Mokros, J. R., & Tinker, R. F. (1987). The impact of microcomputer-based labs on children's ability to interpret graphs. *Journal of Research in Science Teaching*, *24*(4), 369–383. https://doi.org/10.1002/tea.3660240408

Moore, G. (2014). *Crossing the Chasm. Marketing and selling disruptive products to mainstream customers* (3rd ed.). New York: Harper Collins Business.

Mort, P. R. (1953). Educational adaptability. *The School Executive*, *71*, 1–23. https://doi.org/10.1111/j.2044-8279.1953.tb02842.x

Münch, J., Fagerholm, F., Johnson, P., Pirttilahti, J., Torkkel, J., & Järvinen, J. (2013). Creating minimum viable products in industry-academia collaborations. In B. Fitzgerald, K. Conboy, K. Power, R. Valerdi, L. Morgan, & K. J. Stol (Eds.), *Lean enterprise software and systems. LESS 2013. Lecture Notes in Business Information Processing* (vol. 167). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-44930-7_9

Newman, D., Jaciw, A. P., & Lazarev, V. (2018). *Guidelines for conducting and reporting edtech impact research in U.S. K-12 schools*. Washington, DC: Educational Technology Network of Software and Information Industry Association. Retrieved from https://www.empiricaleducation.com/pdfs/guidelines.pdf (Accessed November 7, 2019).

Nicholson, T. W., & Tunmer, W. E. (2010). Reading: The great debate. In C. M. Rubie-Davies (Ed.), *Educational psychology: Concepts, research and challenges* (pp. 50–64). London: Routledge.

Oppenheimer, F. J., & Cole, K. C. (1974). The Exploratorium: A participatory museum. *Prospects*, *4*(1), 21–34. https://doi.org/10.1007/BF02206525

★Penuel, W. R., Bienkowski, M., Korbak, C., Molina, A., Russo, D., Shear, L., Toyama, Y., & Yarnall, L. (2005). Globe year 9 evaluation: Implementation supports and student outcomes. SRI International. https://www.globe.gov/documents/10157/6afb66ed-24de-4d2e-800f-ba26abf7028c

Penuel, W. R., Fishman, B. J., Cheng, B., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation and design. *Educational Researcher*, *40*(7), 331–337. https://doi.org/10.3102/0013189X11421826.

Penuel, W. R., McWilliams, H., McAuliffe, C., Benbow, A., Mably, C., & Hayden, M. (2009). Teaching for understanding in Earth science: Comparing impacts on planning and instruction in three professional development designs for middle school science. *Journal of Science Teacher Education*, *20*(5), 415–436. https://doi.org/10.1007/s10972-008-9120-9

Powell, K., & Wells, M. (2002). The effectiveness of three experiential teaching approaches on student science learning in fifth-grade public school classrooms. *The Journal of Environmental Education*, *33*(2), 33–38. https://doi.org/10.3102/0013189X11421826

Prewitt, K., Schwandt, T. A., & Straf, M. L. (Hrsg.). (2012). *Using science as evidence in public policy*. Washington, DC: National Academies Press.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255. https://doi.org/10.3758/BF03194060

Roehrig, G. H., & Luft, J. A. (2004). Constraints experienced by beginning secondary science teachers in implementing scientific inquiry lessons. *International Journal of Science Education*, *26*(1), 3–24. https://doi.org/10.1080/0950069022000070261

Rogers, E. M. (1962). *Diffusion of innovations* (3rd ed.). New York: A Division of Macmillan Publishing.

Roschelle, J., Knudsen, J., & Hegedus, S. (2010). From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning. In M. J. Jacobson & P. Reimann (Eds.), *Designs for learning environments of the future: International perspectives from the learning sciences* (pp. 233–262). New York: Springer. https://doi.org/10.1007/978-0-387-88279-6_9

★Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*(4), 833–878. https://doi.org/10.3102/0002831210367426

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

Roschelle, J., Tipton, E., Shechtman, N., & Vahey, P. (2018). *Generalizability of a technology-based intervention to enhance conceptual understanding in mathematics* (SimCalc Technical Report 10). Menlo Park, CA: SRI International. https://doi.org/10.13140/RG.2.2.31260.13442

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393. https://doi.org/10.1002/tea.10027

Russ, R. S., & Berland, L. K. (2019) Invented science: A framework for discussing a persistent problem of practice. *Journal of the Learning Sciences*, *28*(3), 279–301. https://doi.org/10.1080/10508406.2018.1517354

Sallis, E. (2002). *Total quality management in education* (3rd ed.). London: Taylor & Francis.

Sarama, J., & Clements, D. H. (2013). Lessons learned in the implementation of the TRIAD scale-up model. Teaching early mathematics with trajectories and technologies. In T. G. Halle, A. J. Metz & I. Martinez Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 173–191). Baltimore, MD: Brookes.

★Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *1*, 89–119. https://doi.org/10.1080/19345740801941332

Scalise, K., & Gifford, B. (2006). Computer-based assessment in E-learning: A framework for constructing "Intermediate Constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, *4*(6). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1653 (Accessed May 15, 2019).

Schneider, B., & McDonald, S. K. (Eds.). (2007a). *Scale-up in education: Vol. 1: Ideas in principle*. Lanham, MD: Rowman & Littlefield.

Schneider, B., & McDonald, S. K. (Eds.). (2007b). *Scale-up in education: Vol. 2: Issues in practice*. Lanham, MD: Rowman & Littlefield.

Schoenfeld, A. H., & Pearson, P. D. (2012). The reading and math wars. In G. Sykes, B. Schneider & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 576–596). London: Routledge.

Scratch Computing Environment. (n.d.). Retrieved from https://scratch.mit.edu/

★Smithsonian Science Education Center. (2015). *The LASER model: A systemic and sustainable approach for achieving high standards in science education. Executive summary.* Washington, DC: Smithsonian Institution. Retrieved from https://ssec.si.edu/laser-i3

Soloway, E., Grant, W., Tinker, R., Roschelle, J., Mills, M., Resnick, M., Berg, R., & Eisenberg, M. (1999). Science in the palm of their hands. *Communications of the ACM*, *42*(8), 21–26. https://doi.org/10.1145/310930.310953

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501. https://doi.org/10.3102/1076998614558486

★Vahey, P., Knudsen, J., Rafanan, K., & Lara-Meloy, T. (2013). Curricular activity systems supporting the use of dynamic representations to foster students' deep understanding of mathematics. In *Emerging technologies for the classroom* (pp. 15–30). New York: Springer. https://doi.org/10.1007/978-1-4614-4696-5_2

Van der Valk, T., & de Jong, O. (2009). Scaffolding science teacher in open-inquiry teaching. *International Journal of Science Education*, *31*(6), 829–850. https://doi.org/10.1080/09500690802287155

Von Hippel, E. (2005). *Democratizing innovation*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/2333.001.0001

Waits, F., & Demana, B. K. (1998). The role of graphing calculators in mathematics reform. Retrieved from https://files.eric.ed.gov/fulltext/ED458108.pdf (Accessed January 31, 2019).

Wilson, S. M. (2013). Professional development for science teachers. *Science*, *340*(6130), 310–313. https://doi.org/10.1126/science.1230725

★WISE Homepage. (2020, October, 13). Retrieved from https://wise.berkeley.edu/features