JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 49, NO. 3, PP. 281–304 (2012)

Position Paper

Large-Scale Science Education Intervention Research We Can Use

William R. Penuel[1] and Barry J. Fishman[2]

[1]*University of Colorado, Boulder, Colorado*
[2]*University of Michigan, Ann Arbor, Michigan*

Received 27 May 2011; Accepted 10 December 2011

Abstract: This article develops an argument that the type of intervention research most useful for improving science teaching and learning and leading to scalable interventions includes both research to develop and gather evidence of the efficacy of innovations and a different kind of research, *design-based implementation research* (DBIR). DBIR in education focuses on what is required to bring interventions and knowledge about learning to all students, wherever they might engage in science learning. This research focuses on implementation, both in the development and initial testing of interventions and in the scaling up process. In contrast to traditional intervention research that focuses principally on one level of educational systems, DBIR designs and tests interventions that cross levels and settings of learning, with the aim of investigating and improving the effective implementation of interventions. The article concludes by outlining four areas of DBIR that may improve the likelihood that new standards for science education will achieve their intended purpose of establishing an effective, equitable, and coherent system of opportunities for science learning in the United States. © 2012 Wiley Periodicals, Inc. J Res Sci Teach 49: 281–304, 2012
Keywords: evaluation and theory; policy; science education; standards; implementation

In the past decade there has been strong support for conducting research on interventions at scale in science education. Funding streams at the National Science Foundation and the U.S. Department of Education have supported dozens of large-scale experimental evaluations of school curricula and programs of professional development aimed at improving elementary, middle, and high school science teaching and learning. Federal agencies and private foundations have also supported quasi-experimental and experimental studies of programs that employ public media or that have been implemented in informal settings. Other private foundations (e.g., the William T. Grant Foundation), professional societies (e.g., the Society for Research on Educational Effectiveness), and intermediary organizations (e.g., the Data Research and Development Center) have sought to build the field's capacity for conducting such research. The commitment to developing an understanding of what works at scale has perhaps never been stronger than it is today.

Much of this emphasis is due to policy changes implemented as part of *No Child Left Behind*. That law required that states, districts, and schools that receive federal funding identify and choose programs that have evidence of results from "scientifically based research." Scientifically based research, according to *NCLB*, "relies on measurements or observational

Correspondence to: W.R. Penuel; E-mail: william.penuel@colorado.edu
DOI 10.1002/tea.21001
Published online 13 January 2012 in Wiley Online Library (wileyonlinelibrary.com).

methods that provide reliable and valid data across evaluators and observers,'' and uses ''experimental or quasi-experimental designs'' (PL 107–110, pp. 126–127). To support implementation of *NCLB*, the law called for the establishment of the Institute of Education Sciences (IES) at the U.S. Department of Education, which has since funded a number of large-scale research studies aimed at identifying effective interventions. The pathway to identifying effective interventions requires a long time frame and employment of range of methodologies, including design-based research studies in classrooms and small-scale field tests to establish the feasibility of implementing interventions in multiple settings (Sloane, 2008). At the same time, the policy strongly emphasizes a trajectory that concludes with gathering evidence of effectiveness from large-scale, randomized controlled trials, and has pushed all researchers to re-think how to make arguments about the external validity or *meaning* of the research we engage in for the broader transformational purposes we hope the field of science education research serves.

The claim developed in this article is that the kind of large-scale intervention research needed to improve science teaching and learning and scale includes both scientifically based research on innovations and a different kind of research, which we call *design-based implementation research* (DBIR; Penuel, Fishman, Cheng, & Sabelli, 2011). DBIR is an emerging form of design research that supports the productive adaptation of programs as they go to scale. As an iterative approach to developing innovations, design research is particularly well suited to informing decision-making about needed adjustments to programs (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). DBIR represents a significant expansion of design research—which typically focuses on classrooms—because the focus is on developing and testing innovations that can improve the quality and equity of supports for implementation of reforms. DBIR complements large-scale efficacy research, in that it seeks to support the development of usable, efficacious interventions in science education and to support implementation of interventions found in efficacy studies to have potential for improving teaching and learning.

This research focuses on what is required to bring interventions and knowledge about learning from research into practice for all students, wherever they might engage in science learning. Where effectiveness studies attempt to estimate average treatment effects across a range of settings under ''typical'' conditions, DBIR instead aims to create conditions more conducive to implementing interventions, especially those that have been found to be efficacious under certain conditions. While typical implementation research seeks to analyze and explain patterns of implementation, DBIR researchers use analyses of implementation to iteratively refine strategies for improving the implementation effectiveness of interventions. DBIR ideally produces more scalable designs and a deeper understanding of the contexts of science education, particularly how these contexts arrangements produce patterns of educational outcomes we observe today and the patterns we hope to produce tomorrow.

The need for DBIR arises specifically from concerns about equity of access to powerful learning opportunities. Main effects studies that identify interventions in science education that *can* work often tell us very little about what we need to do in order to make interventions work for particular groups of students and in particular settings. Making interventions work in diverse settings requires local actors to make adaptations that make student diversity a resource for learning while preserving the integrity of science learning goals (Lee, 2002; Rosebery, Ogonowski, DiSchino, & Warren, 2010). Diversity here refers not simply to students' demographics, although these are important; diversity encompasses the ideas, experiences, and histories that enliven and enrich fixed views of what science learning can and should be. In addition, many of the ''typical'' conditions in education that are

presumed in effectiveness studies themselves help to reproduce inequity, by requiring access to resources or support that are not achievable in many settings, especially lower-SES school settings. High teacher turnover, limited access to high quality instructional materials, and accountability pressures disproportionately affect students in schools with high percentages of low-income students and students of color (Ingersoll, 2001; Jacob, 2007; Oakes, 1990; Scafidi, Sjoquist, & Stinebrickner, 2007). Thus, addressing the needs of students from nondominant communities requires research on strategies for supporting effective adaptation of interventions to relate everyday and scientific forms of meaning making, as well as research on strategies that attempt to create more equitable conditions for all students to learn. For science education, such research is both timely and critical, because the field is at present undergoing a major change in the standards that will be used to organize curriculum, assessments, teaching, and professional development for the coming decade.

Strengths and Limitations of Effectiveness Studies as a Form of Large-Scale Intervention Research

A key aim of effectiveness research is to identify interventions that can work in a wide variety of settings (Flay et al., 2005). In science education, a number of studies in the past decade have analyzed the effectiveness of programs when implemented across a large number and wide variety of settings (Borman, Gamoran, & Bowdon, 2008; Buckley et al., 2004; Lee, Maerten-Rivera, Penfield, LeRoy, & Secada, 2008; Penuel, Gallagher, & Moorthy, 2011; Rethinam, Pyke, & Lynch, 2008; Songer, Kelcey, & Gotwals, 2009; Vanosdall, Klentschy, Hedges, & Weisbaum, 2007). Many of these interventions are grounded in decades of basic research on learning and are intended to instantiate principles derived from that research for organizing coherent sequences of instruction for students (Pea & Collins, 2008).

Effectiveness studies have the potential to help policymakers, state officials, and district leaders make decisions based on evidence regarding how to allocate scarce resources for improving teaching and learning (Dynarski, 2008). For decision makers, an intervention that is easily implemented and has robust, consistent effects across a variety of contexts is desirable, because policies and resources can be marshaled more readily to support a single initiative than multiple, conflicting ones (Rowan, 2002). For researchers, effectiveness studies are beneficial, in that they can help identify what interventions best instantiate principles from basic research on learning that also have potential for broad impact, especially when different treatments or interventions are compared to one another.

In practice, however, treatment effects vary widely from setting to setting, and reliable implementation of new programs that seek to transform teaching and learning is difficult to achieve. In a number of studies that have been conducted in the past decade, researchers found no significant positive average treatment effect, but in some classrooms students learned more than others. Teachers in programs have reported significant barriers to implementation: teachers did not enact all activities in the units, especially when the topics received less coverage in state standards (Borman et al., 2008; Penuel et al., 2011), curricular aims were misaligned with assessments (Borman et al., 2008), and teachers experienced difficulties implementing teaching strategies linked to specific curricula (Lynch, Szesze, Pyke, & Kuipers, 2007). Moreover, in at least one longitudinal study, teacher turnover and organizational churn were major causes of attrition that diminished treatment effects (Shear & Penuel, 2010).

Although researchers conducting effectiveness studies do analyze such barriers, we rarely take aim at *changing* the conditions of the kind that inhibit implementation effectiveness.

