Can Qualitative Researchers Answer Policymakers' What-Works Question?

Robert Donmoyer

Qualitative Inquiry 2012 18: 662
DOI: 10.1177/1077800412454531

The online version of this article can be found at:
http://qix.sagepub.com/content/18/8/662

Published by:

SAGE

http://www.sagepublications.com

Additional services and information for Qualitative Inquiry can be found at:

Email Alerts: http://qix.sagepub.com/cgi/alerts

Subscriptions: http://qix.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://qix.sagepub.com/content/18/8/662.refs.html

>> Version of Record - Sep 6, 2012

What is This?
Can Qualitative Researchers Answer Policymakers’ What-Works Question?

Robert Donmoyer

Abstract
The article asks whether constructivist qualitative researchers have anything to offer policymakers who expect researchers to tell them what works. The first part of the article addresses philosophical objections to characterizing the social world in cause/effect terms. Specifically, it considers whether it is legitimate for qualitative researchers who claim to be employing a constructivist research paradigm to even attempt to provide the sort of simplified causal explanations that policy makers normally expect. The second part of the article takes a more empirical tack by focusing on three recent evaluation studies in which funders wanted to learn what types of programs they should support to produce desired results. The underlying question in this part of the article is pragmatic: Even if there is no paradigmatic prohibition against attempting to answer policymakers’ what-works question, are constructivist qualitative researchers able to answer policymakers’ bottom-line question in a defensible way?

Keywords
causality in qualitative research, what works, qualitative research and policymaking, research utilization

During the so-called paradigm wars (Gage, 1989) that occurred within educational research and certain other fields during the final quarter of the 20th century, qualitative researchers often suggested that the field’s growing interest in qualitative methods represented not only a shift in methodological preferences but also a new way of thinking about research and social life. They even argued that this new way of thinking represented something that was roughly equivalent to the paradigm revolutions that had occurred intermittently within the physical sciences (Kuhn, 1996, c.1977 c.1966) and that the new paradigm that encouraged the use of qualitative methods was incommensurable with the thinking of traditional quantitative researchers (Lincoln & Guba, 1985).

One of the articulated incommensurable differences involved the notion of causality. The argument went something like this: Traditional quantitative researchers assume that the social world is a world of cause and effect relationships; they also assume that, as in physical science fields such as engineering, these relationships can be discovered through qualitative research and that the empirical knowledge quantitative inquiry produces can be used to predict and even control phenomena. By contrast, qualitative researchers—or at least enlightened qualitative researchers who have embraced the so-called constructivist paradigm—assume that social action is not caused; rather it is constructed through an ongoing process of meaning making that occurs within and across individuals.

This article considers whether the storyline about paradigmatic differences—and, more specifically, the idea that qualitative researchers should eschew cause and effect accounts of social phenomena (a key element in the tale of two incommensurable paradigms)—still makes sense. The focus here is on evaluation studies and similar sorts of applied research. More specifically, this article focuses on research for a policy community that, more often than not, expects researchers to answer, in relatively simple terms, a bottom-line question: What works? The article asks whether qualitative researchers who embrace constructivism have anything to offer these sorts of policymakers.

The first part of this article addresses philosophical—or, to use the nomenclature of the recent past, paradigmatic—objections to characterizing the social world in cause/effect terms. In other words, it considers whether qualitative researchers—especially qualitative researchers who claim to embrace a constructivist perspective—ought to even attempt to provide the sort of simplified causal explanations that policymakers desire. The second part of the article takes a more empirical tack by focusing on three recent evaluation studies in which funders wanted to learn what types of

1University of San Diego, CA, USA

Corresponding Author:
Robert Donmoyer, PhD, University of San Diego, Suite 207, 3940 7th Avenue, San Diego, CA 92103, USA
Email: donmoyer@sandiego.edu

Downloaded from qix.sagepub.com at SAGE Publications on January 10, 2014
programs they should support to produce desired results. The underlying question in this part of the article is pragmatic: Can qualitative researchers answer policymakers’ what-works question in any sort of defensible way?

Philosophical/Paradigmatic Objections

Near the start of the 20th century, Thorndike (1910) wrote the following in the lead article of the inaugural issue of *The Journal of Educational Psychology*:

[A] complete science of psychology would tell every fact about everyone’s intellect and character and behavior, would tell the cause of every change in human nature, would tell the result which every educational force—every act of every person that changed any other or the agent himself [or herself]—would have. It would aid us to use human beings for the world’s welfare with the same surety of result that we now have when we use falling bodies or chemical elements. In proportion as we get such a science we shall become masters [and mistresses] of our own souls as we are now masters of heat and light. *Progress toward such a science is being made.* (p. 6, emphasis added)

For the bulk of the 20th century, most social scientists in most, though certainly not all, social science fields uncritically accepted Thorndike’s vision of what social scientists could and should do. In the process, most social scientists also at least tacitly accepted Thorndike’s assumption that there is no fundamental difference between the social and physical sciences. Consequently, most social scientists, at least in the United States, worked hard to adapt the quantitative research designs used in the physical sciences to the study of social phenomena (Campbell & Stanley, 1955).

During the final decades of the 20th century, however, Lincoln and Guba (1985) argued that social scientists should abandon the search for cause and effect generalizations; instead, they should focus on the meaning that human beings attach to activity in the social world, meaning that human beings, themselves, create through social interaction. The job of social scientists, therefore, is to explicate the meaning that different groups of human beings construct. To engage in this explication process (a process that entails reconstructing the constructions of the human beings being studied), Lincoln and Guba argued, social scientists need to use qualitative research methods.

