SPECIAL SECTION

The Similarities Between Research in Education and Research in the Hard Sciences

Carl E. Wieman[1]

In this commentary, the author argues that there is a considerable degree of similarity between research in the hard sciences and education and that this provides a useful lens for thinking about what constitutes "rigorous" and "scientific" education research. He suggests that the fundamental property of hard science research is its predictive power, a property that can equally be applied to large- and small-scale and quantitative and qualitative research in education. Although variables may differ and methods of collection may not be the same, researchers do their best to measure and/or control those variables that matter, and design experiments and subsequent tests to ensure that those that can neither be measured nor fully controlled are unlikely to change the results in significant ways. He concludes that although fields like physics or chemistry are mature sciences, the "cutting-edge" work in these fields is often "messy," as researchers struggle to determine which variables are important. He suggests that education research often resembles the patterns seen in cutting-edge research in the "hard" sciences, as researchers are struggling to identify variables that are important to the problem.

Keywords: educational policy; educational reform; research methodology; research utilization

The criteria listed for federal education research funding again raises the question over what constitutes "rigorous educational research" and "scientific knowledge" in education, and what research designs meet these standards. Frequently these discussions focus on the distinctions between the "hard sciences" (physics, chemistry, biology, etc.) and education, and why these distinctions are important. Here I would like to offer a different perspective, how there is a considerable degree of similarity between research in the hard sciences and education, and how this provides a useful lens for thinking about what constitutes "rigorous" and "scientific" education research. I will also explain how the nature of research (as opposed to "scholarship" writ large) in the two areas is far more similar than researchers in either community recognize, largely because the nature of research in the hard sciences is often misunderstood and mischaracterized. True research in the hard sciences, when it is exploring fundamentally new ground, is much messier, complicated, and less precise than is usually recognized and, thus, more similar to education research. The errors that lead to flawed research also have much the same origin across the different fields.

My perspective on this subject comes from first spending 25 years as a physicist, doing "tabletop" scale experiments that involved working closely with small groups of graduate students. Those interactions led me to try to better understand how my graduate students developed from struggling novices in the lab to expert physicists within a few years. It was particularly puzzling to me that their success in physics courses was such a poor predictor of a student's ultimate success as a physicist. That puzzle led me to start systematically studying research on learning in general as well as specifically the learning of physics. What I learned convinced me that the explanation of the puzzle lay in the shortcomings of undergraduate teaching, and this also fueled my interest in exploring different teaching methods and making comparative measures of learning. That interest grew into my having two parallel research groups for many years, one in experimental atomic physics and the other in research in science education.

A fundamental test of research in the hard sciences is, does the result have predictive power? By that, I mean can one use the results to predict with reasonable accuracy what will happen, or what will be observed, in some new situation (at a minimum in a replication of the experiment as described by the original researcher)? This standard has served the hard sciences well over the years and, I argue, is correspondingly useful to use for education research. Even "unsuccessful" experiments can have value and predictive power. Consider for example, the result that "If you control this particular set of variables and introduce this intervention, there turns out to be no effect on the behavior of the atoms (or, equivalently, educational outcomes)." This is rigorous research and an important contribution to the knowledge base, if that result is correct, particularly if many people would have predicted a different outcome.

Applying this standard to the research does not mean it is necessary to accurately control and predict how every specific student will behave or learn, any more than we can control and predict how every single atom with behave in a physics or chemistry experiment. It means only that one should be able to predict some meaningful measureable outcomes. This is also not a criterion for the importance of the research. Importance depends on a number of other criteria that vary with field and personal opinions.

Considering the predictive power, and corresponding new insights, that a research study will provide is a more meaningful measure of its rigor and value than what particular research design it uses. For example, a good qualitative study that examines only a few students or teachers in depth will allow one to recognize, and hence more accurately predict, some factors that will be important in educational outcomes and important in the design of larger quantitative experiments in similar populations. Such qualitative research provides an important contribution to the knowledge base, albeit of a different sort than a randomized controlled trial that tests the impact of a large-scale intervention on multiple school districts. Similarly, research designs that provide good predictive power when used in the context of learning in university physics courses could be worthless if applied to experiments in diverse elementary school classrooms, because the latter context has so many additional variables that will impact the outcomes.

The way research goes bad is also quite similar between the hard sciences and education. Here, by "bad research," I mean that which provides incorrect or useless predictions. The serious errors in hard science research occur when important variables are overlooked, and this is also true in education research. Usually these variables are overlooked for the same reasons in all fields; the researcher is just sloppy, or more often, the researcher is failing to adequately address his or her inherent biases. Every researcher in every field has a result he or she wants to see and a belief as to what does and does not matter. In all types of research, it is essential to recognize these inherent biases and to have tests and procedures to prevent those biases from unduly influencing the results and conclusions.

Although the common perception is that experimental research in something like physics is much more controlled and precise than in education, and hence such errors are inherently easier to avoid, I do not believe that is actually the case. It is true only if one looks at some area of physics that is very mature and so has been so well studied that all the complications have been sorted out and controlled for. In such situations, there also is no room for significant surprises or breakthroughs, and the research is, at best, incremental. When an area of physics research is truly cutting-edge, pushing advances in very new directions where the behaviors and likely outcomes are quite unknown, that is a very different situation. Then, just like in education, the researchers are struggling to figure out what factors are important and how to control or adequately measure what they hope are the relevant factors. There is the messiness of having many, many quantities that "might" be important, and the experimental results obtained in such circumstances, more often than not, turn out to be irreproducible (or to put it less technically, "wrong"). Everything is much more complicated before you have figured it all out, and the results are far less precise. It is also much messier than usually presented after the fact.