Indeed, the very logic of the effectiveness trial is to estimate treatment effects under typical conditions. The focus within the field has instead largely been on improving treatment integrity by increasing the fidelity of implementation, as a means to increase treatment effectiveness (e.g., O'Donnell, 2008). By improving treatment integrity, researchers believe, the gap between the potential and realized treatment effectiveness can be reduced (Cordray & Pion, 2006). But this strategy leaves unexamined the conditions and unintended effects of intervention that produce unequal access to high-quality curriculum materials and teaching and that make it less likely, for example, that low-income students of color will have new curriculum materials or a teacher who employs reform teaching strategies, such as inquiry-oriented instruction (Smith et al., 2007; Supovitz & Turner, 2000). Our view is that these conditions are likely to shape and partly explain variations in implementation and that we can learn from attempts to understand this variation where we might need to intervene to improve those conditions (Bryk, 2009).

In addition, an overemphasis on fidelity means less giving less consideration to the ways that curriculum developers and professional development leaders could focus their efforts on helping teachers make productive adaptations of materials by being responsive to students. Teaching that takes into account what learners bring to the classroom, both in terms of prior knowledge and relevant experience, is necessarily adaptive and contingent upon student contributions (X. Lin, Schwartz, & Hatano, 2005). Adaptivity is particularly beneficial when teaching students from nondominant communities, who benefit most when teachers are able to leverage students' repertoires for participation in cultural practices (Gutiérrez & Rogoff, 2003) in such a way as to bridge everyday and scientific ways of meaning making (Barton & Tan, 2009; Carlone, Haun-Frank, & Webb, 2011; Gee & Clinton, 2000; Hudicourt-Barnes, 2003; Rosebery et al., 2010; Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001).

Finally, in the predominant model of large-scale intervention research, science education researchers commonly *postpone* in-depth theoretical and empirical study of implementation until an intervention reaches the scale up stage. In fact, sustainability can be theorized at earlier stages in the development of interventions (e.g., Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000; Fishman & Krajcik, 2003), and designers can draw on prior research to anticipate and plan for what are often predictable dilemmas of implementation (Weinbaum & Supovitz, 2010) and for cycles of iteration to improve interventions and the supports within systems for their implementation (Supovitz, 2008).

In summary, effectiveness research is one example of translational research that focuses on the characteristics of interventions that can produce robust learning outcomes across a variety of settings, but effectiveness research does not address three important questions the field faces:

- How can we incorporate considerations of implementation and sustainability earlier in program development?
- How do we change conditions that inhibit implementation of potentially effective programs for improving science learning?
- How do we promote principled adaptation of programs, especially in classrooms with students from nondominant communities?

Models for research that answers these questions are not abundant in education, in part because of how agencies and foundations currently fund research. Fortunately, we can find them in other fields, most notably the health sciences.

A More Comprehensive Typology of Large-Scale Intervention Research

In 2006, the National Institutes for Health established the National Network for Transforming Clinical and Translational Research, with the aim of funding large-scale research centers to design and test strategies that address gaps between research and practice in the behavioral and health sciences. The kind of research that these centers conduct is often called ''translational research,'' because it focuses on the translation of research into practice. The ultimate goal of such research, and of the centers, is to ensure that ''new treatments and research knowledge actually reach the patients or populations for whom they are intended and are implemented correctly'' (Woolf, 2008, p. 211).

Within the health sciences, there are two critical stages of translational research. The first involves the translation of basic science into treatments of different kinds and the study of their effectiveness. The National Institutes of Health (NIH) has named this type of translational research is called ''Type I Translation'' research. It corresponds to the stages of research common in biomedical research that proceeds from basic to applied science. At the same time, it is widely recognized in the health sciences that—despite the strength of evidence available to medical and public health practitioners on effective treatments—research must also focus on a different form of translation, namely the translation of findings from clinical studies to practice and patient decision making (Sung et al., 2003). According to Woolf (2008, p. 211), this ''Type II Translation'' research focuses in closing this gap between research and practice:

> The second area of translational research seeks to close that gap and improve quality by improving access, reorganizing and coordinating systems of care, helping clinicians and patients to change behaviors and make more informed choices, providing reminders and point-of-care decision support tools, and strengthening the patient-clinician relationship.

The design of Type II translation strategies and conducting research on those strategies requires a wide range of expertise, including epidemiology, behavioral science, organizational theory, systems redesign, clinical practice, and mixed-methods research. Perhaps most critically, it requires a systems perspective on the phenomenon of bringing effective treatments to scale, something that Type I translation research does not require in order to develop new and effective treatments. It also requires a design orientation, one that goes beyond analyses of conditions associated with implementation of the kind that can be conducted within the context of large-scale efficacy or effectiveness research. It requires a design orientation because the aim of Type II translation research is not primarily to explain implementation but rather to *improve* it. This is analogous to the interventionist approach embedded in design-based research approaches often employed in science education intervention research (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003).

The clinical research on these interventions includes rigorous tests of the effectiveness of Type II translation strategies. In medicine, for example, researchers have used small-scale experiments to test the efficacy of workshops aimed at improving medical practitioners' use of evidence-based approaches in their practice (Cochrane Collaboration, 2005) and large-scale cluster randomized controlled trials comparing the efficacy of different dissemination strategies (Watson et al., 2002). Public health researchers have also conducted experiments comparing the efficacy of different approaches to implementation support for public health providers. For example, Kelly and coworkers (J. A. Kelly et al., 2000) studied three different approaches to professional development to support the implementation of AIDS prevention programs. In their study, they randomly assigned participants to one of the programs and then compared the impacts of the programs on rates of program adoption and implementation.

Mapping the Type I and Type II Distinction Onto Educational Research

There are of course important differences between research in the clinical sciences and research in science education. There are, however, ways to map this distinction between Type I and Type II translation research in the clinical sciences onto education that can help reveal gaps in current large-scale intervention research. Table 1 shows one possible mapping.

As in the clinical sciences, the two basic forms or tasks of translation are the same for education. The first task is the translation of basic science into interventions. We emphasize the need for interventions to draw upon basic research on how children learn in specific domains, in part because advances in these areas are a key driver of the need for new interventions. In science education, advances in the field—such as the emergence of new fields like nanoscience—are also important drivers in the need for new interventions; however, researchers can make productive use of existing research about children's thinking and reasoning in developing interventions in these areas (e.g., Hsi, Sabelli, Krajcik, Tinker, & Ellenbogen, 2006).

The kinds of research appropriate for each form of translation research in education are distinct from the clinical sciences. In the earliest stages of research, principles derived from laboratory studies or teaching experiments in the field are re-contextualized into new materials and activities. This design activity often involves close collaboration with practitioners and iterative refinements to design goals and strategies, a methodology that learning scientists refer to as design-based research (Cobb et al., 2003). At an intermediate stage of development, researchers may conduct implementation research studies with small samples of classrooms, as a strategy for informing further changes to designs to improve their potential for impact (Confrey, Castro-Filho, & Wilhelm, 2000). In both kinds of studies, researchers and teachers struggle with how to interpret learning situations and develop preliminary evidence sufficient to warrant further investments in programs and interventions (A. E. Kelly, 2004). In education, furthermore, different stakeholder groups, including researchers, regularly contest

Table 1
Major distinctions between Type I and Type II translation research in education

	Type 1 Translation	Type II Translation
What is being translated?	Translating principles from basic learning research into interventions	Translating interventions developed for one or a few settings into interventions that are scalable to many settings
What kind of research is involved?	Design-based research Efficacy and effectiveness trials	Implementation research
What kinds of questions does translational research answer?	What do people learn from this design?	What kinds of capacities are required for organizations to implement this design?
	How do people learn from this design?	What supports are needed for people implementing the design to adapt it in ways congruent with the design's core principles?
	What do problems in learning or implementation suggest about redesign?	
Who is involved?	Learning scientists, classroom teachers, subject matter experts, often also software developers	Learning scientists, organizational researchers, teacher leaders, school and district administrators, often also publishers and enterprise software engineers

not only strategies for improvement but also the very goals for education (K. O'Connor & Penuel, 2010).