Neither Lincoln and Guba’s (1985) social-constructivist view of social life and social action nor their endorsement of qualitative methods was entirely new. Both ideas can be traced back to discussions of hermeneutics (see, for example, Dilthey, 1961) during the advent of the social sciences in 19th-century Europe. There also were 20th-century antecedents. In the United States, for example, virtually the entire discipline of cultural anthropology (see, for example, Geertz, 1973) viewed social action in constructivist terms and used qualitative methods to study social life. And even in the normally more quantitative discipline of sociology, certain subgroups of sociologists such as the symbolic interactionists (Blumer, 1969) championed the use of qualitative methods and a social-constructivist conception of social life.

What was new in Lincoln and Guba’s (1985) thinking, however, was Lincoln and Guba’s claim that those who advocated using qualitative methods and focusing on socially constructed meaning rather than on cause and effect relationships were promoting what Kuhn (1996, c. 1970, 1962), in his discussions of conceptual change in the physical sciences, had called a paradigm shift. This claim changed the debate about the form and function of social science research in important ways.

Incommensurability and Causality

Lincoln and Guba initially called the new paradigm they described the naturalistic paradigm; later they renamed it the constructivist paradigm, a name that continues to be used today and that will be used in this article. Traditional quantitative researchers were said to be working within the positivist paradigm.

Consistent with Kuhn’s description of research paradigms in the physical sciences, Lincoln and Guba (1985) assumed that the constructivist paradigm was incommensurable with the positivist paradigm. Lincoln and Guba (1985), for instance, wrote that naturalism/constructivism “is an entirely new paradigm, not reconcilable with the old . . . just as ‘the world is round cannot be added to the idea that the world is flat’” (p. 33).

Among the incommensurable differences articulated by Lincoln and Guba (1985) was each paradigm’s thinking about causality. Lincoln and Guba indicated that the new paradigm rejected the notion of causality on ontological grounds: Although the positivist paradigm assumed that “every action can be explained as the result (effect) of a real cause that precedes the effect temporally (or at least is simultaneous with it),” the new naturalistic/constructivist paradigm assumed that “all entities are in a state of mutual simultaneous shaping so that it is impossible to distinguish causes and effects” (p. 38). The complexity of social life, Lincoln and Guba claimed, could only be captured in the thick description provided by qualitative case study research.

Clearly, if one accepts Lincoln and Guba’s (1985) characterization of the constructivist paradigm, constructivist-oriented qualitative researchers should not even attempt to answer policymakers’ what-works question. Indeed, answering any kind of question about stable cause and effect relationships would be inconsistent with the assumptions of the
constructivist paradigm, as defined by Lincoln and Guba in 1985. In the next section, I will explore whether attempting to answer causal questions, in general, and policymakers’ what-works question, in particular, are, indeed, inconsistent with a constructivist approach to qualitative research.

Is it Appropriate for Constructivist Qualitative Researchers to at Least Attempt to Answer Policymakers’ What-Works Question?

Here I will argue that that there is nothing inherently problematic with constructivists talking in cause and effect terms and attempting to answer policymakers’ what-works question. (Whether constructivist-oriented qualitative researchers are in a position to actually provide viable answers to policymakers’ what-works question is the topic for the second half of this article.) My argument in this section of the article has four parts.

First, I will demonstrate that, prior to Lincoln and Guba (1985), scholars interested in using qualitative methods to explicate meaning did not assume that their meaning-oriented perspective prohibited characterizing social life in cause and effect terms. Second, I will note that, even Kuhn, in later editions of his book The Structure of Scientific Revolutions, clearly indicated that incommensurability is not a synonym for logical incompatibility. Third, I will argue that Lincoln and Guba’s (1985) paradigm talk, itself, is inherently contradictory on the causality question, for once one embraces a constructivist epistemology and, consequently, assumes that reality is inevitably socially constructed, it makes no sense to talk about the way the world really is and to reject cause and effect conceptions of social life on ontological grounds. Finally, I will make an instrumental rather than an ontological case for constructivist qualitative researchers entering the fray and attempting to answer policymakers’ what-works question.

Other qualitative traditions’ views of causality. I will begin by demonstrating what, thus far, has only been asserted: Although an interest in explicating meaning and using qualitative methods to do this can be traced back through certain segments of 20th-century social science to the advent of the social sciences in the 19th century, earlier researchers did not normally see this interest as being logically incompatible with a focus on cause and effect explanations. Dilthey (1961) and other early hermeneutic scholars who made the case for a meaning perspective and the use of qualitative methods during the time that the social sciences were being created in mid-19th-century Europe, for example, never argued that social phenomena could not also be thought of in cause and effect terms. Dilthey’s argument was simply that social scientists who attempted to discover cause and effect generalizations that cut across different contexts would miss the uniqueness of particular times and places and this uniqueness is what is most interesting, at least in the historical type of research that interested Dilthey.

Even as Dilthey touted the importance of focusing on contextual idiosyncrasy, however, he also indicated that knowledge of general causal relationships and an understanding of general patterns that cut across different time periods and different geographical contexts could inform the hermeneutically oriented scholar (Rickman, 1961). And, of course, other scholars associated with the hermeneutic tradition—most notably Max Weber (1949)—went even further and translated insights garnered from hermeneutical analysis into propositions about cause and effect relationships that could be tested through quantitative research.

Somewhat similar thinking can be seen when one looks closely at the field of cultural anthropology. Geertz (1973), for example, merely described what people who call themselves ethnographers tend to do, not what all types of researchers—or even all types of qualitative researchers—should do. And as was the case with some scholars associated with the hermeneutics tradition, some cultural anthropologists never completely rejected cause/effect thinking or the goal of generating theory that applied to more than a single setting. Historically, in fact, a number of anthropologists engaged in the practice of ethnology. Ethnologists used ethnographers’ findings about idiosyncratic cultures to construct more general theories about culture that often, explicitly or implicitly, invoked the concept of causality. Furthermore, an interest in something akin to old-fashioned ethnology has been resurrected in recent years by anthropologists such as Markus (1995) who have argued that ethnographers who exclusively focus on the socially constructed meaning in individual cultures inevitably overlook global forces that delimit and constrain the construction-of-meaning process. (This argument also is articulated in the Anderson and Scott article in this issue of Qualitative Inquiry.)