Physics and chemistry are quite mature sciences, and so most of the research that gets presented in textbooks, classes, and even in the media has all this messiness understood and cleaned up. It is much like seeing a child only through official portraits taken after they are grown, cleaned up, and dressed in formal wear, rather than seeing them as they really were, climbing trees and splashing through mud puddles. However, that clean situation comes in an area of physics research only after long exploration. Many of the complications that at one time were tremendous intellectual challenges have been sorted out long ago and are now largely forgotten.

Only a small fraction of what I did in my 30 years of physics research falls in this cutting-edge "messy and complex" category, and my fraction is probably larger than that for most physicists or chemists. In contrast, much of modern biology is working in much less well-studied and less understood areas of research, and there, the results tend to be far less incremental and correspondingly less reproducible, because the complexities of the systems involved have not yet been so well studied and understood. In general, education research is more like biology research. In some respects, I found these differences make education research more fun and in some ways "easier" than my physics research. Fun and easier in the sense there is so much unplowed ground, so many unanswered questions, and so many potential experiments and possible surprises. Of course, in other respects, it is harder; for example, we know a lot more about the contextual influences on the behavior of atoms than on students and, hence, what contextual elements do and do not have to be controlled in designing experiments. Also, atoms do not require institutional review board approval and consent forms.

In cutting-edge research in the hard sciences, there are always things that one wants to know or measure or control that one cannot, just as there are in education research. I have found that the basic intellectual challenges of designing and executing good cutting-edge research that meet these criteria of predictive power are much the same across fields. There is an enormous number of ways to get the wrong answer by overlooking some relevant variable, and the mark of a good researcher is to recognize, with limited information, which variables are relevant. Figuring out what to measure, and how well to measure it, is critical in all fields. The best researchers do their best to measure and/or control those variables that matter, and they design experiments and subsequent tests to ensure that those that can be neither measured nor fully controlled are unlikely to change the results in significant ways.

On the other hand, it is possible to be too careful. If a researcher is determined to examine, measure, and carefully control every conceivable variable, he or she will be a failure, be it in hard sciences or education, because he or she will never finish

[1]Stanford University, Stanford, CA

Educational Researcher, Vol. 43 No. 1, pp. 12–14
DOI: 10.3102/0013189X13520294
© 2014 AERA. http://er.aera.net

anything. The measure of a great researcher is the one who understands how to do enough, and only just enough, to obtain important results that are reproducible and have adequate predictive power to advance the field. Although the specifics for how to do this are different between physics and education, the basic methods are much the same. One must have a complex model of the system that is used to analyze which factors are important or not and sophisticated criteria for testing one's model and biases.

The difference in how extensively the different research areas have been studied also leads to a difference in terms of what types of research are publishable. This difference in publication standards contributes to the inaccurate perception of the nature of hard-science research. Although there is descriptive, hypothesis-generating research/observations carried out in all fields, in fields like physics or chemistry, such work is seldom considered publishable until it is followed up by quantitative controlled experiments, typically with proposed mechanisms and explanations. Many areas of both biology and education research are similar to where many areas of chemistry and physics were 100 to 150 years ago, in that descriptive observations that generate new hypotheses for basic models of phenomena are recognized as valuable and necessary and hence publishable as a stand-alone results. They are seen as necessary precursors to more extensive controlled experiments that may require major further investments to carry out.

I have argued that the similarities between research in education and research in the hard sciences are greater than usually recognized, largely because of a mischaracterization of the latter. In both cases, a basic standard for research should be the predictive power of the results; and in both cases, the underlying basic intellectual challenges in experimental design, and reasons for flawed research, are much the same.

AUTHOR

CARL E. WIEMAN is a professor in both Department of Physics and the Graduate School of Education at Stanford University, Stanford, CA 94305; *cwieman@stanford.edu*. His research has focused on atomic physics, problem solving and attitudes about science, and the effect of different pedagogies on learning science. In 2001, he won the Nobel Prize in Physics, along with Eric Allin Cornell and Wolfgang Ketterle, for fundamental studies of the Bose-Einstein condensate.

SPECIAL SECTION

Why Understanding Science Matters: The IES Research Guidelines as a Case in Point

John L. Rudolph[1]

The author outlines the rise of a hard-science model advocated by the Institute for Education Sciences, including the application of research and development approaches to education following the Second World War, and describes the attraction of these hard-science approaches for education policymakers. He notes that in the face of complex and persistent educational problems, these approaches seem to promise objective results, uniform solutions, and standardized interventions less prone to ideological distortion. He argues that this particular view of science, however, represents only a narrow slice of the myriad intellectual, social, and cultural practices that fall under the rubric of science and ignores a good deal of the contextual nuance of educational phenomenon. The author highlights the consequences of adopting a narrow vision of science in educational policy, including the marginalization of swathes of research, and the constraint of educational activities to make them more amenable to experimental research.

Keywords: educational policy; educational reform; research methodology; research utilization

Within the science education research community, there have been persistent calls for greater attention to questions in the policy domain. What comes to mind typically is work on *science education* policy, that is, looking at issues of science, technology, engineering, and mathematics recruitment or how statewide science assessments might reduce the achievement gap, and so on. But a more neglected line of inquiry centers on the manner in which science teaching relates to the formation of public policy more generally. The focus here is on how our collective understanding of what science is and how it works shapes the policy decisions we make that, in turn, have consequences for how we live. The research funding guidelines from the Institute for Education Sciences (IES; 2013) provide an object lesson of the way in which our perceptions of science matter in this regard. The language of scientific rigor and references to reliable intervention and progress betray just the sort of public misunderstanding of science that has implications both for education research as a field and, more importantly, for the classroom experiences that are likely to be developed and implemented for children as a result.

