2014

When and How Qualitative Methods Provide Credible and Actionable Evidence; Reasoning with Rigor, Probity, and Transparency

Sharon F. Rallis

7

When and How Qualitative Methods Provide Credible and Actionable Evidence

Reasoning with Rigor, Probity, and Transparency

Sharon F. Rallis

Evaluation is applied research, so as evaluators, we want our evaluations to make a difference in policy or practice. To produce evidence that stakeholders trust enough to use, we reason through the many choices we make in designing and conducting studies. Our first decisions focus on which evaluation questions and which methods effectively yield data to inform those questions. Then, implementing the method requires numerous selections and ongoing decision making as we collect, analyze, and interpret data to turn them into findings. Finally, we choose how and to whom we report findings. Throughout the entire process, we attend to *rigor* (adherence to established standards for process), *probity* (wholeness, integrity, and moral soundness), and *transparency* (open and detailed documentation or display of all decisions and actions). If the methods we use match the purpose of the evaluation, if we

I want to acknowledge and thank several people: Rachael Lawrence, my trusty graduate assistant, who was my right arm in conducting the PLBSS study and who provided valuable feedback on drafts of this chapter; Bethany Rallis, a clinical psychologist-in-training who expands my perspectives and forces me to clarify my arguments; and Gretchen Rossman, my colleague and friend, with whom I have explored and written about many of these ideas over the years.

employ these methods ethically with technical competence, and if our decisions and the underlying reasoning are apparent, the evaluation will meet our ultimate goal—to produce credible and actionable evidence. Therefore, we must choose our methods wisely.

Qualitative evaluation is a prudent choice when research questions aim to illuminate the program practices: the *how* and *why* of program implementation on intended effects. When the appropriate method is qualitative, the evaluator's decisions and corollary actions become especially important because the evaluator *is* the "instrument" (Guba & Lincoln, 1989, p. 175; Lincoln & Guba, 1985, p. 236; Patton, 2002, p. 14) or "means through which the study is conducted" (Rossman & Rallis, 2012, p. 5), and the participants are individual (or groups of individual) human beings, each with unique values, perceptions, and experiences. With qualitative methods, we collect data in natural settings, often through face-to-face encounters; thus the moral principles of *respect, beneficence,* and *justice** shape how we treat participants. Our procedures are emergent rather than prefigured. Our analyses and interpretations, while grounded in theories of action or conceptual frameworks, are openly subjective—that is, meanings are generated within perspectives and experiences of the context (see, for example, Blumer, 1969, pp. 78–89; Rallis & Rossman, 2012). Thus articulating our standards and revealing and explaining what we do and why is imperative. If we employ qualitative methods with rigor, ethical practices, and transparency, the evaluation illuminates program practices and effects, and stakeholders are highly likely to use the findings for program improvement, decision making, and informing policy.

In this chapter, I describe the reasoning behind the use of qualitative methods, demonstrating why and how these methods in an evaluation can produce credible and actionable evidence. When are qualitative tools appropriate for meeting the purpose of the evaluation and exploring the evaluation questions? How do qualitative data become information—and what technical rigor is involved? What composes ethical practice in data collection, analysis, and interpretation? I suggest that by addressing the *how* and *why* questions, qualitative methods have the potential to bridge the research-to-practice gap. I ground the reasoning in principles articulated in the National Research Council's *Scientific Research in Education* (Shavelson & Towne, 2002) and in moral principles underlying ethical practice.

---

*These are three basic principles, generally accepted in in Western traditional culture, that guide the ethics of research involving human subjects (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979).