Even Blumer (1969), the leader of the symbolic-interactionism movement within sociology, a movement that rejected the notion that attitudes and predispositions can be treated as causes that produce certain effects, suggested that qualitative researchers still are capable of answering questions like policymakers’ what-works question. Blumer simply argued that, because action is “fashioned, constructed, and directed by the process of definition [of meaning] that goes on in the individual or the group. . . , a knowledge of this [definition] process would be of far greater value for prediction, if that is one’s interest, than would any amount of knowledge of tendencies or attitudes” (p. 98).

So, the claim that the goal of explicating meaning and the use of constructivist-oriented qualitative methods is logically inconsistent with causal or causal-like explanations of social phenomena was not shared by many research traditions that focused on explicating meaning and used qualitative research methods. Claims about inconsistency only emerged when Lincoln and Guba (1985) appropriated
Kuhn’s (1996, c. 1970, 1962) paradigm construct to make sense of the expanding interest in qualitative methods and the growing concern with explicating meaning that was visible during the last decades of the 20th century in a number of social science fields that previously had used primarily quantitative methods and focused on establishing cause and effect relationships.

**Clarifying the concept of incommensurability.** Whether or not the emerging interest in explicating meaning and using qualitative methods was, indeed, equivalent to the sorts of paradigm shifts that Kuhn described in the physical sciences is something that continues to be debated (see, for example, Donmoyer, 2006; Nespor, 2006; Wright & Lather, 2006). What no longer seems debatable, however, is the fact that Lincoln and Guba (1985) misinterpreted Kuhn’s notion of incommensurability. For example, in later editions of The Structure of Scientific Revolutions, Kuhn (1996, c. 1970, 1962) clearly indicated that incommensurability is not a synonym for logical incompatibility. In other words, one need not be irrational to embrace different paradigms at different times to accomplish different purposes.

In fact, traditionally, the term incommensurability simply referred to the fact that one cannot determine the relative correctness of competing paradigms empirically. One cannot, for example, design a critical experiment—or even a series of experiments—to determine whether Einstein’s view of the physical world is more accurate than Newton’s because the framing and language of the competing paradigms differ so substantially.

Of course, differences in framing and language also mean that one cannot simultaneously employ different paradigms. Kuhn (1996, c. 1970, 1962), for example, relies on a gestalt psychology metaphor to make this point: “The marks on paper that were first seen as a bird are now seen as an antelope, or vice versa” (p. 85), he writes. Obviously, with only a little effort, one can see both the bird and the antelope (albeit sequentially, not simultaneously), and, although Kuhn acknowledges that scientists do not normally shift their perspective when working within the ideal typical world of their academic disciplines, there is no reason why they—or anyone else, for that matter—must suffer from paradigm myopia. Indeed, even physicists who consistently embrace Einstein’s view of the physical world when working with their linear accelerators, almost certainly think in Newtonian terms when they leave work and put their feet on the accelerators of their cars.

**Reconsidering Lincoln and Guba’s ontological argument for rejecting causality.** Physicists think in Newtonian terms when driving their cars because it is functional for them to do so. Indeed, once we embrace a constructivist epistemology—that is, once we assume that our knowledge of the world is inevitably filtered through and influenced by socially constructed cognitive schema—we have few other options than to use functional or instrumental criteria to decide which of a number of potentially credible paradigms to employ at any particular point in time to accomplish a particular purpose.

What I have just said, of course, conflicts with what Lincoln and Guba wrote in their highly influential 1985 book, Naturalistic Inquiry. Their incommensurable-paradigm objection to using the concept of causality to characterize social phenomena is based, primarily, on the assumption that traditional researchers (the so-called positivists) and adherents to the new paradigm (i.e., constructivists) embrace distinctly different (and, supposedly, logically inconsistent) ideas about the nature of reality. Although positivists supposedly embrace an ontology that assumes that “reality is single, tangible, and fragmentable” into independent and dependent (and, at times, intervening) variables and, ultimately, causes and effects, constructivists assume that “realities are multiple, constructed and holistic,” and must be represented by detailed descriptions of the mutual, simultaneous shaping of social action (Lincoln & Guba, 1985, p. 37).

Now, the problem with constructivists making their case against the use of the notion of causality on ontological grounds is that they, also, embrace a constructivist epistemology. Once one assumes that knowledge of the social world is inevitably socially constructed, it makes little sense to talk about the way the world really is. In short, for social constructivists, epistemology trumps ontology.

At best, we can use the complete absence of a sense of experiential verisimilitude to eliminate conceptions of the real world that are totally outlandish. (Quantitative researchers would label this a face-validity strategy.) But when we must choose among potentially credible socially constructed visions of reality, we must invoke instrumental or functional criteria. We cannot ask which view of reality is more correct, because we cannot make judgments about the nature of reality independent of our constructions of it. We can only ask: What do different conceptions of social reality allow us to do; which conception is most helpful—that is, most functional—in accomplishing the particular task (e.g., driving a car, making social policy) in which we are engaged? When we focus on this sort of question, it becomes clear that some human activities cannot be undertaken in any sort of thoughtful way without thinking in cause/effect terms.