In reading the text from the IES call for proposals, the effort to apply a particular vision of science is readily apparent. The fact that education, as the guidelines note, "has always produced new ideas, new innovations, and new approaches" is cast as part of the long, unproductive history of educational practice that ebbs and flows without clear direction. What is needed, according to the authors, are things like "appropriate empirical evaluation" that can identify those things that "are in fact improvements," that will in turn "contribute to the bigger picture of scientific knowledge" (p. 11). From this one short paragraph a reader can easily sense the frustration of policymakers who appear to be striving for a model of research that will finally break the cycle of fads and fashion and generate hard, reliable knowledge that will ensure reproducible results across all classroom settings. This desire is certainly understandable; who, after all, would argue against tangible progress in our knowledge about how students learn?

The hard-science research model advanced by IES, though, is far from a new, game-changing innovation. As far back as the 1950s (as many education researchers well know), federal efforts to reshape education drew on research practices and organizational approaches from the natural sciences, where instrumental success and cumulative progress are the norms. There were, of course, the National Science Foundation–funded curriculum reform projects, first in the sciences but quickly spreading to all academic subject areas, that borrowed heavily from the research and development methods pioneered during World War II. In applying the R&D approaches that were so spectacularly successful in the building of weapon systems during the war to the problems of education, the directors of the curriculum projects

[1]University of Wisconsin, Madison, WI

Educational Researcher, Vol. 43 No. 1, pp. 15–18
DOI: 10.3102/0013189X13520292
© 2014 AERA. http://er.aera.net

(physical scientists being the most prominent of the group) sought to achieve similar levels of success. The early education research centers, such as the Learning Research and Development Center at Pittsburgh and the Wisconsin Center for Educational Research, were in the same way assembled following the institutional models set by the national science laboratories and centers that proliferated in the United States in the mid-20th century.[1]

More recently, we have individuals like the Nobel Prize–winning physicist–turned–education researcher, Carl Wieman (2007), who has argued that efforts to produce effective teaching at scalable levels can be had only by applying "practices that are essential components of scientific research" (p. 10). Such practices, he notes, "explain why science has progressed at such a remarkable pace in the modern world" (Wieman, 2007, p. 10). If only education research, he seems to suggest, were able to progress in a similar way. The research funding preferences of IES follow this familiar pattern. Early on when research guidelines were being developed, IES leaders explicitly drew on experimental models from the fields of medicine and agriculture, where randomized controlled trials were the standard (see, e.g., U.S. Department of Education, 2003).

The allure of these scientific research models is obvious. In the face of complex and persistent educational problems, they seem to promise objective results, uniform solutions, and standardized interventions less prone to ideological distortion that will actually "work" in our nation's classrooms. Outcomes such as these make it easier to argue that the tax dollars going to IES are being spent in a wise and efficient manner. Moreover, these experimental research models conform closely to conceptions of science widely held by the public. "Science" for the average citizen typically entails some activity centered around experimentation whereby a hypothesis or conjecture is demonstrated to be true (or false) with absolute certainty, often revealing in the process knowledge of how this or that corner of nature "really works" (Lederman, 1992). From a policy perspective, it is easy to garner support for methodologies and approaches that align with the dominant perception of what "real" science is. Science, at least this particular version of it, possesses a level of cultural authority that is unmatched in modern society; it should come as no surprise that conclusive demonstration, or experimental confirmation, carries significant weight with the general public (Gauchat, 2012; Shapin, 2007; Toumey, 1996).

This particular view of science, however, represents only a narrow slice of the myriad intellectual, social, and cultural practices that fall under the rubric of science more broadly considered. This is true even if we limit ourselves to the natural sciences. As scholars from the field of science studies have demonstrated in recent years, science is far from a single, unified enterprise or endeavor. Rather, the work of scientists is distributed among a diverse number of smaller research communities, each of which organically fashions its own set of methods, standards of evidence, types of representation, forms of argumentation, social and institutional arrangements, and the like depending on both the nature of the phenomena being studied and the questions deemed worthy of exploring. That is to say that the methods of inquiry are highly contextual, contingent, and emergent over time. As such, it should be obvious that many of these methods fall outside the narrow band of those recognized as experimental. This however, makes them no less scientific.[2]

If we accept the fact that our research methods are dependent on what it is we are seeking to understand, it makes sense, then, to ask ourselves whether the phenomena of teaching and learning are best examined using an experimental approach. Let me state up front that I am certainly not against the use of "rigorous" methods that can ensure reliable and reproducible outcomes. It would go against common sense to devote resources to any line of work that does not eventually result in knowledge that enables us to make decisions, create environments, or interact with others in ways that align with our intentions. Empirical research, in the end, is all about understanding how things work so that we might have them work to our advantage at some time in the future.

But, while we all likely share a commitment to research that enables us to understand the results of our actions, the kind of knowledge we are able to produce is clearly dependent on the phenomena under study. Consider the differences between targets of interest in the natural sciences compared to education. In the former, one finds physical or biological systems, for example, that are relatively simple, capable of isolation in a laboratory setting, manipulable in real time, and (even more important) invariant—that is, they operate following rules that remain constant over time. In the field of education, however, researchers are trying to understand things that are far more complex: the way students learn from text, lecture, visual representations, or combinations thereof. They seek to find out, in addition, how that learning is shaped by prior knowledge and experience and by interactions with teachers and student peers. Complicating the picture is the fact that the object of study in education research is a knowing participant who can resist, cooperate, or simply not engage in the instruction being observed (a characteristic not typically found among subatomic particles or even most mammals).