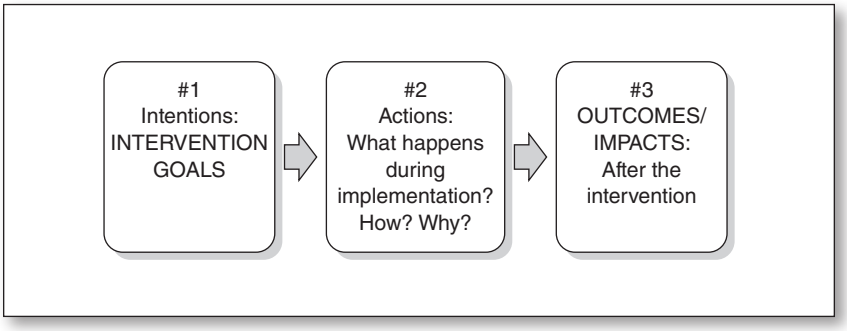
Matching Method to Program Theory and Questions

Producing credible and actionable evidence begins with choosing the appropriate method—and this choice does not depend on a preference for a particular method (quantitative or qualitative) or on a belief that one method is considered more rigorous than another. Rather, the decision depends, either implicitly or explicitly, on purposes of the evaluation. We ask the following questions: What do stakeholders want to know about the program? What are the evaluation questions? The program's logic or theory of change generates a series of questions, and stakeholders—whether policy makers, program leaders, practitioners, or users—direct the evaluation toward the particular focus of interest. We choose methods with the tools to collect and analyze the data that will best inform those questions. Making explicit the program's theory of change is therefore critical to designing and conducting the evaluation.

Decision makers choose programs based on theories or arguments that suggest that the programmatic intervention will make a positive difference in the context. "Social programs are based on explicit or implicit theories about how and why the program will work" (Weiss, 1995, pp. 66–67): *If we implement X intervention, then Y outcomes will result.* The following figure illustrates the links between intention, action, and result in a theory of action.

Articulating the theory brings to the surface assumptions and causal inferences and allows evaluators to locate where the research questions fall. Are we interested in the connection between goals and outcomes (Boxes #1 and #3)—What happens as a result of the intervention? If so, we probably choose a method (such as an experiment) that allows causal inference. Do we want to know whether the program was implemented with fidelity? Or do we

Figure 7.1 Theory of Action



want to explore the complex interactions that take place within what is referred to as the *black box* (Box #2) as the goals become program activities? These latter questions, which fall into explanatory or exploratory contexts, demand descriptive answers that require descriptive statistics or qualitative methods. Given stakeholders and resources, we may choose to look at several questions.

Most evaluations I conduct focus on Box #2, exploring what happens during program implementation:

- What actions take place in the setting? How do people perform these actions? Why?
- What do these actions mean to the participants—both during and afterward?
- What are perceived effects of these actions?
- In what ways do the actions relate to each other and to external forces?
- Are these actions and meanings organized into any patterns that indicate norms for this setting?*

These questions demand descriptive data—data that tell a story—so I use qualitative tools with participants in the setting. I observe the program in operation; I listen to and read the words of program participants; I examine artifacts related to the program operation. Capturing these images, words, voices, objects, and ideas requires intensive interaction with the people and the program activities and materials in order to tell a story that participants can recognize and that sheds light on the program operation and connections. Such descriptions honor the idiosyncratic and contextual nature of participants' experiences in the program and allow complex and dynamic interpretations of those experiences. I borrow tools from ethnography to understand interactions and relationships, from phenomenology to understand perspectives and perceptions about the lived experience, and from sociolinguistics/semiotics to understand how people communicate their meanings for activities, events, objects, and people (see Rallis & Rossman, 2012 for discussion of these genres). I also draw on the program's theory or logic to make sense of what people say and do, I document and display my decisions, and (to the best of my ability) I treat participants with respect and fairness while addressing their needs and concerns. I aim to achieve wholeness, integrity, and moral soundness. In short, while still adhering to technical rigor, probity, and transparency, my method allows the program participants to tell their own story.

For example, in a two-part evaluation we conducted (in 2011 and 2013) of the Professional Learning-Based Salary System (PLBSS), a professional

*These questions are loosely adapted from the work of Erickson (1986).