**An instrumental case for qualitative researchers talking in cause/effect terms.** Peter Cohen (1968), in fact, demonstrated the virtual necessity of cause and effect thinking for human activities such as policymaking years ago while critiquing Peter Winch’s (2007; c. 2003, 1990, 1958) classic book, The Idea of a Social Science. In his book, Winch made the same sort of ontological argument against causal explanation in social science research that scholars like Lincoln and Guba (1985) articulated several decades later. Cohen’s argument did not so much reject Winch’s ontological argument;
Cohen, in fact, conceded that “one would have to agree that the use of the term ‘causation’ does not have as precise a reference in the social world as it does in the natural world” (p. 416). Rather Cohen made his case on instrumental grounds: “If one is to use such criteria,” Cohen wrote, “one wonders what is to be offered in place of ‘causation’. . . . In fact, one begins to wonder how social policy would be possible without some idea of causation” (p. 416).

To state Cohen’s point another way: The notion of causality can be thought of as a functional fiction rather than as an ontological claim (which constructivists would never be able to prove or disprove anyway, given that they assume that all views of reality are social constructions), there is no defensible paradigmatic reason for them to automatically reject characterizing social reality in cause–effect terms. Or, to stick closer to home, how would a scholar write what hopefully will be a persuasive essay without thinking, at least at times, about the essay’s effect on the reader.

In short, if qualitative researchers view the notion of causality as a functional fiction rather than as an ontological claim, then they have anything to offer policymakers. Rather Cohen made his case on instrumental grounds: “If one is to use such criteria,” Cohen wrote, “one wonders what is to be offered in place of ‘causation’. . . . In fact, one begins to wonder how social policy would be possible without some idea of causation” (p. 416).

In recent years, members of the policy community have been exceedingly optimistic about researchers’ ability to gauge the impact of policies and programs they enact. In recent years, members of the policy community have been exceedingly optimistic about researchers’ ability to gauge the impact of policies and programs they enact. In recent years, members of the policy community have been exceedingly optimistic about researchers’ ability to gauge the impact of policies and programs they enact. In recent years, members of the policy community have been exceedingly optimistic about researchers’ ability to gauge the impact of policies and programs they enact. Now that researchers are being asked to provide what works, it is clear that they must be able to provide a credible answer. In recent years, members of the policy community have been exceedingly optimistic about researchers’ ability to gauge the impact of policies and programs they enact. The most recent study was an evaluation study of an urban school district’s federally funded in-service development efforts for principals (Galloway & Donmoyer, 2010). A second study also focused on a federally funded principal development effort; in this case, however, the focus was on the impact of a preservice preparation program developed and implemented by a university/school district partnership on program graduates who had been fast-tracked by the district into principal positions (Donmoyer, Yennie-Donmoyer, & Galloway, 2012). The final study was funded by a foundation and looked at a reform initiative developed by a school district/foundation partnership more than a 3-year period (Donmoyer & Galloway, 2010).

All of the studies used primarily qualitative methods, though in each study quantitative survey data were collected to triangulate qualitative interview data. In addition, in the third study mentioned above, there was an attempt to use more sophisticated qualitative techniques when funders and school district officials began attributing higher student performance on state tests to the initiative they had funded. My co-principal investigator and I did what we could to determine whether there was convincing evidence to support such claims. Also because of funder expectations, there was an attempt, in the second study listed above, to analyze quantitative student achievement data in fast-tracked principals’ schools and to explore whether the principal and, even, the preparation program that the principal had been a part of might have had an impact on student achievement.

My game-plan in the second part of the article has two parts: First, I will use the studies listed above to demonstrate that policymakers have, in fact, overestimated quantitative researchers’ ability to answer their what-works question; one of the studies also will be used to demonstrate at least part of the reason this overestimation continues to occur. Then, I will use the studies to demonstrate that policymakers often have underestimated the importance of qualitative research in gauging the impact of the policies and programs they enact and fund. I begin, however, with a brief description of current thinking about research within the policy community.

Practical Concerns

In considering the question of whether constructivist-oriented qualitative researchers have anything to offer a policy community interested in learning about “what works,” I will draw on three recent studies for which I served as a co-principal investigator. The studies were funded by policy-making groups either within the federal government or, in one case, the foundation world. The most recent study was an evaluation study of an urban school district’s federally funded in-service development efforts for principals (Galloway & Donmoyer, 2010). A second study also focused on a federally funded principal development effort; in this case, however, the focus was on the impact of a preservice preparation program developed and implemented by a university/school district partnership on program graduates who had been fast-tracked by the district into principal positions (Donmoyer, Yennie-Donmoyer, & Galloway, 2012). The final study was funded by a foundation and looked at a reform initiative developed by a school district/foundation partnership more than a 3-year period (Donmoyer & Galloway, 2010).

All of the studies used primarily qualitative methods, though in each study quantitative survey data were collected to triangulate qualitative interview data. In addition, in the third study mentioned above, there was an attempt to use more sophisticated qualitative techniques when funders and school district officials began attributing higher student performance on state tests to the initiative they had funded. My co-principal investigator and I did what we could to determine whether there was convincing evidence to support such claims. Also because of funder expectations, there was an attempt, in the second study listed above, to analyze quantitative student achievement data in fast-tracked principals’ schools and to explore whether the principal and, even, the preparation program that the principal had been a part of might have had an impact on student achievement.

Great Expectations

In recent years, members of the policy community have been exceedingly optimistic about researchers’ ability to
definitively tell them what works and, consequently, which policies and programs to fund. The only stipulation has been that researchers use quantitative methods—for example, quantitative measures of impact, experimental designs in which research subjects are randomly assigned to control and experimental groups (Whitehurst, 2003)—to link processes (i.e., causes) with measures of impact (i.e., effects), and, when experimental designs are not feasible, sophisticated analysis techniques such as multiple regression analysis or hierarchical linear modeling (National Research Council, 2002) to discover the apparent impact of different independent variables on specified dependent variables.