Context matters as well in ways that are irrelevant to physical systems. It is one thing to limit study to what happens within a classroom, but classrooms exist in a broader matrix of institutions, political systems, and cultures. Whether and what learning takes place are highly contingent on everything from immediate and long-term educational goals to local and national politics, considerations of the global economy, and the allocation of resources (again, to highlight only a handful of factors). And, if this level of complexity were not enough, I would add that all of this changes over time. What we might count as learning today—and certainly what society deems *worth* learning—is not likely to be the same 20 years from now. This fact seriously complicates any effort to establish some record of cumulative knowledge or progress. Clearly, teaching and learning, as they happen in classrooms and lecture halls all over the world—and through time—are far different phenomena from those studied in controlled laboratory settings or even in field studies of naturally occurring plant or animal populations. Education, as an empirical phenomenon, is just the sort of context-sensitive, dynamically responsive complex system that philosopher of science Sandra Mitchell (2009) argues requires research methods that are locally applied, tolerant of uncertainty, and pragmatically adopted to meet particular social ends at a given point in time—methods that go well beyond randomized controlled trials.

There are without a doubt aspects of learning and educational practice that are amenable to experimental and quasi-experimental study, and with some questions, real progress can indeed be made. But we should rightly be concerned when policymakers (supported by a public operating with an incomplete understanding of what science is and how it works) seek to push particular research models and methodological approaches in a misguided attempt to secure knowledge outcomes (reliable, predictable, uniform, etc.) that are unlikely to be obtained given the nature of the activities and enterprise in question. The consequences of such actions should be carefully considered. The most immediate (and self-interested) concern centers on the allocation of resources. When government agencies distribute research money based on what counts as legitimate research, there are naturally winners and losers. Drawing on some definition of "science" (in this case, a methodological one) in making these distinctions is not uncommon, historically speaking. These are just the kind of things that fall into what the sociologist of science Thomas Gieryn (1999) has termed "boundary work." In such work, selective representations of science are typically invoked that reward some and deprive others. But, while worries about diminished federal funding for certain types of research are real for those on the outside looking in, there are at least two potential outcomes of far greater significance to the broader public.

First, if we accept the fact that physical and educational systems operate at fundamentally different levels of complexity, then any attempt to move education toward a hard-science research model in which we try to measure and/or control key variables is bound to fail at the broad-based societal levels that really matter. Put simply, the things we are able to measure in such a system rarely conform to the learning outcomes we most highly value as a society. Moreover, if we hold up an experimental model as our standard, a good deal of other education research (that which relies on more descriptive or qualitative approaches) will likely be viewed as deficient by comparison—a result that has the potential to discredit education research in general. One unhappy consequence might be that educational policymakers and administrators give up any hope that research can meaningfully inform decisions about how best to educate our students and begin to rely instead on folk theories, personal bias, or political ideology to guide their actions.[3]

The second possibility relates more directly to the power the experimental model has to shape the world outside the laboratory. If we believe that progress can be made in education only if we embrace something similar to the experimental models of physics, medicine, or agriculture, then to get these models to work in the real world requires us to constrain our educational activities so that they more closely match the research models we use to generate knowledge. We would need to, in other words, make the naturally occurring system more like the experimental system, a change that would require the simplification of natural learning environments. This might entail things like the standardization of learning goals, scripted instructional plans, the reduction of individual and institutional autonomy, and so on. Only by extending the conditions of the laboratory to the settings we seek to improve can the power of the knowledge p duced in that context be realized.[4]

It does not escape my attention that we are already living w both these undesirable outcomes in one form or another. M work clearly needs to be done to better understand how resear policy, and practice might be most productively integrated for t broader goal of social improvement. Helping the public and p cymakers (who, after all, in our democratic political system members of the public) understand just how various scienti research practices work to generate reliable knowledge seems to a logical first step. It seems that the field of science education both in research and practice—has much work to do on this fro

NOTES

[1]For a more complete historical overview, see Rudolph (2002), Li (2003), and Westwick (2003).

[2]On this point, see Galison and Stump (1996), Cartwright (199 Longino (2002), and Pickering (1995), among others.

[3]This is close to the point Mitchell offers more generally in her 20 book, *Unsimple Truths*.

[4]This consequence of the rise of science and laboratory work in world is described in Latour (1983).

REFERENCES

Cartwright, N. (1999). *The dappled world: A study of the boundarie science*. New York, NY: Cambridge University Press.

Galison, P., & Stump, D. J. (Eds.). (1996). *The disunity of science: Bou aries, contexts, and power*. Stanford, CA: Stanford University Pres

Gauchat, G. (2012). Politicization of science in the public sphere: study of public trust in the United States, 1974 to 2010. *Ameri Sociological Review, 77*, 167–187.

Gieryn, T. F. (1999). *Cultural boundaries of science: Credibility on line*. Chicago, IL: University of Chicago Press.

IES. (2013). *Institute for Education Sciences: Request for applicatio Education research grants*. U.S. Department of Education. Washi ton, DC: Government Printing Office

Latour, B. (1983). Give me a laboratory and I will raise the world. In Knorr-Cetina & M. Mulkay (Eds.), *Science observed: Perspectives the social study of science* (pp. 141–169). London, UK: Sage.

Lederman, N. G. (1992). Students' and teachers' conceptions of nature of science: A review of the research. *Journal of Research Science Teaching, 29*, 331–359.