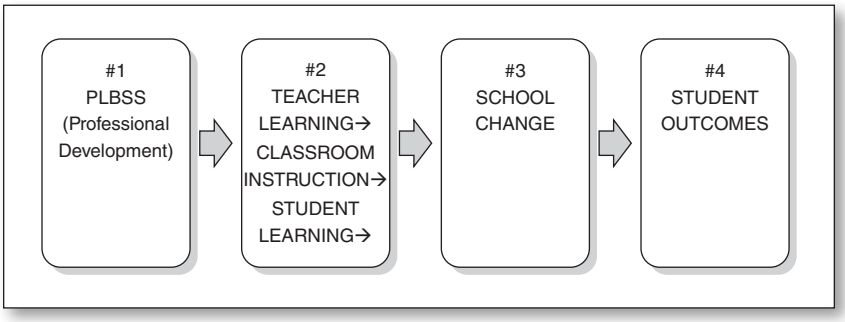
development initiative implemented in the Portland (Maine) Public Schools, the theory of action read as follows:

> If teachers are compensated on the basis of their professional learning, their salaries will increase and they will become agents of their own learning. They will build skills and knowledge, both individually and collaboratively, to improve their instructional practices and a broader culture of learning in the schools. These improvements will result in increased student learning. (Rallis, Churchill, & Lawrence, 2011)

Each conditional action represents a cause and intended effect, so to design the evaluation, we worked with two primary stakeholder groups (the district leadership and the teachers' union) to specify the focus and generate evaluation questions.

The first questions fell into Boxes #1 and #4: Did teachers participate and did their salaries rise? Did student scores increase? An initial study answered these questions with surveys and analyses of district records. Since the program was operationalized during the 2007–2008 school year, participation (and satisfaction) levels in PLBSS were high, salaries rose, and student scores improved. However, since district and union leaders agreed that conducting controlled experiments were impractical, they acknowledged that the achievement increases might not be directly attributed to the professional development intervention—but they still wanted to know how the intervention might be related to the student outcomes. Essentially, they wanted the evaluation to look into Boxes 2 and 3 to explore any causal relationships among teacher learning, classroom instruction, student learning, and school change.

Figure 7.2 PLBSS Theory of Change



SOURCE: Rallis, Keller, Lawrence, & Soto, 2013

District and union leaders therefore directed their questions toward teacher and student actions related to the professional development activities. Did participation in these activities build skills and knowledge, both individually and collaboratively? Did teachers use their learned skills in their instructional practices? Did student work change relative to instruction? In what ways did teachers work together in the schools? Did the school culture change—and if so, in what ways? We needed evidence of teacher learning, use of that learning, and its affect on student work. Quantifying or measuring teacher learning or use of learning was not possible, given the varying purposes and formats, number, and potential uses of professional development opportunities, so at this point, we chose qualitative tools: observations in schools and classrooms, interviews, reviews of professional development curriculum and teacher lesson plans, and analyses of student work. Put simply, what did both teachers and students *experience* and then *do* with their experience in their instruction and classroom work? In this case, our evaluation design sought to illuminate the black boxes between the intervention and the impact on student scores, thus directing us to qualitative methods.

While this example comes from the education field, my assertions can be generalized to other practical fields that serve social needs, because evaluation is an applied science. For example, current discussion among clinical psychologists considers the need to recognize individual differences in both patients and clinicians in the development of best practices and that naturalistic methods help bridge the chasm between research and practice, as these methods answer questions that quantitative methods cannot (Beutler, 2009; Reed, Kihlstrom, & Messer, 2006).

From Data to Use: Clear Criteria and Thick Description

When using designs that match method with program logic and evaluation questions, evaluators collect and combine *data* (images, words, numbers, impressions) into *information*; if policy makers or program participants use that information, it becomes *knowledge*. Since use may take several and varied forms and since evaluators seldom have full control over how the evaluation information will be used, we seek to provide evidence that can inform the evaluation questions and possible decisions or actions. The relationships among data, information, and knowledge, while not completely linear, might best be illustrated as follows:

> *Data* (analyzed and synthesized) become → *Information* (interpreted and used) becomes → *Knowledge*
>
> (Rallis & Rossman, 2012, p. 10)

When using qualitative methods, evaluators collect data as words and images and artifacts to capture participants' actions or reports of actions and their perspectives of the experienced actions. Through analysis, we combine these data to build *thick descriptions* that make interpretation possible, suggesting intentions, causal connections, and patterns of behavior (see Geertz, 1973). The descriptions are so detailed that readers (potential users of the evaluation) can see what the evaluator sees. Often best portrayed in narrative form, the combined images and words tell a story that constructs understandings of and generates insights into program operations and effects. The narratives are not intended to be generalizable; they are detailed descriptions of one program or setting. However, they allow potential users to judge the value of the findings and decide which aspects might apply to similar settings and in what ways—and thus potentially suggest future actions.