It is because of the fact that different stakeholder groups—from district leaders to teachers, parents, and students—can contest the goals, strategies, and conclusions of researchers that we prefer not to use the term ''translational'' to characterize the forms of research we have described so far. The mapping we have done shows ways that health sciences have developed more elaborate ways for characterizing and conducting research on implementation that we think can be applied to science education research. The metaphor of ''translation,'' however, is problematic in education, because it can imply a one-way of translation from educational research *to* practice. In our work, practitioners can be and are often engaged in shaping not just the use of research but also its production. In some cases, practitioners challenge researchers' suggestions; in others, by enacting suggestions from researchers, practitioners clarify those suggestions in ways researchers could not anticipate. Further, the metaphor of translation obscures too much of the work that must be done to achieve equity in educational systems. As Willinsky (2001) puts it,

> [I]mproving the educational situation of such challenging contexts will ultimately be about the allocation of scarce resources-good teachers, well-equipped classrooms, and other educational opportunities-which will always be about more than translating the best research into the best practice. It will entail the hotly contested politics of equity and entitlement, the advocacy of dedicated leaders and interest groups, all of whom could be better informed, presumably, by better access to the relevant research. (p. 10)

At present, it is principally policy researchers and sociologists of education who are likely to focus on issues of equity at scale, conducting analyses of implementation of programs that examine the associations among organizational context, implementation, and outcomes and the capacities required to implement interventions well (McDonald, Keesler, Kauffmann, & Schneider, 2006). Further, educational research that focuses on the design and impact of strategies for *improving* implementation has largely been limited to the domain of professional development, with little attention to the improvement of organizational processes (Halverson, Feinstein, & Meshoulam, 2011). Such research, if it is to develop as an area within science education, will require the formation of interdisciplinary teams with broad expertise in the learning sciences, organizational studies, leadership, engineering, and educational practice. Further, it will need to employ methods from the learning sciences, specifically that employ an iterative, collaborative process of design and research. Hence, we have called this form of research ''design-based implementation research,'' because it is a form of design-based research that is aimed simultaneously at developing interventions and at improving their implementation.

In contrast to models of large-scale intervention research that postpone the systematic study of implementation until after a program has been demonstrated to be efficacious, we suggest here that investigation of implementation can be incorporated within early-stage DBIR by addressing the question: What do problems in learning or implementation suggest about redesign? Here, we can draw on computer science and engineering. In those fields, implementation research has influenced design through a family of practices that go by various names: contextual design (Beyer & Holtzblatt, 1997), rapid prototyping (Gorden & Bieman, 1995), and rapid ethnography (Millen, 2000). Although the specifics of each practice differ, practitioners of this family of approaches to design all employ social science methods ''up front'' to identify the dilemmas and problems of practice that could inform the design and development of products. As we describe below, there are also examples of similar kinds

of studies within science education and the learning sciences. A key assumption, in focusing on implementation early, is that examples of many of the problems one finds in implementing interventions at scale are predictable (Weinbaum & Supovitz, 2010), consistent across a variety of school contexts, and revealed in early stages of design-based research.

An Example of DBIR in the Early Stages of Intervention Development

A good example of DBIR as part of research in a systemic context is the work of the Center for Learning Technologies in Urban Schools (LeTUS). LeTUS was a long-term research and development collaboration between university researchers and school teachers and administrators to develop inquiry-oriented science materials, with the end-goal of improving middle grade students' science performance in urban settings (Geier et al., 2008; Marx et al., 2004). Relatively early in the development of LeTUS interventions, a set of linked research studies was conducted to refine Earth science curriculum materials and professional development. Researchers on the project studied the implementation of the unit and then worked collaboratively with curriculum developers and science educators to improve the quality of professional development supports for teachers such that their enactments led to improved student learning. The study focused on a single concept, students' map reading skills necessary to interpret maps of watersheds, and researchers used items from the team's proximal assessments of student learning as the basis for making judgments about where to focus improvements on professional development and about the success of those improvements (Fishman, Marx, Best, & Tal, 2003).

The research unfolded as a sequence of multi-method studies. The researchers collected pre- and post-student learning data from their assessments, observation data, surveys, and interviews with teachers during one enactment of the unit. Next, the team made refinements to the professional development to give teachers practice with the aspects of the curriculum related to map reading skills and to engage them in analyzing student responses to the items and in brainstorming strategies for developing students' map skills. Then, the research team conducted a second study of the enactment, to determine whether or not the revised professional development had produced improvements in the enactment and in student learning. Results for the second enactment were significantly better for students, suggesting the promise of their iterative approach to studying implementation and refining professional development on the basis of the research.

The series of studies in the LeTUS work and the broader research context into which they fit reflect the distinctive goals of DBIR. First, a key goal of the research was to enhance the usability of the curriculum materials. Their efforts to enhance usability were guided by their theory of usability that emphasized the importance of the fit between the local capacity to support a curricular intervention (e.g., by providing needed resources, access to expertise useful for implementation) and the requirements of the curricular intervention itself (Blumenfeld et al., 2000). The scope of the research included both classrooms and professional development settings, seeking to trace links across these two settings that could improve learning outcomes for students. In that respect—and in contrast to much intervention research—the series of studies looked across contexts of adult and student learning and sought to coordinate better the learning opportunities in such a way as to improve outcomes for students.

An Example of DBIR With Published Curriculum Materials

Examples of DBIR conducted on mature programs and published curricula are more difficult to find in science education. Primary examples of such research are studies that have

focused on teacher professional development (e.g., Desimone, Porter, Garet, Yoon, & Birman, 2002; Roth et al., 2011; Supovitz & Turner, 2000; Supovitz & Zief, 2000) and on teachers' curriculum use (e.g., Drake & Sherin, 2006; Schneider & Krajcik, 2002). What distinguishes these studies from most scale-up research is their focus on the *conditions* under which particular interventions can be effective.

A study completed recently by Penuel and coworkers (Penuel & Gallagher, 2009; Penuel et al., 2011) illustrates how a large-scale experiment can systematically compare different models of teacher support aimed at *improving* implementation effectiveness. This study focused on conditions under which teachers' adaptations of curriculum might support, rather than hinder, making improvements to both teaching and learning. Instead of viewing adaptation as a problem to be solved, the study put at the heart of its inquiry a central question in policy debates today: Should we prepare teachers to adopt, adapt, or create curriculum materials for students? The study did not set out to resolve this debate but rather to inform it by a study of what happens when we randomly assign teachers to different support conditions that correspond to these alternatives in one subject area in a single school district.

This efficacy trial compared the impacts of three different programs for preparing teachers to teach for deep understanding of Earth science concepts, following the Understanding by Design (Wiggins & McTighe, 1998) model for unit creation. All three programs tested in the study reflected research-based principles for professional development (e.g., were of a significant duration, involved teachers in active learning strategies), but they differed with respect to the role they gave to teachers with respect to the curriculum. In the Adopt program, teachers learned how to *adopt* high-quality curriculum materials developed by experts in Earth science and curriculum design. In the Design program, teachers learned how to *design* curriculum experiences aligned to local standards using available materials and lessons they developed themselves. In the Principled Adaptation program, teachers learned how to *adapt* expert-developed materials in a principled way to align to local standards. To test the efficacy of the program, teachers who volunteered for the study were randomly assigned to one of the three programs or to a "business-as-usual" control group, and changes to teaching and learning were documented using a combination of surveys, observation, analyses of lesson plans, and standards-aligned tests of student learning. Teachers in all four groups came from a district seeking to promote the Understanding by Design model, but the district had not yet made significant investments in professional development to support implementation.

Results of the analyses of teacher survey and observations at the end of the first year of implementation (Penuel et al., 2009) indicated all three programs affected teachers' instructional planning and practice in Earth science relative to controls, though effects differed by program, and the program was not as effective in some areas as in others. After a year, teachers in the Design and Principled Adaptation programs reported significant changes to their unit planning process, a finding that is also consistent with intent of the professional development designs for those conditions. In particular, teachers reported that the programs had affected both the process by which they planned and the content of their units. Consistent with the idea that all adoption involves some adaptation, teachers assigned to the Adopt program also reported making some changes to how they planned units of instruction. Qualitative data from the implementation survey revealed the nature of their changes was different from that of the teachers assigned to the Design and Principled Adaptation programs, in that their changes were largely limited to creating pacing guides to go with the curriculum materials they were expected to adopt. Observational data showed that all three programs produced students who could provide explanations for why their teacher had them engage in particular activities with reference to a big idea in the unit. At the same time, none

of the designs had an impact on the probability that students would be observed engaging in explanation or application. Further, in contrast to the data on instructional planning that would suggest a greater attention to assessments, observers did not find teachers making use of preconceptions in their instruction. Results of the study of the programs' impacts on student learning also showed significant impacts of the two programs that provided explicit models for teaching with the materials and that prepared teachers by engaging them in the activity of design (Penuel et al., 2011).

This study of adaptation illustrates the potential of experimental research to advance scientific understanding of how best to support teacher enactment of curriculum. This study explicitly compares different models of teacher support, and despite the study's reliance on a volunteer population, the employment of a random assignment design helps reduce bias associated with teacher selection (Shadish, Cook, & Campbell, 2002). In contrast to experiments that are focused mainly on the effects of particular programs and curricula on student learning, this study devoted considerably more resources to teacher professional development and its effects on teaching and learning. As such, the study provides some preliminary evidence in answer to the question of how to productively support teacher adaptation.