Light, J. S. (2003). *From warfare to welfare: Defense intellectuals a urban problems in cold war America*. Baltimore, MD: Jo Hopkins University Press.

Longino, H. E. (2002). *The fate of knowledge*. Princeton, NJ: Princet University Press.

Mitchell, S. D. (2009). *Unsimple truths: Science, complexity, and poli* Chicago, IL: University of Chicago Press.

Pickering, A. (1995). *The mangle of practice: Time, agency, and scien* Chicago, IL: University of Chicago Press.

Rudolph, J. L. (2002). From world war to Woods Hole: The use wartime research models for curriculum reform. *Teachers Coll Record, 104*, 212–241.

Shapin, S. (2007). Science and the modern world. In E. Hackett, Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *Handbook science and technology studies* (3rd ed., pp. 433–448). Cambrid MA: MIT Press.

Toumey, C. P. (1996). *Conjuring science: Scientific symbols and cultu meanings in American life*. New Brunswick, NJ: Rutgers Univers Press.

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Government Printing Office.
Westwick, P. J. (2003). *The national labs: Science in an American system, 1947–1974*. Cambridge, MA: Harvard University Press.
Wieman, C. (2007, September/October). Why not try a scientific approach to science education? *Change*, pp. 9–15.

AUTHOR

JOHN L. RUDOLPH is a professor of science education in the Departments of Curriculum and Instruction, Educational Policy Studies, and History of Science at the University of Wisconsin-Madison, 225 N. Mills St, Madison, WI 53706; *jlrudolp@wisc.edu*. His research focuses on the history of science education in the United States and public understanding of science.

Relevance to Practice as a Criterion for Rigor

Kris D. Gutiérrez[1*] and William R. Penuel[1]

The authors argue for a reconceptualization of rigor that requires sustained, direct, and systematic documentation of what takes place inside programs to document how students and teachers change and adapt interventions in interactions with each other in relation to their dynamic local contexts. Building on promising new programs at the Institute of Education Sciences, they call for the formulation of collaborative research standards that must require researchers to provide evidence that they have engaged in a process to surface and negotiate the focus of their joint work, and to document the ways participation in this process was structured to include district and school leaders, teachers, parents, community stakeholders, and, wherever possible, children and youth. They close by describing how this new criterion—"relevance to practice"—can ensure the longevity and efficacy of educational research.

Keywords: educational policy; educational reform; research methodology; research utilization

When Congress passed and the president signed the Education Sciences Reform Act in 2002, it called for scientifically based research that would "apply rigorous, systematic, and objective methodology to obtain reliable and valid knowledge relevant to education activities and programs" (Pub. L. No. 107-279, p. 116). That same year, the National Research Council (2002) produced a report and *Educational Researcher* published a related article, "Scientific Culture and Education Research" (Feuer, Towne, & Shavelson, 2002), written by several of the report's authors. There is much with which we agree in both of the publications. However, there was and still remains a concern from the field about the narrow set of criteria used to define rigor. Erickson and Gutiérrez (2002) questioned the publications' call for a "scientific culture" that prescribed and relied primarily on "gold standard" random assignment studies of program effects as the remedy for the failures of education research to offer credible guidance for policy and practice. As we (Erickson & Gutiérrez, 2002) argued then, *rigor* in studies that aim to draw causal inferences about policies, programs, and practices requires in-depth qualitative research. In particular, scientifically rigorous research on what works in education requires sustained, direct, and systematic documentation of what takes place inside programs to document not only "what happens" (cf. National Research Council, 2002) but also how students and teachers change and adapt interventions in interactions with each other in relation to their dynamic local contexts.

Today, we see even greater need for the field to take up broader questions about what works to include questions about a study's relevance to transforming practice. Studies of "what works" should be concerned with the specific mechanisms by which outcomes for teachers and students are accomplished within specific structural and ecological circumstances. Rigorous research on "what works" also must take up seriously the questions, "Who does the design and why?" (Engeström, 2011, p. 3), "How can practice and research inform one another?" "What are the unintended consequences of change?" and, importantly, "Who benefits?" (Erickson & Gutiérrez, 2002; Gutiérrez & Vossoughi, 2010; O'Connor & Penuel, 2010). For us, consequential research on meaningful and equitable educational change requires a focus on persistent problems of practice, examined in their context of development, with attention to ecological resources and constraints, including why, how, and under what conditions programs and policies work.

New Programs at the Institute of Education Sciences

Recently, the Institute of Education Sciences (IES) within the U.S. Department of Education created new promising programs of research that address important problems of practice and provide the time to build relevance into the design of research and development projects.[1] These and other IES studies have begun to incorporate more direct observation into research on policies and programs, in ways that have generated productive insights

[1]University of Colorado, Boulder, CO
*Kris D. Gutiérrez is a current member and vice chair of the National Board of Educational Sciences.

Educational Researcher, Vol. 43 No. 1, pp. 19–23
DOI: 10.3102/0013189X13520289
© 2014 AERA. http://er.aera.net

into what works where, when, why, and for whom. For example, Continuous Improvement Research in Education grants fund well-established partnerships to adapt, study, and iteratively refine tested interventions for improving teaching and learning. The Researcher-Practitioner Partnerships in Education Research program provides funding for developing new research-practice partnerships. As director John Easton (2013) recently noted, through these and other initiatives, IES is "promoting research use, but not in a unidirectional 'research to practice' sense but in a more reciprocal 'practice to research' pathway" (Easton, 2013, p. 18). This new research program calls for "empirical tinkering" (Morris & Hiebert, 2011) in which partners collaborate "to fine tune programs, interventions or regimens of activities through iterative processes that rely heavily on measurement, quick studies and refinement" (Easton, 2013, p. 18). This approach represents a significant move toward meeting IES's charge to support research that is "relevant to education practice and policy."