Credibility depends on how we as evaluators combine those data, how we analyze and synthesize pieces of data into information. *Actionability*—that is, *use*—lies in how the information is interpreted and then presented to audiences. Stakeholders use findings, depending on whether they understand and accept how we have created and told the story, that is, how we made meaning of the words and images. If they can see how and where the data were collected and how conclusions grew out of these data, they become interested in and engage with the findings. As evaluators using qualitative tools, we pay special attention to clearly articulating and revealing the criteria we use to organize and analyze our data and to make judgments (interpretations) about the meaning and value of actions and perceptions.

Whatever the method, technically rigorous analysis and interpretation of these actions—or perceptions of these actions—rely on identified criteria, the underpinnings of an analytic framework. Evaluation is the "systematic assessment of the operation and/or outcomes of a program, compared to a **set of explicit or implicit standards** [bold added for emphasis], as a means of contributing to the improvement of that program or service" (Weiss, 1998, p. 4). Because not all stakeholders are likely to share the same set of beliefs, values, goals, and definitions, we use the program logic to reach agreement on criteria for assigning meaning to data. The standards or criteria emerge from the program logic and serve to bind the selection, analyses, and interpretation of data: What exactly is the intervention intended to do to yield which specific results? How will we know if the intervention happens as intended? How do we judge effects?

In choosing the focus of the evaluation questions and articulating criteria, stakeholders and evaluators emphasize what they consider important, thus making explicit their subjectivity. The criteria serve to guide analysis: What categories of action can we expect? How will we know if or what actions occur? What characteristics describe or define expected quality/success? We specify indicators and details that enable us to recognize these indicators in the

images, words, or pictures we collect. Deductively using the program logic, we detailed descriptions and patterns. We compare what we see, hear, and read to the defined criteria. Identifying expected operations and outcomes also facilitates inductive analysis; the unexpected is highlighted, and we consider how these unexpected actions relate to the program goals and implementation. Throughout the process, we reason through choices and make decisions, embedding our evaluative thinking in program theory and criteria.

To illustrate how criteria are developed and used to support analysis and interpretation, we refer to our evaluation of the PLBSS, in which several levels of criteria emerged. The levels and criteria were articulated together with program and district personnel, beginning with descriptions of the categories of offerings (traditional university courses; district-created, including teacher-led, courses; and individual projects): What counts as an activity in each category to receive salary credit? Criteria included required structure (e.g., length, meeting time, materials) and content. Content had to address either district priorities (technology, English language learners [ELLs], early childhood needs, adolescent literacy, poverty) or instructional needs (literacy and language acquisition, math, cultural competency). Finally, to connect local criteria to national criteria, we cross-referenced content, structure, and process to the *Standards for Professional Learning,* a set of standards and indicators developed collaboratively and adopted by various professional educational associations (Learning Forward, 2011). These criteria provided categories for organizing and coding our data. The images, words, actions, or artifacts did not need to fit the categories; rather, the criteria allowed us to agree on meanings and to build a convincing narrative about what was happening in professional development activities, among teachers, and in classrooms and schools.

We analyzed the data collected in response to the following questions:

- What choices did teachers make among professional development offerings?
- What did the teachers learn in their professional development choices?
- How do teachers use these learnings in their classroom instruction?
- What student performance examples in classrooms illustrate use of or response to instruction related to specific professional development of the teacher (e.g., how student writers develop their characters)?

We asked how teachers' choices met instructional needs or district needs. We sought descriptions of the activity, its format, and its content. We captured stories in which teachers described what they learned and how they used these strategies in their classroom instruction. We spent time in their classes observing instruction and students working. We looked at lesson plans. The stories were enhanced by collecting evidence that illustrated students' performance in response to specific instruction related to their teacher's professional development.