A Focus for DBIR for the Next Decade: The Next Generation Science Standards

There are still too few examples of the kind of efforts described above to conduct research focused on improvement of the scalability and sustainability of interventions that aim especially at improving equity in access and learning outcomes. Furthermore, the past few decades have revealed the persistence of inequities, rather than their reduction or transformation, both in education and in society more broadly (Reich, 2011). These inequities persist, despite major investments in standards-based reforms in science education. We can learn from them, in order to anticipate the kinds of challenges that new reforms are likely to face in going to scale and becoming self-sustaining. Similarly, we can learn from them what might be termed a ''critical stance'' toward improvement efforts, namely how we must attend to and change the conditions that inhibit implementation of potentially effective programs for improving science learning, especially within schools and other settings where there are high concentrations of students of color and families living in poverty. At the same time, we are likely to need to invent and test new strategies for change that can be adapted to multiple contexts and that promote educators' adaptivity to student diversity.

The coming implementation of next generation science standards offers a potential laboratory for developing and testing new methods of translational research. The new science education framework (National Research Council, 2011) presents a new vision for science education intended in part to guide development of a next generation of standards. The new vision highlights blending of ''science and engineering practices,'' ''core ideas in science disciplines,'' and ''crosscutting themes'' within science (National Research Council, 2011). Though the next generation science education standards that are intended to reflect this vision from the framework are still in development, it is not too early to anticipate and plan for a DBIR agenda related to them. The past fifteen years of policy research should lead us to expect wide variability in standards implementation and provides insights into mechanisms for improving implementation and barriers to achieving equity. In addition, an emerging body of research focused on relating everyday and scientific ways of thinking and reasoning offers a promising approach to the question of ''what scales'' when the primary goal of interventions shifts to focus on promoting productive adaptation of curriculum materials. In the remainder of this article, we illuminate some of the insights from past research on standards and emerging research on relating everyday and scientific ways of knowing and describe how

both can inform a research and development agenda focused on improving the likelihood that implementing the new science standards will successfully broaden access to high-quality opportunities to learn for all learners.

Insights From Implementation Research on the First Generation of Science Standards

To address the question of how we can incorporate considerations of implementation and sustainability earlier in program development, we must first draw on lessons from the recent past. During the 1990s, national committees developed two standards documents: the *National Science Education Standards* (National Research Council, 1996) and the *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993). These documents were intended to inform state and local science education policy, curriculum development, and science teaching. To support these goals, additional committees formed to elaborate on specific aspects of the standards, including science inquiry (National Research Council, 2000) and classroom assessment (National Research Council, 2001). States developed their own standards, modeled after the national committees' standards. These efforts were supported simultaneously by federal agencies' investments in programs of professional development and systemic reform initiatives aimed at helping educators across different levels of systems (both districts and states) bring curriculum, teaching, and assessment into alignment with new standards.

These coordinated efforts to support standards implementation fell short of achieving broad scale improvements to science education. Educational leaders' interpretations of standards for inquiry varied widely, in ways that were consequential for how standards were implemented in different settings (Spillane & Callahan, 2002). The most widely available curriculum materials reflected few of the ideals of the standards, both with respect to providing students opportunities to learn from direct encounters with phenomena (Kesidou & Roseman, 2002) and with respect to assessment (Stern & Ahlgren, 2002). In some instances, the new standards conflicted with dominant forms of teaching, reducing opportunities to learn, especially for young people in low-income, urban settings (Songer, Lee, & Kam, 2002). As a consequence, students from urban neighborhoods, from nondominant cultural communities, and from low-income families became less likely to experience curricula that provide them with direct encounters with phenomena and opportunities to make sense of them (Smith et al., 2007).

Even so, research in the past decade has identified a number of promising strategies for promoting effective implementation of standards. For example, educational leaders who activate material, human, and social resources for improving science learning can be successful in achieving steady improvements to science teaching and learning (Spillane, Diamond, Walker, Halverson, & Jita, 2001). Curriculum materials organized around learning goals can provide significant supports for teachers' learning about standards and enacting practices that promote students learning science (Krajcik, McNeill, & Reiser, 2008). Teachers who receive significant, sustained, and content-focused professional development may be more likely to adopt teaching practices consistent with the standards (Garet, Porter, Desimone, Birman, & Yoon, 2001; Smith et al., 2007) and make effective use of curriculum materials (Penuel et al., 2011). These different strategies—activating technical and social capital through leadership, developing aligned curriculum materials, and providing coherent, sustained, content-focused professional development—are all likely to be key to implementing the next generation of science standards as well.

Since the first generation of standards was introduced, more science education researchers have turned their attention to the challenge of broadening participation in science. A key

goal of this research has been to develop an understanding of how to expand opportunities for students from nondominant groups, including girls and youth of color, to learn science. A consistent finding, moreover, is that apprenticeship to scientific practices is neither a simple nor straightforward matter for many students; rather, it is necessary to consider how learners understand themselves (i.e., their identity) as an essential aspect of science learning (Carlone et al., 2011; Hazari, Sonnert, Sadler, & Shanahan, 2010; National Research Council, 2009). Below, we review key findings from this research and suggest ways that it can inform DBIR for the next generation of science standards.

Insights From Research on Relating Everyday and Scientific Thinking

Strengthening students' competency with scientific practices related to specific science content is a key goal of the framework for the next generation of science standards in the United States (National Research Council, 2011). Accomplishing this goal is likely to be challenging, because participation in the social practices of science requires the mastery of specialized forms of discourse that often differ from everyday ways of speaking about familiar phenomena in the natural and social world (Gee, 2004; Lemke, 1990). Not only do scientists use specialized vocabulary and lexical forms (Halliday & Martin, 1993; Martin, 1989), they also use evidence from models and direct investigations to construct explanations that are intended to convince peers of both their point of view and conclusions (Latour, 1987). In most science classrooms, students have few opportunities to engage in these practices: far more frequent are discursive practices that require students to provide short answers in written or oral form to questions posed ahead of time by the teacher (National Research Council, 2007). Tests question students' recall of scientific facts, rather than providing them with opportunities to construct explanations in ways that approximate scientific practices (Lemke, 1990).

Meeting the challenge presented by the new framework for science standards will thus require most teachers to introduce new discursive practices into their classrooms. One line of inquiry into how to facilitate students' productive participation in scientific practices focuses on *talk moves* as tools for teachers. Talk or conversational moves are discursive practices teachers can use to elicit student thinking, promote scientific reasoning, encourage students to explain their thinking so others can understand, and build knowledge within classroom communities (M. C. O'Connor & Michaels, 2011). Some talk moves, such as revoicing, position students differently vis-à-vis one another and scientific knowledge in ways that support these goals (M. C. O'Connor & Michaels, 1993). Emphasizing talk moves as tools for facilitating student participation in scientific practices highlights their function or purpose within classroom teaching and elevates their importance as a potential resource for improving teaching quality (Windschitl, Thompson, & Braaten, 2008).

For some students, learning to participate in these new forms of discourse will not be easy. The forms of talk characteristic of both school and school science are unfamiliar, and diverge from those that characterize the forms characteristic in their families and communities (Ballenger, 2009; Rosebery, McIntyre, & Gonzales, 2001). A related line of research on talk moves focuses on this challenge, emphasizing moves that help students relate familiar, everyday ways of speaking and reasoning to more unfamiliar and scientific ways of speaking and reasoning. Researchers in TERC's Chèche Konnen Center have engaged students from nondominant cultural communities in a form of ''science talk'' they named ''Sherlock,'' in which students are encouraged to develop arguments as a group from classroom and personal experience (Rosebery et al., 2010). Some of the forms of talk also relate specifically to forms of talk and reasoning familiar to the many Haitian immigrant students whom Center researchers

and teachers teach (Hudicourt-Barnes, 2003). In this center's projects and other studies, the practice of helping students think with others through participation in classroom discourse has had positive effects on student learning, particularly among students from nondominant groups (Ballenger, 2009; Gallas, 1995; Rosebery, 2005; Rosebery & Warren, 2008; Tzou, Bricker, & Bell, 2007).