This emphasis on quantification was especially strong during the recent Bush administration when Grover Whitehurst (2003) headed the federal government’s Institute of Educational Sciences (IES) and declared that randomized trials were the federal government’s new gold standard in research. The gold-standard rhetoric has largely disappeared as a new administration has come to town, but the Reading Excellence Act, a law that stipulated that all federally funded reading research had to use experimental designs and that passed Congress during the Clinton administration with strong bi-partisan support, indicates that a strong belief in and preference for quantitative research is not simply associated with one political party (Donmoyer, 2006). The so-called new philanthropy indicates that a strong belief in and preference for quantitative research is not simply associated with one political party (Donmoyer, 2006). The so-called new philanthropy focus of many foundations also has been translated not only into a general concern for accountability but also a specific interest in quantitative measurement (Eikenberry & Kluver, 2004).

Given this emphasis on the use of quantitative methods, it is hardly surprising that both governmental and nongovernmental policymakers often relegate qualitative methods to a supporting role. They are to be used, for example, “when plausible hypotheses are scant” (Feuer, Towne, & Shavelson, 2002, p. 8) and case studies are needed for hypothesis generation purposes. The other role for qualitative methods that has been sanctioned in recent years is to help make ex post facto sense of apparent anomalies in a study’s quantitative data (Cronbach, 1975; Institute of Education Sciences, 2006).

A Request for Proposals (RFP) issued by the IES in 2006, for example, actually solicited case study proposals, but the RFP made clear that the purpose of the case studies was to develop hypotheses that would be confirmed, later, by large-scale quantitative studies. The RFP also stipulated that even the case studies needed to be quantitative case studies. Qualitative data could only be used “as a complement to quantitative measures of student outcomes. . . [i.e., to help] explain the effectiveness or ineffectiveness of the intervention. . . [and] identify conditions that hinder implementation of the intervention” (Institute of Education Sciences, 2006, p. 67, emphasis added).

The Overestimation of Quantitative Research’s Ability to Answer the What-Works Question

The studies I have been involved with in recent years suggest that policymakers’ faith in the use of quantification to answer their what-works question is not justified. For example, whenever we attempted to do anything resembling sophisticated quantitative analysis, either because our funders expected it or because justifying a funders’ claim about their program’s impact required it, we encountered a host of practical problems, many of which undoubtedly are not unique to our particular studies. I will discuss two problems here, the multiple-reforms problem and the problem of selection effects.

The multiple-reforms problem. As I noted above, by the third year of our evaluation of a school reform initiative developed and implemented by a foundation/school district partnership, the partners were attributing achievement gains in schools that had opted to be part of the reform to the reform initiative they had helped develop and, in the case of the foundation, fund. Consequently, my colleague and I, in our roles as evaluators of the reform, felt compelled, in what previously had been primarily a qualitative study, to use statistical modeling techniques to determine whether these claims could be defended.

Soon after we began this part of our investigation, however, we encountered a host of problems. We eventually labeled one problem the reform-overload problem. Quite simply, the reform we were researching was not the only reform that was being implemented in district schools at the time of our study.

The reform overload problem was especially acute in schools that were classified as low performing. During an informal lunchroom interview with a teacher in one such school, for example, the teacher kept getting confused about which of the many reform initiatives being implemented in the school at the time we were asking about. A member of the school’s leadership team who was nearby kept telling our interviewee, “No, it’s the one [i.e., the reform initiative] about [improving students’] vocabulary [the particular reform focus this school had selected as part of the reform initiative we were investigating].” We were not convinced that, even after such prompting, the teacher we were talking with actually understood which of the school’s many reform initiatives we were asking about.

Unfortunately for us, the reform-overload problem was not limited to low-performing schools. During a formal two-on-one interview with the principal of a highly successful school, for instance, we commented that we were somewhat surprised that the school had increased its math scores dramatically, given that the school had selected to focus all of its resources in the reform effort we were
studying on improving student writing. The principal, at that point, leaned forward and let us in on a little secret: In defiance of reform initiative rhetoric about needing to focus resources narrowly, the principal had clandestinely implemented a rather extensive professional development initiative focused on improving the teaching of mathematics.

Thus, in this school, as in the low-performing school discussed previously, we could not quantitatively link documented achievement gains back to the specific reform effort we were studying. Furthermore, it is likely that the multiple-reforms problem will arise in studies in other school districts as well, since it unlikely that most districts will agree to cease additional reform activity—especially in low-performing schools—simply to make it easier for researchers to solve the attribution-of-causality problem. And, although techniques such as multiple regression analysis and hierarchical linear modeling might be used in studies that have much larger samples than ours had to estimate the percentage of variance that could be attributed to a particular reform, the sample would have to very large to successfully do this and, in the end, we would still have only an estimate of a reform’s apparent impact.

The selection-effects problem. At least as problematic as the reform-overload problem were the problems we encountered with selection effects. To be sure, the most obvious selection-effects problem we encountered was a consequence of a particular design feature of the reform we were studying: the reform’s commitment to voluntary participation by the schools that were part of the reform initiative. Because of this commitment, we certainly could not even consider randomly assigning district schools to control and experimental groups, a key component of Whitehurst’s (2003) gold standard for educational research.

We also could not implement the statistical modeling techniques that quantitative researchers tend to use when experimental designs are not feasible, because the process and the criteria used to select schools to participate in the reform varied in important ways from year to year. During the first year of our study, for example, when the developers’ theory of action dictated that participation had to be voluntary and the number of volunteer schools was so small that all schools that had volunteered were selected, the sample of participating schools tended to be skewed toward already successful schools with highly effective principals (who in some cases had used their considerable expertise to “sell” participating in the reform project to their staffs). During the second year, the situation was different. At that point, word of the reform’s positive features had spread, and, consequently, there were more volunteer schools than available slots. The group of schools that was selected from the pool of volunteer schools during the second year was skewed toward high-needs schools that had applied to participate.