For these new programs to be successful, relevance to practice must be an explicit criterion for judging the quality of research proposals. For example, there should be documentation that the problem of focus is perceived by multiple stakeholders to be significant, persistent, and worthy of investigation. Standards must also require researchers to provide evidence that they have engaged in a process to surface and negotiate the focus of their joint work, and to document the ways participation in this process was structured to include district and school leaders, teachers, parents, community stakeholders, and, wherever possible, children and youth.

Developing such evidence of relevance for a research proposal is not likely to be easy. The problems that researchers initially think important to address are not likely to be the same ones that diverse education stakeholders perceive as important. What is needed are specific methodologies for bringing relevant stakeholders together and *deliberating* about the problems that can and should be addressed through research and development projects. The process is time-intensive, and it must begin well before the final weeks before researchers submit proposals to IES.

Interventions as Contested Spaces

Educational systems have multiple layers of infrastructure that have accumulated over time and that must be engaged directly if they are to support, rather than obstruct, transformation (Penuel & Spillane, in press). As Engeström (2011) reminds us, interventions take place in complex and multilayered activity systems rife with recurring problems that are conceptualized as contradictions inherent in the structuring of the system. Interventions themselves are contested spaces, filled with tensions and resistance from a range of stakeholders. Supporting and engaging more diverse stakeholder engagement in defining the focus of research and development will require researchers and reviewers to rethink the nature of educational interventions. In contrast to closed or top-down notions of designed collaborations, the approaches to interventions we discuss here are systems that are subject to revision, disruptions, and contradictions (Gutiérrez & Vossoughi, 2010). This dialectic of "resistance and accommodation" in practice is what Pickering terms "the mangle of practice" (Pickering, 2010, p. 10).

We want to generate and support robust teaching and learning practices, but we want to do so by addressing this dialectic and redesigning functional systems that open up new pathways and social futures for youth, particularly, youth from nondominant communities (Gutiérrez, 2008; O'Connor & Allen, 2010). An emphasis on what is happening in the day-to-day life of participants in those systems helps make visible the structural and historically existing contradictions inherent in complex activity systems, like schools, and refocuses our analytical lens and objects of design. Studying the "social life of interventions" moves us away from imagining interventions as fixed packages of strategies with readily measurable outcomes and toward more open-ended social or socially embedded experiments that involve ongoing mutual engagement.

As education researchers committed to studying persistent problems of practice, we step into the messiness and uncertainty that problem-oriented work and rigorous scientific inquiry requires. Following Erickson (2006, p. 225), rigorous and consequential study of the efficacy of educational interventions involves sustained firsthand observation, sharing in the action and cognition of practitioners (Gutiérrez & Vossoughi, 2010). Studying "side by side" with research partners jointly engaged in work to transform systems is more likely to produce more sensitive and robust measurement and ecologically valid accounts of cultural production and institutional change.

Need for New Approaches to Research and Development

The success of projects within new programs at IES will also depend on the development of new approaches to research and development as we describe above. These approaches must encompass participatory design tools and practices for deliberating about and negotiating problems of practice and for engaging in iterative design. They must also make use of findings from implementation research to improve interventions.

There are extant models of this kind of formative intervention research in the field. The "change laboratory," for example, involves the collaboration of practitioners and researchers around an important and consequential problem of practice within an existing activity system (Cole & Engeström, 2006; Engeström, 2008; Engeström, Virkkunen, Helle, Pihlaja, & Poikela, 1996). Within this approach, researchers observe everyday practices and conduct interviews with stakeholders to identify contradictions within and across the various levels of the system under study. The researcher, then, is a collaborative partner and a reflective "observant participant" who helps make visible the practices, meanings, and contradictions that often become invisible to those closest to the action (Erickson, 1986, p. 157; Gutiérrez & Vossoughi, 2010). The researcher is then positioned to re-present what is learned to a design team as part of an iterative process.

A number of scholars are hard at work developing and testing other approaches to collaborative research and development that are consistent with the principles of change laboratories. Some of these approaches are place-based efforts in school districts or communities that engage in collaborative design to improve teaching and learning at scale (Cobb & Jackson, 2012; Penuel, Fishman, Cheng, & Sabelli, 2011). Other approaches highlight the value of engaging in rapid cycles of iterative design and research to improve practice across networks of geographically dispersed institutions (Bryk, Gomez, & Grunow, 2011). Still others engage historically excluded communities in efforts to reclaim connections between cultural and disciplinary forms of learning (Bang, Medin, Washinawatok, & Chapman, 2010), and transforming practice across multiple systems of activity, while attending to people's history of involvement in practices (Vossoughi & Gutiérrez, in press).

These models do not require researchers to specify ahead of time all the elements of an intervention, since practitioners participate in design, and implementation data inform an iterative design process that often transforms interventions. It is important to ask, What is a partnership if the research plan is fully predefined by researchers? And how might we address the emergent tension between the importance of starting with a "germ cell"—an emergent model that is examined experimentally and analytically and elaborated across iterations with and by participants—and the push for a fully developed design?