Adapting these tools effectively does require both skill and the adoption of a particular perspective on students' thinking. Talk moves' success depend on their being responsive to both affective and cognitive dimensions of student learning (Aguiar, Mortimer, & Scott, 2010; Herrenkohl & Mertl, 2010). Likewise, it requires teachers to take the heterogeneity of forms of speaking and thinking that students bring to the classroom in a way that does not imply that difference constitutes a deficit in students' thinking (Bang, Medin, & Atran, 2007; Gutiérrez & Orellana, 2006; Rosebery & Warren, 2008). A central goal of instruction becomes to help students navigate among different epistemologies, rather than adopt one over another in a way that may lead to devaluing of particular cultural practices (Bang & Medin, 2010) or identities (Carlone et al., 2011).

Preparing teachers to use these moves is not a simple or easy matter (Michaels & O'Connor, 2011), and students themselves may meet their teachers' efforts with resistance (Carlone, 2004). Significant professional development and practice are both required to develop skill in using the moves and in addressing student concerns about their positioning within new classroom participation structures. Teachers may need to develop new cultural competencies and an understanding of the meanings of everyday practices linked to particular communities (Ares, 2010). Even so, the effort to develop clear guidance for teachers with respect to the purposes and forms of talk moves and as the growing evidence base about their effects suggest that talk moves represent one promising response to the question of ''What scales?'' when a major goal of improvement becomes to promote teachers' adaptivity in teaching, especially to the contributions of students from nondominant communities.

Four Possible Areas of DBIR With respect to the New Science Standards

The field of science education faces significant challenges in the coming decade. Budget crises at the state and federal levels mean that we face the likelihood of threats to infrastructures in education that are essential to improvement, from professional development to curriculum research and development. We are likely to be asked to ''do more with less,'' despite the significant time and resources that are required to bring about systemic change. At this very time, the field is seeking to implement new standards in science to guide the development of a more coherent system of curriculum, assessments, and professional development (National Research Council, 2010).

As with all new policies, the implementation of the standards will present significant learning challenges to teachers, schools, and districts (cf., Cohen & Hill, 2001). Teachers will need to reorganize instruction to emphasize fewer ideas and develop strategies for integrating content, science and engineering practices, and crosscutting themes. Schools and districts will need to coordinate instructional programs and materials across grade levels, to reflect the ways that standards will attempt to build progressively sophisticated understandings of content, practices, and cross-cutting themes. Teachers will need to develop a new language and strategies for teaching focused on supporting students' engagement with ''practices to learn content'' rather than with ''inquiry.''

To address these challenges, we propose four areas in which DBIR might focus on identifying strategies for supporting broad and deep implementation of the new standards. The first two areas help us to address the broad question we introduced at the beginning of this article,

''How do we change conditions that inhibit implementation of potentially effective programs for improving science learning?'' They pertain to research that could be conducted on organizational strategies for crafting coherent science programs and strategies for supporting diverse learners' trajectories towards meaningful STEM participation. The final two areas help us address the broad professional development that prepares teachers to adapt high-quality science curriculum materials, instructional strategies to promote content learning through productive engagement in STEM practices. We argue that these four arcs of research are an ideal fit for DBIR, in which the primary research questions are not just about ''what works,'' but also about ''what works when, for whom, and under what conditions.'' In making this argument, we refer the reader back to Table 1. Each of these four areas take challenges that are faced by many schools, districts, and states, and attempts to address those challenges through explicit partnership or connection with programs of research that have tried to understand these problems in one or a few settings. In order to address these problems, it is necessary to identify needed organizational capacities and develop an understanding of how to develop these capacities. Finally, to address each of these areas will require broad coalitions and partnerships that engage science educators, learning scientists, teacher leaders, school administrators, policy makers, and others.

Organizational Strategies for Crafting Coherent Science Programs. In the face of constantly shifting policy and program demands, effective school and district leaders must ''craft coherence'' in their instructional programs by selectively appropriating and modifying resources to achieve local goals for improvement (Honig & Hatch, 2004). Research and development on strategies that schools and districts can pursue that will improve the coherence of their science programs is needed.

One such strategy for which there is ongoing research is the cultivation of professional learning communities in science (Gerard, Bowyer, & Marx, 2008; Knapp & Plecki, 2001). There is much to draw on from other disciplines about the conditions for promoting effective professional learning communities. In the area of writing, for example, researchers have pointed to the value of networks of teachers that span school boundaries for promoting reflection on practice (Lieberman, 2000; Lieberman & Wood, 2002). Within schools, research in mathematics education has focused on the need for well-designed protocols as anchors for fostering deep, critical conversations about student work (Horn, 2010; Horn & Little, 2010). Still other professional communities have been organized at the district level around the improvement of formative assessment practices (McMunn, McColskey, & Butler, 2004). In science education, we have focused relatively little research on how these organizational strategies can support standards implementation or how such strategies can best be coordinated with professional development and leadership activities.

More recently, there has been strong policy interest in the creation of schools that are focused on or have a Science, Technology, Engineering, and Mathematics (STEM) theme (PCAST, 2010). The concept is already being implemented at scale in a number of states, districts, and charter networks, and new schools that seek to develop STEM interest, engagement, and achievement among students who are traditionally underrepresented in STEM fields (Means, Confrey, House, & Bhanot, 2008). At the same time, to date there has been limited systematic empirical research on the impacts of these schools on student learning or on the kinds of course offerings and organizational processes of inclusive STEM school models that produce positive outcomes.

Of course not all science reforms presume that schools will undertake school-wide efforts to improve teaching and learning. In these contexts, the coherence of the instructional

program is no less likely to shape the course and success of reforms. Additional tools for helping leaders in these schools analyze the coherence of current materials and programs will be needed (e.g., Bessell, Burke, Plaza, Lee, & Schumm, 2010; Newmann, Smith, Allensworth, & Bryk, 2001), as will be policy documents targeting actors at different levels of the system that help them make sense of the new standards (see, e.g., Tabak, 2006).

Strategies for Supporting Diverse Learners' Trajectories Toward Meaningful STEM Participation. The next-generation science standards reference the idea of organizing coherent sequences of instruction around learning progressions, which are cognitive models for how students might grow in understanding and skill across their school careers (Corcoran, Mosher, & Rogat, 2009). Fostering both interest and excitement is of critical importance in developing future scientists and engineers (Tai, Liu, Maltese, & Fan, 2006; Tai & Maltese, 2010). In addition, professionalization requires young people to navigate numerous dilemmas associated with becoming a scientist or engineer and to gain access—often through sponsorship by mentors—to participation in disciplinary practices (Stevens, O'Connor, Garrison, Jocuns, & Amos, 2008). Cultural brokering—the activity of helping young people bridge values and cultural norms of nondominant communities to those of a scientific discipline—may be particularly critical at different points of childhood, adolescence, and young adulthood (Cooper, Denner, & Lopez, 1999). So, too, may be the creation of hybrid cultural spaces, where participants can explore, confront, and transform dilemmas they face in constructing discipline-based identities (Gutiérrez, Baquedano-Lopez, & Tejada, 2000).

We need research that can help us understand the conditions under which students from culturally and linguistically diverse backgrounds gain access to and have opportunities to contribute to disciplinary practices. Some of this research can take place within efficacy and effectiveness research, or as part of observational studies that seek to model the extent to which students from different backgrounds or different schools are exposed to reform-based instructional practices (Maerten-Rivera, Penfield, Myers, Lee, & Buxton, 2009). In addition, so-called ''social design experiments'' (Gutiérrez & Vossoughi, 2010), in which the aim of research is to help young people from nondominant communities engage in collective activity to promote their learning and development, may be particularly important for realizing the promise of learning progressions. In social design experiments, the aim is not simply to expand access to existing trajectories of opportunity for young people: it is to create entirely new trajectories of participation, new ways to be productive, engaged, and successful citizens (K. O'Connor & Allen, 2010). Such experiments can help develop tentative answers about what is needed for interventions to work at scale with diverse student groups (Lee & Luykx, 2005).

Professional Development That Prepares Teachers to Adapt High-Quality Materials. Developing students' proficiency in science depends on teachers' skill in sequencing instructional experiences that build understanding over time (National Research Council, 2007). Curriculum materials are important resources for teachers, both in developing their skill in design and in providing models of standards-aligned activities (Krajcik et al., 2008). Though researchers in the past decade have focused principally on maximizing fidelity of teachers' implementation of curriculum materials, research indicates that teachers can benefit from professional development that prepares them to make principled adaptation of curriculum materials (Penuel & Gallagher, 2009). Additional research and development on decision support tools that help teachers adapt materials in ways that align to standards and support learning goals could be a significant aid for new science education standards implementation.