This sort of skewing in the direction of troubled schools was even greater during the final year of our study: At that point, the reform developers’ commitment to voluntary participation often appeared to be trumped by the need to do something significant with the district’s most problematic schools before the negative consequences stipulated in the No Child Left Behind law kicked in; consequently, according to our interviews with a number of principals, a number of academically problematic schools were “nudged” into the “voluntary” initiative.

In short, in all 3 years, the procedures and criteria that were used to select schools into the reform initiative we were studying were quite different. These differences meant that we could not answer questions about impact statistically. Had we simply ignored the selection effects problem and proceeded with the modeling and testing processes we had planned to use, we would have been doing pseudoscience for the purpose of public relations.

Undoubtedly, some of the selection effects problems we encountered in our study of a particular reform were unique to the reform we were studying. Still, it is likely that other sorts of impossible-to-manage selection-effects issues will pop up in other studies conducted in any complex environment outside of the controlled environment of a laboratory, and it is unlikely that practitioners will ignore legitimate concerns simply to accommodate researchers’ pristine research designs.

Why quantitative research’s inability to answer the what-works question may not be noticed. Most members of the policy community seem quite unaware of the sorts of problems discussed above or a host of other problems that are likely to arise when quantitative methods are used to answer the what-works question. The most recent study being discussed here (and the one listed first in the list of studies presented above) provides some insight into why a lack of awareness might occur: Quantification often serves to create the illusion that we can know what is and is not working without actually providing accurate information about what works. In this particular study, the illusion involved meeting requirements of the federal Government Performance and Results Act (GPRA) that became law during the Clinton administration as part of that administration’s reinventing government initiative.

As the passage of GPRA, federal programs must articulate so-called GPRA measures that can be used to determine the impact—or lack of impact—of initiatives the government has funded. Measuring programmatic outcomes is always a difficult business, of course, and measurement problems become even more acute when funders use a RFP process to make funding decisions. Such processes at the federal level normally provide some degrees of freedom for those submitting proposals so that the projects that are funded can respond to local situations and local needs as well as more general federal priorities.
The federal program that funded the local initiative we were evaluating in the first study listed above attempted to manage the tension between the need for simplification in measuring impact, on the one hand, and the desire to accommodate the diverse needs and goals of those who submit grant proposals, on the other hand, by requiring those who submit funding proposals to identify in their proposals a measurement tool that they believe is consistent with their goals. If they receive a grant, the proposal writers must agree to administer the identified instrument at the beginning and end of each of the 5 years of the grant to recipients of grant-funded interventions. The measures generated at the beginning of the year function as pre-assessment data, while the end-of-the-year measures serve as post-assessment data.

Gain scores for each participant are then calculated to determine impact (or lack of impact) of the intervention on particular individuals.

The federal program streamlined the gain-score calculation process by providing a spreadsheet to grant recipients. Recipients (or their designated evaluators, in our case) were told simply to insert the pre- and post-assessment scores of each grant participant on the spreadsheet, and the number of participants with significant gains would automatically (and rather mysteriously) be calculated. Program officials in Washington can then aggregate these data from each funded program and report to policymakers—who make decisions about continuing, expanding, or terminating federal programs—the number of grant-funded participants that demonstrated significant improvement during the year in question.

When we looked at the process of generating the so-called GPRA measures up close, we saw a host of problems. Many of the problems related to the instrument being employed (an instrument, incidentally, that had been listed in the RFP as an example of the sort of instrument that could be selected); others involved aspects of the administration of that instrument. Four of these problems are briefly discussed below.

First, the instrument was a 360 degree assessment tool in which a principal’s effectiveness was gauged by having the principal’s teachers, his or her supervisor, and the principal himself/herself respond to survey items about the principal’s performance. The instrument was originally designed to be used formatively to assess individuals within a professional development context. It had not been designed to calculate gains scores as part of a summative assessment of program impact. In fact, we discovered that the highly touted psychometric evidence about the instrument’s worth had been generated in contexts that reflected the instrument’s original intent, not the new summative purpose for which it was now being marketed.

Second, in our study, for a variety of reasons, supervisors did not participate in the assessment except at the high school level. Consequently, at the elementary and middle school levels, a principal’s self-assessment accounted for half of the effectiveness measure. Unfortunately, we generated a good deal of qualitative data about principals “gaming the system,” for example, intentionally rating themselves low in the pre-assessment and high in the post-assessment.

Third, response rates were quite low, especially in the post-assessment. Only 23 of the 111 district schools had teacher response rates high enough to meet the instrument-developer-established validity threshold of 50%. Furthermore, when we examined these 23 schools to check for principal participation, we found that 6 of the 23 schools that had high enough response rates to stay in the validity game (as defined by the instrument developers) had nonparticipating principals.

This low response rate produced a fourth problem: In many of the schools, the number of participants was so low that we could not formally assess inter-rater reliability, something that seems important to do when using a 360 degree instrument for summative purposes. Assessing inter-rater reliability seems especially important when self-assessment accounts for half of the pre- and post-assessment scores.

I could go on. What is important here, however, is the fact that there was no established procedure for feeding back information about the problems outlined above (and a number of other problems, as well) to federal officials through the GPRA assessment system they had devised. Our task was simply to input pre- and post-assessment data—data that, in many cases were highly suspect—and let the spreadsheet magically calculate whether gains were adequate to demonstrate that involvement in the initiative had had a significant impact on a participant.