In one school, for example, the data told the story of a turnaround that was shaped largely by teachers' participation in professional development that changed the way they worked with each other and with children in their classrooms. With a majority population of ELLs, the school was identified as failing; the district chose to use a school improvement grant to develop teachers' skills for teaching reading and writing to ELLs through job-embedded language and literacy acquisition workshops that met during the school day with push-in literacy coaching. We heard, "We did not choose [to participate]—at first, we hated it. You cannot teach these kids." Soon, teachers began to say, "I need to learn how to teach these children" and "I'm learning strategies I can use in my classroom." Their words described how planning and instruction of reading and writing concepts and skills became coherent across years and grades by using common calendars and consistent writing prompts and sentence frames, which changed developmentally from grade to grade. Teachers showed us student work (displayed throughout the school) that represented evidence of activities related to the workshop curriculum; for example, at every grade level, we saw student essays on how writers develop characters. We documented, often with pictures, teachers working together, moving between classrooms with ease and relating to students in any classroom. Their words added credibility to our findings:

- I find that when taking a class or course together with other teachers from the same school or district, discussions often develop about what we learned. It's constructive adult talk with peers.
- It's not just that we like each other; it is about the work. The learning and sharing continues.
- We use more common language, more common strategies . . . [and understand] what other teachers at different levels need you to do at your level. And I know what to expect from the children before they come to my class.
- Again, all of this, I never could have done any of this if I hadn't spent those lousy credit hours, and the payback is really good . . . for the kids. I have been teaching for quite a few years, but I needed to enhance my ability to teach ELLs. . . . [I] t's been a wonderful experience for me because our school has grown in diversity immensely over the last few years and will continue to do so, and I feel it's important for me. . . . You really learn how to teach—much more so than the college courses that certify you as [an] ELL [teacher].
- The [workshops] changed the content and focus of independent reading. I built an extensive classroom library, created a blog for students to discuss their independent reading, [and] revised the focus of the writing prompts related to reading, and as a result, increased my students' interest in and practices around independent reading. These practices, I believe, transfer to more in-depth and comprehensible reading in all content areas.
- Triangulating the assessment data on my ELL . . . offers evidence of success that I can credit to some degree to the specific emphasis I have placed on the workshops—differentiation, tiered instruction, and language frames. Building back-

ground knowledge has enhanced my work around curriculum mapping, which in turn has generated clearer learning targets for students who, for their part, have demonstrated stronger work products [writing pieces] over time.

- When you walk around our school, you see the evidence and volume of student writing, and you see students in primary grades writing nonfiction reports, which is such a keystone of the common core [standards]. It's amazing, when you walk the school [to see all of their work].
- Really, what is it students need to do to comprehend the text at a higher level of thinking, critical thinking[?] . . . I actually learned what moves a kid from a level . . . to the next band of reading that you come to, what kinds of things do I need to teach them—to move a student to the next level. . . . An example is being able to have them give evidence, like a character trait, that they know who a character is. Humor, as well, what is the evidence that shows it . . . or what is the author [saying]. . . . [I]f he isn't coming right out and saying it, can the student infer what the author was implying?

Analyzed and synthesized using the criteria, these images and words become the rich and detailed story we offered as evidence of what happened with the PLBSS. When we presented the findings, leaders and practitioners recognized the story and engaged with questions, additions, critiques, and comments such as: "That's a lot to think about."

From Data to Use: Trust, Transparency, and Everyday Ethical Practices

The relationship we built with the district and program leadership as well as with the teachers and other participants contributes to the credibility and usability of the stories offered as evidence. Turning qualitative data into findings that illuminate practices and linkages between intents and outcomes requires evaluators to interact with other human beings—namely, the participants (policy makers, funders, planners, and practitioners at various levels). As evaluators using qualitative methods, we enter the program world; we both shape and are shaped by the actions and interactions. We are satisfied that program logic, criteria, and thick description have kept our interpretations bounded and relevant. Still, why should the stakeholders and potential users trust that interpretations and findings bear any relationship to reality and have any potential for use? The answer is this: Transparency and ethical practices are imperative. Can our audiences see and understand how we reached these conclusions? Is the chain of reasoning evident and coherent and grounded in moral principles? Are the relationships continually assessed as respectful, beneficial, and just?