With limited budgets for new textbooks, districts may turn increasingly to online curricula. In this context, digital tools that help teachers to customize curricula may be important to develop. Some of these tools can provide feedback to teachers on the impact of their decisions to exclude or adapt materials, in terms of students' opportunity to learn a network of related concepts in science (H.-T. Lin & Fishman, 2004). Research is also needed on tools that allow teachers to adapt and then share high-quality materials with coworkers, a strategy that has shown some recent promise (Maull, Saldivar, & Sumner, 2011).

Instructional Strategies to Promote Content Learning From Productive Engagement in Science and Engineering Practices. Many teachers have little direct experience with the kinds of science and engineering practices promoted in the new standards, and they will need opportunities to learn about them. Of particular importance is preparing teachers to give students access to the discursive practices of science (Windschitl et al., 2008). Academic language is critical to success, and bridging everyday language to scientific practices for all students will require a sensitivity to the ways of thinking, speaking, and valuing characteristic of students' families and communities, as research on relating everyday and scientific ways of thinking and reasoning reviewed above shows (Warren et al., 2001).

Our view is that DBIR that investigates supports for teachers to develop a repertoire of tools for engaging in science and engineering practices is needed. Some of this research can focus on tools that will be useful to a broad range of students and teachers. These include research on preparing teachers to use "talk moves" in science to promote productive disciplinary engagement in practices such as argumentation (Chin & Osborne, 2010; M. C. O'Connor & Michaels, 2011; Penuel & DeBarger, 2011; Thompson, Windschitl, & Braaten, 2010). Other DBIR research will need to focus specifically on strategies that help teachers elicit and make productive use of forms of talk that their particular students employ at home and in their communities. The challenge of this research will be to produce strategies that can be used across different classroom contexts and in ways that do not presume that belonging to a particular community or cultural group means that all individuals adopt particular (stereotyped) ways of thinking and reasoning in relation to a particular science topic. Cultural communities are characterized by varied and changing repertoires of practices, and individuals within them appropriate different forms and ways of participation, depending on circumstances and their own experiences (Gutiérrez & Rogoff, 2003).

What's Needed to Develop DBIR in Science Education

In this article, we have argued for the need to expand what we consider to be a dominant model for large-scale intervention research in science education to include a sharper focus on implementation, both as a support for improving the scalability of designs and as an object of design. The proposed areas of DBIR outlined above should be seen as complementary to, and developing alongside, design, efficacy, and effectiveness studies that are sure to continue and are needed to create and yield evidence about the potential of programs and curricula that will be developed in the coming years to reflect the new standards. In other words, both early-stage DBIR of the kind currently supported both within the U.S. Department of Education and the National Science Foundation *and* this new kind of later-stage DBIR are needed to advance both research and practice in science education. Our intent is not to argue for one over the other.

As a field, we stand to learn much from implementation research that may be conducted in the coming decade that focuses on adoption of the next-generation science education standards. However, sustaining improvements to particular systems and gains made by teachers

and students who participate in this research is unlikely unless we develop a stronger research and development infrastructure (Bryk & Gomez, 2008; Donovan, Wigdor, & Snow, 2003). Especially critical are enduring partnerships between research and practice, as well as substantive, sustained investments at the federal and state levels in school reform and in professional development. Researchers pursuing these lines of inquiry should seek simultaneously to advance the aims of specific improvement initiatives and to build a more robust infrastructure for research and development.

A critical needed aspect of such an infrastructure is the development of a network focused on improving models of DBIR. This network should include researchers, practitioners, curriculum and program developers, and public and private investors with a stake in improving educational systems. Just as a network of scholars formed in the past decade into professional organizations such as the Society for Research on Educational Effectiveness to advance methods in conducting experimental research in education, we need a network for DBIR that shares common principles and a commitment to refinements in approaches to design and methods for research and development.

Looking to the future, we see some common principles across DBIR projects in science education today around which a network could form. First, we see a common commitment to solving problems of practice as constructed by educators and educational leaders, that is, from the perspective of those who will ultimately be responsible for implementing interventions. Second, DBIR engages in iterative, collaborative design of solutions targeting multiple levels of the system, design that is informed by ongoing and systematic inquiry into implementation and outcomes. Third, we see a common commitment to building theory and knowledge within the research community. The object of that theory is learning, but across scales of a system, where ''learning'' applies not just to students in classrooms, but to individual adult actors (e.g., teachers, principals), organizational units (e.g., schools, curriculum departments in districts), and systems. Finally, we see a commitment on the part of teams to develop capacity of practitioners to further improve science teaching and learning. This final principle is one for researchers to remember, in part to remind ourselves that within the current system, it is not we who implement interventions, but science educators.

This material is based upon work supported by the National Science Foundation under grant number BCS-0624307. 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. We thank Nora Sabelli and Kris Gutiérrez and two anonymous reviewers for their comments on earlier versions of this manuscript.

References

Aguiar, O. G., Mortimer, E., & Scott, P. (2010). Learning from and responding to students' questions: The authoritative and dialogic tension. Journal of Research in Science Teaching, 47(2), 174–193.

American Association for the Advancement of Science. (1993). Benchmarks for science literacy. New York: Oxford University Press.

Ares, N. (2010). Political and cultural dimensions of organizing learning around funds of knowledge. In W. R. Penuel & K. O'Connor (Eds). Learning research as a human science. National Society for the Study of Education Yearbook, 109(1), 192–206.

Ballenger, C. (2009). Puzzling moments, teachable moments: Practicing teacher research in urban classrooms. New York, NY: Teachers College Press.

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. Science Education, 94(6), 1008–1026.

Bang, M., Medin, D., & Atran, S. (2007). Cultural mosaics and mental models of nature. Proceedings of the National Academy of Sciences, 104(35), 13868–13874.

Barton, A. C., & Tan, E. (2009). Funds of knowledge, discourses and hybrid space. Journal of Research in Science Teaching, 46(1), 50–73.

Bessell, A. G., Burke, M. C., Plaza, M. P., Lee, O., & Schumm, J. S. (2010). The educational reform rating rubric: Example of a new tool for evaluating complex school reform initiatives. Field Methods, 20(3), 283–295.

Beyer, H., & Holtzblatt, K. (1997). Contextual design: A customer-centered approach to systems designs. San Francisco, CA: Morgan Kaufmann.

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

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. Journal of Research on Educational Effectiveness, 1(3), 237–264.

Bryk, A.S. (2009). Support a science of performance improvement. Phi Delta Kappan, 90(8), 597–600.

Bryk, A. S., & Gomez, L. M. (2008). Reinventing a research and development capacity. In F. M. Hess (Ed.), The future of educational entrepreneurship: Possibilities for school reform (pp. 181–187). Cambridge, MA: Harvard Educational Press.

Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlitz, B., & Willett, J. (2004). Model-based teaching and learning with BioLogicaTM: What do they learn? How do they learn? How do we know? Journal of Science Education and Technology, 13(1), 23–41.

Carlone, H. B. (2004). The cultural production of science in reform-based physics: Girls' access, participation, and resistance. Journal of Research in Science Teaching, 41(4), 392–414.

Carlone, H. B., Haun-Frank, J., & Webb, A. (2011). Assessing equity beyond knowledge- and skills-based outcomes: A comparative ethnography of two fourth-grade reform-based science classrooms. Journal of Research in Science Teaching, 48(5), 459–485.

Chin, C., & Osborne, J. (2010). Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. Journal of Research in Science Teaching, 47(7), 883–908.

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

Cochrane Collaboration. (2005). Continuing education meetings and workshops: Effects on professional practice and health care outcomes (Cochrane review). Journal of Continuing Education in the Health Professions, 21(3), 187–188.

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

Confrey, J., Castro-Filho, J., & Wilhelm, J. (2000). Implementation research as a means to link systemic reform and applied psychology in mathematics education. Educational Psychologist, 35(3), 179–191.

Cooper, C. R., Denner, J., & Lopez, E. M. (1999). Cultural brokers: Helping Latino children on pathways toward success. The Future of Children, 9, 51–57.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). Learning progressions in science: An evidence-based approach to reform. CPRE Research Report # RR-63. New York, NY: Center on Continuous Instructional Improvement, Teachers College, Columbia University.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), Strengthening research methodology: Psychological measurement and evaluation (pp. 103–124). Washington, DC: American Psychological Association.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. Educational Evaluation and Policy Analysis, 24(2), 81–112.

Donovan, S., Wigdor, A. K., & Snow, C. E. (2003). Strategic education research partnership. Washington, DC: National Research Council.

Drake, C., & Sherin, M. G. (2006). Practicing change: Curriculum adaptation and teacher narrative in the context of mathematics education reform. Curriculum Inquiry, 36(2), 153–187.

Dynarski, M. (2008). Bringing answers to educators: Guiding principles for research syntheses. Educational Researcher, 37(1), 27–29.