What we have here, in short, is the illusion that quantitative data are telling us what is and what is not working. And while the illusion may have been functional for the federal officials who were in the untenable position of complying with the GPRA, on the one hand, and attempting to accommodate local idiosyncrasy, on the other hand, the fiction also obscures the limitations of quantitative methods in terms of answering policymakers’ what-works question. One consequence of this illusion is that policymakers end up using unrealistic criteria when assessing the potential contribution that qualitative methods can make in answering their what-works question.

Summary. The analysis in this part of the article “lowers the bar” substantially for answering the bottom-line question being addressed in the second half of this article: Can qualitative researchers answer policymakers’ what-works question in a credible and defensible way? The discussion now focuses directly on this bottom-line question.
What, if Anything, Can Qualitative Researchers Offer Policymakers Interested in What Works?

Qualitative data and the preponderance-of-evidence strategy.

The discussion in the prior section demonstrates quite clearly that, in many cases, quantitative data and analysis cannot definitively answer policymakers’ what-works question. At best, they can only create the illusion that the question has been answered. Consequently, if we want to know what works, we are forced to rely on our best estimates of what the answer to the what-works question might be. To state this point another way: Unless we are willing to be deluded by the mystification that numbers can, at times, provide, we are forced to use what my colleagues and I call a preponderance-of-evidence strategy (Donmoyer & Galloway, 2010) when determining what worked in particular settings and what is likely to work in other places.

The good news here for qualitative researchers is that there is no reason that qualitative data cannot be used in playing the preponderance-of-evidence game. Indeed, in the studies being discussed here, explanations of impact (or lack of impact) could never have been provided, even when using the preponderance of evidence decision rule, had we not reviewed the extensive qualitative data we had gathered. But because we had gathered extensive qualitative data and used these data to construct relatively thick descriptions of what had been happening in each school we studied, we were able to predict, in virtually all cases, those schools in which achievement scores would increase and those in which scores would decline long before the scores were released by the states in which the schools we were studying were located and later explain the likely reasons for what had happened. This was certainly the case in the study of principal preparation program graduates who had been fast-tracked into principal positions. In all cases we predicted which principals’ schools would produce higher scores than the school had produced before the new principal arrived and which schools would have declining scores.

More importantly, we could use the qualitative data, along with quantitative survey data gathered primarily for triangulation purposes, to make sense of either the apparent impact or lack of impact. During member checking, most readers of the documents we produced in each of the districts in which impact-oriented questions were asked indicated that they found our sense making—which as noted, relied primarily on the qualitative data we had gathered—quite convincing.

One final point before moving on: What I have written here is not really all that different from the IES RFP writers’ comments quoted above about qualitative data functioning “as a complement to quantitative measures of student outcomes (Institute of Education Sciences, 2006, p. 67, emphasis added). There is, however, one rather important difference: In our studies, at least, qualitative data were essential for making preponderance-of-evidence cases that linked outcomes with processes, and it was the quantitative data that played the supporting role. In other words, we could not have predicted a priori what would happen with respect to student achievement scores—and we certainly could not have explained what happened ex post facto—had we attempted to build our preponderance-of-evidence cases exclusively, or even primarily, on quantitative data.

An argument for reporting thick description as thick description. A confession undoubtedly is in order, at this point: To make the thick description in the study being discussed above (and in the other studies, as well) palatable to policymakers who want simple, easy-to-digest answers to their what-works question, we used a cross-case analysis process to construct exceedingly distilled storylines. These distilled storylines were reasonable facsimiles for the sorts of regularity type of causal explanations (Macdonald, 2012) that quantitative researchers provide.

For example, in the beginning principal study that was alluded to in the previous section, after making a highly detailed and quite nuanced preponderance-of-evidence case that accounted for the achievement gains and losses in a number of principals’ schools, we stripped away most of the complexity and idiosyncrasy during a cross-case analysis process and ended up telling the story of principals whose success was associated with the principals working closely with their staffs to extensively analyze student achievement test data and the implications of these analyses for modifying classroom instruction. We also told another distilled tale of test scores that appeared to have declined, at least in part, because of the school principal’s apparent lack of interest in playing the instructional leader role in any form and to teacher morale problems that occurred, at times, because the principal relied on district officials to lead improvement efforts in her school that felt punitive to teachers.

Now these distillations were certainly supported by evidence. They were, however, exceedingly incomplete. For example, not all principals who did the test-data-analysis routine with their staffs were equally successful, but the additional factors that made some principals more successful than others were generally lost in translating what Macdonald (2012) calls process causality into what he refers to as regularity causality. Before concluding this discussion, therefore, I want to argue that there is value in reporting relatively un-distilled thick description in the form of case studies (Geertz, 1973, 1983; Lincoln & Guba, 1985) even to a policymaking community that prefers simple answer to its what-works question. Once again, I will rely on one of the studies listed at the start of the second part of this article to make my point.
During the first year of the 3-year study of school reform (the third study listed above), only five of 33 district schools had opted to participate in the reform initiative that was being developed and implemented by a school district/foundation partnership. Consequently, it was possible, during the first year of the project, to use a case study/cross-case analysis research design. Furthermore, because the reform developers were, in a very real sense, designing the reform as they were implementing it, my colleague and I had little choice but to focus on creating thick-description accounts of what was happening in the five schools that had opted to participate.

If nothing else, the five thick-description cases demonstrated the incredible amount of diversity that exists even within a seemingly homogeneous school district. One school’s teachers, for example, had just graduated from college and frequently engaged in adolescent-like spats that, at times, interfered with reform activity. After one emotional meeting, for example, one of the teachers planted her tongue firmly in her cheek and suggested that all the school’s teachers join hands and sing “Kumbaya.” By contrast, another volunteer site had a mature staff that, during the previous year, had succeeded in getting the school’s previous principal fired. A third school was a district-sponsored bilingual charter school so committed to parent involvement that parents made up a more than 50% of the membership of the school leadership team.