This is a particularly relevant dilemma for education researchers who rely on extramural funding to support their empirical work. Addressing this problem requires far more elaboration than we can provide here. However, we believe that there is some middle ground that funding agencies might consider, as we understand that review panels need criteria to ensure rigorous, systematic examination of an educational problem with a probability of success in its execution. Our basic argument here is that funders could demand more attention to the process side, asking researchers to address what research would look like and what methods and co-designing processes are relevant to the study at hand.

Generalization in Theory as Organizing for Relevance to Practice

IES recently indicated its intent to support work across the agency that yields knowledge about the effects of specific interventions and that also contributes to "the bigger picture of scientific knowledge and theory of learning, instruction, and education systems" (IES, 2013, p. 11). For us, this represents an important advance for the agency, because greater weight is given to the importance of *generalization* from research findings. The challenges to generalization in education research are many (Berliner, 2002); here we highlight two challenges that strike us as particularly challenging for IES, given the kinds of projects the agency has funded in the past.

First is that efficacy and even effectiveness trials engineer contexts that are not easily replicable without sustained funding beyond the life of a research project. Research projects focused on curriculum design typically provide teachers with materials that must be replenished and updated, with professional development, and grant-supported incentives for implementation. When the research ends, teachers may discontinue programs found to be effective with a wide variety of students. What is more, the teachers who discontinue use may disproportionally serve students from nondominant backgrounds who stand to benefit from these programs.

An instructive example is the SimCalc study, funded through the Interagency Education Research Initiative at the National Science Foundation. SimCalc aims to provide middle school students with access to key foundational ideas related to the mathematics of change (Roschelle, Kaput, & Stroup, 2000). Evidence from a large-scale randomized controlled trial showed that students of diverse backgrounds can learn from SimCalc (Roschelle, Pierson, et al., 2010; Roschelle, Shechtman, et al., 2010). Analyses of the generalizability of the treatment effects indicated, too, that findings could generalize to most counties in the state where the research took place (Roschelle, Hedges, Tipton, & Shechtman, 2012).

When the initial scale-up study concluded, however, not all students of teacher participants in the study continued to have access to SimCalc MathWorlds, even though the teachers still had the materials (Fishman, Penuel, Hegedus, & Roschelle, 2011). About half the teachers used the materials the year following the research. Some teachers discontinued use of the materials because of perceived policy pressures from within their school or district to adopt different materials and approaches to teaching mathematics. Low-income students were less likely to have access to materials because their teachers discontinued use of the materials. This particular finding is not unique to SimCalc; the story echoes decades of policy and program implementation research.

Our point is neither to criticize programs like SimCalc nor to diminish the potential value of research on the effects of programs like SimCalc. Evidence indicates that SimCalc is a potentially powerful program, and there is strong evidence that a wide range of students can benefit from it. In our view, sustaining nearly any robust intervention will require ongoing work, work of the kind that went into making the SimCalc study a success and the program a good temporary fit to the goals of teachers for their students in the study. This includes work to craft professional development, curriculum, and technology into a coherent "curricular activity system" that could be used in a wide variety of classrooms, work to align this system to standards, and work to support implementation of the system in the field (Roschelle, Knudsen, & Hegedus, 2010).

The work of mutual adjustment of powerful interventions and local contexts does not end when the research ends, but sustaining an intervention requires uptake by schools and districts (Coburn, 2003). For us, we define the *generalizability* of findings and theories developed through research as contingent on the uptake of research by local actors who must sustain programs. Local actors' productive adaptation of interventions or use of theories from research and the documentation of the work they must do to sustain change are important sources of evidence for generalizability.

This uptake may include researchers as part of the activity, or it may be sustained entirely by practitioners. Researchers can continue to partner with schools and districts to adapt and test which supports are most needed (Penuel et al., 2011). Alternately, professional communities inside schools and districts can sustain programs through frequent, deep interaction, provided they have sufficient access to expertise relevant to program implementation (Coburn, Russell, Kaufman, & Stein, 2012).

We also need to understand the limits of generalizability by answering questions of what works, under what conditions, and for whom. The challenge is that the effects of any instructional program as estimated in an efficacy trial are likely to vary widely, as is implementation, requiring identifying and mastering variation. In this connection, "mastering variation" does not mean attempting to minimize variation in implementation but, rather, to learn from of productive adaptations teachers make with learners from variety of backgrounds. It means developing and testing supports to broaden capacity of teachers to make such productive adaptations themselves, to increase the effectiveness of programs, and to promote equity.

This requires a shift in focus of research and development efforts, away from innovations designed to be implemented with fidelity in a single context and toward cross-setting interventions that leverage diversity (rather than viewing it as a deficit). It also suggests the need to focus some research and development projects on the design of new organizational routines and infrastructures for improvement (Bryk et al., 2011; Penuel & Spillane, in press). It also implies the need for efficacy and effectiveness research that addresses how to *make programs work* under a wide range of circumstances and for all groups (Bryk, 2009; Bryk et al., 2011).

With Engeström and Sannino (2010), we view the ultimate benchmark for any program or learning theory is how well it helps us to organize conditions for learning in a way that takes up present and future problems society faces. Making relevance to practice a key criterion of rigor is an important step toward more equitable and consequential research. This is a high standard, but it is not just up to researchers to accomplish. We see the aim of intervention research as facilitating participants in activity to deal with the historically accumulated tensions and contradictions of the systems within which they work in order to transform the activity of teaching and learning.

NOTE

We would like to thank our colleagues Ben Kirshner, Kevin O'Connor, and Susan Jurow, Learning Sciences, University of Colorado Boulder, for their thoughtful comments.