Fishman, B. J., & Krajcik, J. (2003). What does it mean to create sustainable science curriculum innovations? A commentary. Science Education, 87(4), 564–573.

Fishman, B. J., Marx, R. W., Best, S., & Tal, R. (2003). Linking teacher and student learning to improve professional development in systemic reform. Teaching and Teacher Education, 19(6), 643–658.

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.

Gallas, K. (1995). Talking their way into science: Hearing children's questions and theories, responding with curricula. New York: Teachers College Press.

Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, 38(4), 915–945.

Gee, J. P. (2004). Language in the science classroom: Academic social languages as the heart of school-based literacy. In E. W. Saul (Ed.), Crossing borders in literacy and science instruction: Perspectives on theory and practice (pp. 10–32). Newark, DE: International Reading Association, Inc.

Gee, J. P., & Clinton, K. (2000). An African American child's 'science talk': Co-construction of meaning from the perspective of multiple Discourses. In M. Gallego & S. Hollingsworth (Eds.), What counts as literacy: Challenging the school standard (pp. 118–135). New York: Teachers College Press.

Geier, R., Blumenfeld, P., Marx, R. W., Krajcik, J., Fishman, B. J., & Soloway, E. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. Journal of Research in Science Teaching, 45(8), 922–939.

Gerard, L., Bowyer, J., & Marx, R. (2008). A community of principals: Building school leadership for scaling technology-science curriculum reform. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Gorden, V. S., & Bieman, J. M. (1995). Rapid prototyping: Lessons learned. IEEE Software, 12(1), 85–95.

Gutiérrez, K. D., Baquedano-Lopez, P., & Tejada, C. (2000). Rethinking diversity: Hybridity and hybrid language practices in the third space. Mind, Culture, and Activity, 6(4), 286–303.

Gutiérrez, K. D., & Orellana, M. (2006). The problem of English learners: Constructing genres of difference. Research in the Teaching of English, 40(4), 502–507.

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. Educational Researcher, 32(5), 19–25.

Gutiérrez, K. D., & Vossoughi, S. (2010). Lifting off the ground to return anew: Mediated praxis, transformative learning, and social design experiments. Journal of Teacher Education, 61(1–2), 100–117.

Halliday, M. A. K., & Martin, J. (1993). Writing science: Literacy and discursive power. London: Falmer.

Halverson, R., Feinstein, N., & Meshoulam, D. (2011). School leadership for science education. In G. E. DeBoer (Ed.), The role of public policy in K-12 science education (pp. 397–430). Charlotte, NC: Information Age Publishing.

Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M.-C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. Journal of Research in Science Teaching, 47(8), 978–1003.

Herrenkohl, L. R., & Mertl, V. (2010). How students come to be, know, and do: A case for a broad view of learning. New York, NY: Cambridge University Press.

Honig, M. I., & Hatch, T. C. (2004). Crafting coherence: How schools strategically manage multiple, external demands. Educational Researcher, 33(8), 16–30.

Horn, I. S. (2010). Teaching replays, teaching rehearsals, and re-visions of practice: Learning from colleagues in a mathematics community. Teachers College Record, 112(1), 225–259.

Horn, I. S., & Little, J. W. (2010). Attending to problems of practice: Routines and resources for professional learning in teachers' workplace interactions. American Educational Research Journal, 47(1), 181–217.

Hsi, S., Sabelli, N., Krajcik, J., Tinker, R. F., & Ellenbogen, K. (2006). Learning at the nanoscale: Research questions that the rapidly evolving interdisciplinarity of science poses for the learning sciences. In S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), Proceedings of the 7th International Conference of the Learning Sciences (Vol. 2, pp. 1066–1072). Mahwah, NJ: Erlbaum.

Hudicourt-Barnes, J. (2003). The use of argumentation in Hatian Creole science classrooms. Harvard Educational Review, 73(1), 73–93.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. American Education Research Journal, 38(3), 499–534.

Jacob, B. (2007). The challenge of staffing urban schools with urban teachers. The Future of Children, 17(1), 129–153.

Kelly, A. E. (2004). Design research in education: Yes, but is it methodological. The Journal of the Learning Sciences, 13(1), 113–128.

Kelly, J. A., Somlai, A. M., DiFranceisco, W. J., Otto-Salaj, L. L., McAuliffe, T. L., Hackl, K. L., & Rompa, D. (2000). Bridging the gap between the science and service of HIV prevention: Transferring effective research-based HIV prevention interventions to community AIDS service providers. American Journal of Public Health, 90(7), 1082–1099.

Kesidou, S., & Roseman, J. E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. Journal of Research in Science Teaching, 39(6), 522–549.

Knapp, M. S., & Plecki, M. L. (2001). Investing in the renewal of urban science teaching. Journal of Research in Science Teaching, 38(10), 1089–1100.

Krajcik, J. S., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. Science Education, 92(1), 1–32.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Cambridge, MA: MIT Press.

Lee, O. (2002). Science inquiry for elementary students from diverse backgrounds. In W. G. Secada (Ed.), Review of research in education (pp. 23–69). Washington, DC: American Educational Research Association.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up innovations in elementary science instruction with nonmainstream students. American Educational Research Journal, 42(3), 411–438.

Lee, O., Maerten-Rivera, J., Penfield, R., LeRoy, K., & Secada, W. G. (2008). Science achievement of English language learners in urban elementary schools: Results of a first-year professional development intervention. Journal of Research in Science Teaching, 45(1), 31–52.

Lemke, J. L. (1990). Talking science: Language, learning, and values. Norwood, New Jersey: Ablex.

Lieberman, A. (2000). Networks as learning communities: Shaping the future of teacher development. Journal of Teacher Education, 51, 221–227.

Lieberman, A., & Wood, D. (2002). Inside the National Writing Project: Connecting network learning and classroom teaching. New York: Teachers College Press.

Lin, H.-T., & Fishman, B. J. (2004). Supporting the scaling of innovations: Guiding teacher adaptation of materials by making implicit structures explicit. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), Proceedings of the Sixth International Conference of the Learning Sciences (pp. 617). Santa Monica, CA: Erlbaum.

Lin, X., Schwartz, D. L., & Hatano, G. (2005). Toward teachers' adaptive metacognition. Educational Psychologist, 40(4), 245–255.

Lynch, S. J., Szesze, M., Pyke, C., & Kuipers, J. (2007). Scaling up highly-rated science curriculum units for diverse student populations: Features that affect collaborative research and vice versa. In B. Schneider & S-.K. McDonald (Eds.), Scale-up in education: Issues in practice (Vol. II, pp. 91–122). Lanham: Rowman & Littlefield Publishers.

Maerten-Rivera, J., Penfield, R., Myers, N., Lee, O., & Buxton, C. A. (2009). School and teacher predictors of science instruction practices with English language learners in urban elementary schools. Journal of Women and Minorities in Science and Engineering, 15(2), 93–118.

Martin, J. R. (1989). Factual writing: Exploring and challenging social reality. London: Oxford University Press.

Marx, R. W., Blumenfeld, P. C., Krajcik, J., Fishman, B. J., Soloway, E., Geier, R., & Revital, T. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. Journal of Research in Science Teaching, 41(10), 1063–1080.

Maull, K. E., Saldivar, M. G., Sumner, T. (2011 June). Understanding digital library adoption: A use diffusion approach. Paper presented at the ACM/IEEE Joint Conference on Digital Libraries, Ottawa, Canada.

McDonald, S.-K., Keesler, V. A., Kauffmann, N. J., & Schneider, B. (2006). Scaling up exemplary interventions. Educational Researcher, 35(3), 15–24.

McMunn, N., McColskey, W., & Butler, S. (2004). Building teacher capacity in classroom assessment to improve student learning. International Journal of Educational Policy, Research, & Practice, 4(4), 25–48.

Means, B., Confrey, J., House, A., & Bhanot, R. (2008). STEM high schools: Specialized science, technology, engineering, and mathematics secondary schools in the United States. Menlo Park, CA: SRI International.

Millen, D. R. (2000). Rapid ethnography: Time deepening strategies for HCI field research. Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (pp. 280–286). New York, NY: ACM Press.

National Research Council. (1996). National Science Education Standards. Washington, DC: National Academy Press.

National Research Council. (2000). Inquiry and the National Science Education Standards. Washington, DC: National Academy Press.

National Research Council. (2001). Classroom assessment and the National Science Education Standards. Washington, DC: National Academy Press.

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Academies Press.

National Research Council. (2009). Learning science in informal environments: People, places, and pursuits. Washington, DC: National Academies Press.