In fact, the other two schools also exhibited important features that were not like the features of the other schools that had opted to participate in the reform during Year 1, and, so, my partner and I quickly began to think of the case studies we were generating as miniature versions of the “thick” cultural descriptions anthropologists like Geertz produced and for which education scholars such as Lincoln and Guba provided a methodology. And what did these case studies have to offer policymakers? If nothing else, they demonstrated how complex schools are, how simple-minded policymakers’ what-works question is, and how important it is to provide more than a little “elbow” room to accommodate contextual variation when writing policy and developing programs.

But thick description can do more than provide a cautionary tale for policymakers too intent on standardization. It also can provide policymakers with new frames for thinking about the issues they are confronting and the problems they are attempting to solve.

In our study of the five schools that had opted into the reform effort during Year 1, for example, we also looked closely at the school district in which the reform was occurring, largely because the district seemed, at first glance, to be a bundle of contradictions. The district, for example, sincerely valued decentralization and letting each of its schools experiment with different approaches to reform that the schools’ staffs had, themselves, selected. The evidence we gathered was unequivocal on this point. Other evidence, however, also suggested that the district was simultaneously engaging in a fair amount of top-down control. The district staff, for example, informally monitored what was happening in each of the district’s schools and constantly communicated what they had learned with other district-level staff. We also discovered that a surprising number of principals had lost their jobs during and prior to our study because of perceived inadequate performance.

We spent a good deal of time and effort making sense of how the district managed, for the most part, to seamlessly integrate decentralization and top-down-control approaches to management, approaches that traditionally have been seen as antithetical. We suspect that policymakers in other districts and, possibly, also at the state and federal levels, would find our thick-description account of the sense we made (see Donmoyer & Galloway, 2010) useful as they think about the management of other schools. To be sure, it is unlikely that our case study could be translated into the sort of “formulas for practice” that those who ask the what-works question seem to desire. Our little thick-description account, however, could certainly be used heuristically to think about what might work in another district setting and to generate new approaches to old problems. At the very least, our account challenges conventional dichotomous thinking about managing schools and other organizations.

Of course, the idea of research serving a heuristic function requires a different and more nuanced interpretation of policymakers’ what-works question than policymakers normally have in mind when they pose this question. Furthermore, it is unlikely that many policymakers will take the time to actually read the thick descriptions researchers produce. Other people who interact with policymakers may read some of this material, however. I, for example, have spent a good part of my professional life advising policymakers about what might—and what might not—work, and my advice has almost always been informed by the thick descriptions scholars have produced. I would be remiss if I did not also acknowledge that much was lost in translating thick description insights into a form that meshes comfortably with most policymakers’ less-than-subtle mindset. But, at the very least, I normally have been able to make policymakers feel at least a tad uncomfortable about the simplistic way they view the world. On a good day, I have even encouraged them to rethink their what-works question from a somewhat different—and, at times, more nuanced—frame of reference.

Conclusion

This article considered whether qualitative researchers have anything to offer members of the policy community who expect researchers to tell them, in relatively simple terms, what works. The first part of the article explored philosophical/paradigmatic objections to qualitative researchers...
attempting to answer any cause–effect question, including policymaker’s what-works question. The conclusion was that the notion of causality is a functional—and, indeed, an indispensable—fiction in the policy arena and that, in a research tradition that assumes that we can only know the social world through our own constructions of it, paradigmatic decisions can only be made by invoking functional or utilitarian criteria. Consequently, there should be no paradigmatic prohibition against qualitative researchers attempting to answer policymakers’ what-works question.

The second part of the article took a more pragmatic tack and considered whether constructivist-oriented qualitative researchers who view the notion of causality as a functional fiction within the policy arena have anything useful to offer policymakers who want to know what works. After lowering the bar by demonstrating that members of the policy community often overestimate the answers quantitative researchers can provide, the article demonstrated that questions of policy and program impact can only be answered by using a preponderance of evidence strategy and that there is no reason that qualitative evidence cannot be included in a preponderance-of-evidence case. Indeed, often qualitative data are required for the evidence to make sense. The heuristic value of traditional thick-description approaches to qualitative research also was demonstrated.

Acknowledgments

The author wants to acknowledge the contributions of Fred Galloway, the other principal investigator on all three studies, and of June Yennie Donmoyer who was a key figure in the data collection and analysis processes of the two federally funded studies.

Author’s Note

An earlier version of this article was presented at the 2011 meeting of the American Educational Research Association held in New Orleans, LA.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Two of the studies discussed in the second half of this article were funded by the School Leadership Program of the United States Department of Education; the third study was funded by the Ball Foundation. The writing of this particular article also was supported, in part, by a Faculty Research Grant from The University of San Diego. The author appreciates the support he has received from funders and acknowledges that funders do not necessarily endorse the ideas articulated in this article.

Notes

1. Initially scholars such as Lincoln and Guba (1985) used the term naturalistic paradigm to refer to thinking that they later renamed the constructivist paradigm.

2. The term triangulate has a number of meanings. For the most part, in the studies being discussed here, the term refers to the use of quantitative survey data to determine whether interview findings from a subset of organizational actors were representative of the thinking of organizational members in general.

References


Bio

Robert Donmoyer is currently professor of leadership studies at the University of San Diego. He has written extensively about qualitative research and evaluation methods, educational leadership and reform, and the philosophical issues that ground research practice. His recent publications include “Reconsidering the Utility of Case Study Designs for Researching School Reform in as Neo-Scientific Era: Insights from a Multiyear, Mixed-Methods Study,” which was published in the February 2010 issue of Educational Administration Quarterly and “Cultural Connections in Leadership Education and Practice,” which appeared in a 2011 issue of the International Journal of Leadership in Education.