REFERENCES

Bang, M., Medin, D., Washinawatok, K., & Chapman, S. (2010). Innovations in culturally based science education through partnerships and community. In M. S. Khine & M. I. Saleh (Eds.), *New science of learning: Cognition, computers, and collaboration in education* (pp. 569–592). New York, NY: Springer.

Berliner, D. C. (2002). Comment: Educational research. The hardest science of all. *Educational Researcher, 31*(8), 18–20.

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

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M. Hallinan (Ed.), *Frontiers in sociology of education* (pp. 127–162). Dordrecht, Netherlands: Verlag.

Cobb, P. A., & Jackson, K. (2012). Analyzing educational policies: A learning design perspective. *Journal of the Learning Sciences, 21*, 487–521.

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

Coburn, C. E., Russell, J. L., Kaufman, J. H., & Stein, M. K. (2012). Supporting sustainability: Teachers' advice networks and ambitious instructional reform. *American Journal of Education, 119*(1), 137–182.

Cole, M., & Engeström, Y. (2006). Cultural-historical approaches for designing for development. In J. Valsiner & A. Rosa (Eds.), *The Cambridge handbook of sociocultural psychology* (484–507). Cambridge, UK: Cambridge University Press.

Easton, J. (2013, June). *Using measurement as leverage between developmental research and education practice*. Talk given at the Center for the Advanced Study of Teaching and Learning, Curry School of Education, University of Virginia, Charlottesville.

Engeström, Y. (1996). Development as breaking away and opening up: A challenge to Vygotsky and Piaget. *Swiss Journal of Psychology, 55*, 126–132.

Engeström, Y. (2008). *From teams to knots: Activity theoretical studies of collaboration and learning at work*. New York, NY: Cambridge University Press.

Engeström, Y. (2011). From design experiments to formative interventions. *Theory & Psychology, 21*(5), 598–628.

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review, 5*, 1–24.

Engeström, Y., Virkkunen, J., Helle, M., Pihlaja, J., & Poikela, R. (1996). Change laboratory as a tool for transforming work. *Lifelong Learning in Europe, 1*(2), 10–17.

Erickson, F. (2006). Studying side by side: Collaborative action ethnography in educational research. In G. Spindler & L. Hammond (Eds.), *Innovations in educational ethnography: Theory, methods and results* (pp. 235–257). Mahwah, NJ: Lawrence Erlbaum.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119–161). New York, NY: Macmillan.

Erickson, F., & Gutiérrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher, 31*(8), 21–24.

Feuer, M., Towne, L., & Shavelson, R. (2002). Scientific culture and education research. *Educational Researcher, 31*, 4–14.

Fishman, B. J., Penuel, W. R., Hegedus, S., & Roschelle, J. (2011). What happens when the research ends? Factors related to the sustainability of a technology-infused mathematics curriculum. *Journal of Computers in Mathematics and Science Teaching, 30*(4), 329–353.

Gutiérrez, K. (2008). Developing a sociocritical literacy in the third space. *Reading Research Quarterly, 43*(2), 148–164.

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.

Morris, A., & Hiebert, J. (2011). Creating shared instructional products: An alternative approach to improving teaching. *Educational Researcher, 40*(1), 5–14.

National Research Council. (2002). *Scientific research in education*. Washington, DC: National Academy Press.

O'Connor, K., & Allen, A. (2010). Learning as the organizing of social futures. In W. Penuel & K. O'Connor (Eds.), *Yearbook of the National Society for the Study of Education* (Vol. 108, pp. 160–175). New York, NY: Teachers College Press.

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: Yearbook of the National Society for the Study of Education* (Vol. 109, No. 1, pp. 1–16). New York, NY: Teachers College Press.

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

Penuel, W. R., & Spillane, J. P. (in press). Learning sciences and policy design and implementation: Key concepts and tools for collaborative engagement. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed.). Cambridge, UK: Cambridge University Press.

Pickering, A. (2010). *The mangle of practice: Time, agency, and science*. Chicago, IL: University of Chicago Press.

Roschelle, J., Hedges, L. V., Tipton, E., & Shechtman, N. (2012). *Generalizability of integrating dynamic representation technology with curriculum and teaching to enhance understanding of mathematics*. Unpublished manuscript.

Roschelle, J., Knudsen, J., & Hegedus, S. J. (2010). From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning. In M. J. Jacobson & P. Reimann (Eds.), *Designs for learning environments of the future: International perspectives from the learning sciences* (pp. 233–262). New York, NY: Springer.

Roschelle, J., Pierson, J., Empson, S., Shechtman, N., Dunn, M., & Tatar, D. (2010). Equity in scaling up SimCalc: Investigating differences in student learning and classroom implementation. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences* (Vol. 1, pp. 333–340). Chicago, IL: International Society of the Learning Sciences.

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

Vossoughi, S., & Gutiérrez, K. (in press). Toward a multi-sited ethnographic sensibility. In J. Vadeboncoeur (Ed.), *NSEE yearbook*. New York, NY: Teachers College Press.

AUTHORS

KRIS D. GUTIÉRREZ holds the Inaugural Provost's Chair and is professor of learning sciences and literacy in the School of Education at the University of Colorado Boulder, Education Building 124 249 UCB, Boulder, CO 80309-0249; *kris.gutierrez@colorado.edu*. Her research focuses on learning in designed learning environments, with particular attention to students from nondominant communities and English learners.

WILLIAM R. PENUEL is a professor of educational psychology and learning sciences at the University of Colorado, School of Education, UCB 249, Boulder, CO 80309; *william.penuel@colorado.edu*. His research focuses on teacher learning and organizational processes that shape the implementation of educational policies, school curricula, and after-school programs.