National Research Council. (2010). A framework for science education: Preliminary public draft. Washington, DC: Committee on Conceptual Framework for New Science Education Standards, Board on Science Education, National Research Council.

National Research Council. (2011). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: National Research Council.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. Educational Evaluation and Policy Analysis, 23(4), 297–321.

O'Connor, K., & Allen, A-R. (2010). Learning as the organizing of social futures. In W. R. Penuel & K. O'Connor (Eds). Learning research as a human science. National Society for Studies in Education, 109(1):160–175.

O'Connor, K., & Penuel, W. R. (2010). Introduction: Principles of a human sciences approach to research on learning. In W. R. Penuel & K. O'Connor, (Eds). Learning research as a human science. National Society for Studies in Education, 109(1):1–16.

O'Connor, M. C., & Michaels, S. (1993). Aligning academic talk and participation status through revoicing: Analysis of a classroom discourse strategy. Anthropology and Education Quarterly, 24(4), 318–355.

O'Connor, M. C., & Michaels, S. (2011). Problematizing dialogic and authoritative discourse, their coding in classroom transcripts, and realization in the classroom. Paper presented at the ISCAR Congress, Rome, Italy.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. Review of Educational Research, 78(1), 33–84.

Oakes, J. (1990). Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science. Santa Monica, CA: RAND.

(2010). Prepare and inspire: K-12 education in science, technology, engineering and math (STEM) for America's future. Washington, DC: Executive Office of the President.

Pea, R. D., & Collins, A. (2008). Learning how to do science education: Four waves of reform. In Y. Kali, M. C. Linn, & J. E. Roseman (Eds.), Designing coherent science education (Vol. 3–12). New York: Teachers College Press.

Penuel, W. R., DeBarger, A. H. (2011). Supporting teacher learning to improve classroom assessment in science. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Penuel, W. R., & Gallagher, L. P. (2009). Comparing three approaches to preparing teachers to teach for deep understanding in Earth science: Short-term impacts on teachers and teaching practice. The Journal of the Learning Sciences, 18(4), 461–508.

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in Earth science: A comparison of three professional development programs. American Educational Research Journal, 48(4), 996–1025.

Penuel, W. R., McWilliams, H., McAuliffe, C., Benbow, A., Mably, C., & Hayden, M. M. (2009). Teaching for understanding in Earth science: Comparing impacts on planning and instruction in three professional development designs for middle school science teachers. Journal of Science Teacher Education, 20(5), 415–436.

Reich, R. B. (2011). Aftershock: The next economy and America's future. New York: Vintage.

Rethinam, V., Pyke, C., & Lynch, S. (2008). Using multi-level analyses to study the effectiveness of science curriculum. Evaluation & Research in Education, 21(1), 18–42.

Rosebery, A. (2005). ''What are we going to do next?'' A case study of lesson planning. In R. Nemirovsky, A. Rosebery, B. Warren, & J. Solomon (Eds.), Everyday matters in mathematics and science: Studies of complex classroom events (pp. 299–328). Mahwah, NJ: Erlbaum.

Rosebery, A., McIntyre, E., & Gonzales, N. (2001). Connecting students' cultures to instruction. In E. McIntyre, A. Rosebery, & N. Gonzales (Eds.), Classroom diversity: Connecting curriculum to students' lives (pp. 1–13). Portsmouth, NH: Heinemann.

Rosebery, A., Ogonowski, M., DiSchino, M., & Warren, B. (2010). ''The coat traps all your body heat'': Heterogeneity as fundamental to learning. The Journal of the Learning Sciences, 19(3), 322–357.

Rosebery, A., & Warren, B. (Eds.). (2008). Teaching science to English language learners. Arlington, VA: National Science Teachers Association.

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Videobased lesson analysis: Effective PD for teacher and student learning. Journal of Research in Science Teaching, 48(2), 117–148.

Rowan, B. (2002). The ecology of school improvement: Notes on the school improvement industry in the United States. Journal of Educational Change, 3(3–4), 283–314.

Scafidi, B., Sjoquist, D. L., & Stinebrickner, T. R. (2007). Race, poverty, and teacher mobility. Economics of Education Review, 26(2), 145–159.

Schneider, R. M., & Krajcik, J. (2002). Supporting science teacher learning: The role of educative curriculum materials. Journal of Science Teacher Education, 13(3), 221–245.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton-Mifflin.

Shear, L., & Penuel, W. R. (2010). Rock-solid support: Florida district weighs effectiveness of science professional learning. Journal of Staff Development, 31(5), 48–51.

Sloane, F.C. (2008). Through the looking glass: Experiments, quasi-experiments, and the medical model. Educational Researcher, 37(1), 41–46.

Smith, T. M., Desimone, L. M., Zeidner, T. L., Dunn, A. C., Bhatt, M., & Rumyantseva, N. L. (2007). Inquiry-oriented instruction in science: Who teaches that way? Educational Evaluation and Policy Analysis, 29(3), 169–199.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. Journal of Research in Science Teaching, 46(6), 610–631.

Songer, N. B., Lee, H.-S., & Kam, R. (2002). Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? Journal of Research in Science Teaching, 39(2), 128–150.

Spillane, J. P., & Callahan, K. E. (2002). Implementing state standards for science education: What district policy makers make of the hoopla. Journal of Research in Science Teaching, 37(5), 201–225.

Spillane, J. P., Diamond, J. B., Walker, L. J., Halverson, R., & Jita, L. (2001). Urban school leadership for elementary science instruction: Identifying and activating resources in an undervalued school subject. Journal of Research in Science Teaching, 38(8), 918–940.

Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. Journal of Research in Science Teaching, 39(9), 889–910.

Stevens, R., O'Connor, K., Garrison, L., Jocuns, A., & Amos, D. M. (2008). Becoming an engineer: Toward a three dimensional view of engineering learning. Journal of Engineering Education, 97(3), 355–368.

Sung, N. S., Crowley, W. F., Jr., Genel, M., Salber, P., Sandy, L., Sherwood, L. M., ... Rimoin, D. (2003). Central challenges facing the national clinical research enterprise. Journal of the American Medical Association, 289(10), 1278–1287.

Supovitz, J. A. (2008). Implementation as iterative refraction. In J. A. Supovitz & E. H. Weinbaum (Eds.), The implementation gap: Understanding reform in high schools. New York, NY: Teachers College Press.

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. Journal of Research in Science Teaching, 37(2), 963–980.

Supovitz, J. A., & Zief, S. G. (2000). Why they stay away: Survey reveals invisible barriers to teacher participation. Journal of Staff Development, 21(4), 24–28.

Tabak, I. (2006). Prospects for change at the nexus of policy and design. Educational Resercher, 35(2), 24–30.

Tai, R. H., Liu, C. Q., Maltese, A. V., & Fan, X. (2006). Planning early for careers in science. Science, 312, 1143–1144.

Tai, R. H., & Maltese, A. V. (2010). Eyeballs in the fridge: Sources of early interest in science. International Journal of Science Education, 32(5), 669–685.

Thompson, J., Windschitl, M., Braaten, M. (2010). Developing a theory of teacher practice. Paper presented at the National Association for Research in Science Teaching Annual Conference, Philadelphia, PA.

Tzou, C. T., Bricker, L. A., & Bell, P. (2007). Micros & Me: A fifth-grade science exploration into personally and culturally consequential microbiology. Seattle, WA: Everyday Science and Technology Group, University of Washington.

Vanosdall, R., Klentschy, M. P., Hedges, L. V., Weisbaum, K. S. (2007). A randomized study of the effects of scaffolded guided-inquiry instruction on student achievement in science. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. Journal of Research in Science Teaching, 38(5), 529–552.

Watson, M. C., Bond, C. M., Grimshaw, J. M., Mollison, J., Ludbrook, A., & Walker, A. E. (2002). Educational strategies to promote evidence-based community pharmacy practice: A cluster randomized controlled trial (RCT). Family Practice, 19(5), 529–536.

Weinbaum, E. H., & Supovitz, J. A. (2010). Planning ahead: Make program implementation more predictable. Phi Delta Kappan, 91(7), 68–71.

Wiggins, G., & McTighe, J. (1998). Understanding by design. Alexandria, VA: Association for Supervision and Curriculum Development.

Willinsky, J. (2001). The strategic education research program and the public value of research. Educational Researcher, 30(1), 5–14.

Windschitl, M., Thompson, J., & Braaten, M. (2008). How novice science teachers appropriate epistemic discourses around model-based inquiry for use in classrooms. Cognition and Instruction, 26(3), 310–378.

Woolf, S. H. (2008). The meaning of translational research and why it matters. The Journal of the American Medical Association, 299(2), 211–213.