Transparency and ethical practices are important at all stages of the evaluation. As the evaluation begins, participants must know what evaluators will be

doing and why; they must trust that the evaluator's actions will be respectful and not bring harm and that the motivation to evaluate this program is fair. At the end, whether they are practitioners or policy makers, potential users want to trust the integrity and social value of the evaluation; as Phillips (2007) argued,

> [a] policymaker no doubt will consult the empirical data, but rightly will be swayed by ethical premises and by notions of what constitutes a good and caring society, and by his or her conceptions of the rights that all individuals possess. (p. 383)

Ultimately, according to Weiss (1998), the purpose of evaluation *is* ethical. Evaluators are not judges; rather, "with the information and insight that evaluation brings, organizations and societies will be better able to improve policy and programming for the **well-being of all** [bold added for emphasis]" (p. ix).

Grounding evaluation in ethical purposes and relationships adds a moral dimension to the evaluator's reasoning: What moral principles guide my decisions and interactions with stakeholders? "The point of moral principles is to regulate interactions among human beings" (Strike, Haller, & Soltis, 1998, p. 41). Again, the relational aspect is especially present when using qualitative methods, which rely on human interactions. While human subject reviews have emphasized procedure over relationships in real settings, the principles can provide guidance for ongoing moral reasoning. The *principle of respect for human persons* implies a moral obligation to honor participants' rights and to treat them as an end in themselves, not as mere means to an end. The *principle of beneficence* is a moral obligation to act to benefit others and protect them from risk, recognizing that an evaluator cannot know what is best for participants; it does not imply a balance of benefits and risks. Finally, the *principle of justice* ensures that both selection and receipt of benefit should be equitable and that procedures are nonexploitative. Only through a morally reasoned relationship between the evaluator and the participants can the dimensions, parameters, and expectations of these principles be ethically defined, negotiated, and shared (see, for example, Hemmings, 2006; Rallis, 2010; Rallis, Rossman, & Gajda (2007); and Rossman & Rallis, 2010 for further discussion of ethical practice and moral principles).

Ongoing ethical dialogue about standards for conduct necessarily occurs because not everyone agrees upon or acts on a common ethic. In fact, ethical theories can be grouped into two broad categories that may conflict. One set of moral principles form consequentialist theories that focus on results of actions: Any particular action is neither intrinsically good nor bad; rather, it is good or bad because of its results in a particular context—its consequences. If the end has value, the means are less important. Such a view argues for programs that result in the greatest good for the greatest number. Nonconsequentialist ethical

theories, on the other hand, advocate ethics of individual rights and responsibilities and of justice. These ethics recognize universal standards or rights to guide all behavior, regardless of the consequences in a specific context. The protection of these rights may not be denied, even for the greatest good for the greatest number. Principles of fairness and equity are used to judge whether program purposes and actions are right and wrong, seeking to ensure the well-being of all, even though the allocation of benefit may differ (Rawls, 1971).

Evaluators' reasoning moves from abstract principles to concrete circumstances of the evaluation, asking where the evaluation questions and criteria actually focus: on the consequences, on the outcomes, or on the rights and treatment of participants as the intervention is implemented? *Communitarianism* ethics (MacIntyre, 1981) acknowledge that communities differ on what is morally good or right, and evaluators know that stakeholders often hold different fundamental values. Those values may even conflict with the evaluator's own values. Which values define the evaluation questions, guide the decisions, and shape interpretations? Nonconsequentialist theories of justice and individual rights raise questions of power and representation: Who defines what is right in this program? Whose values are used in rating or in choosing criteria? Are all voices given opportunity to be heard? Different stakeholders in a given situation might call for an evaluator to draw on more than one moral principle or from various ethical perspectives. For example, influential stakeholders or clients might demand consideration of outcomes over process in the program while program recipients would be concerned with potential benefits, costs, or harm.

Addressing ethical issues is not a single event nor is it facilitated by any codified procedure. Rather, ethical reasoning consists of ongoing decisions made throughout the evaluation in *caring* interactions with those affected by the decisions at the moment. Instead of turning to a principle for guidance, the caring evaluator returns to the participants: What do they need in this place at this time? Will filling this need harm others in the network of care? Am I, as evaluator, competent to fill this need? Is the expressed need really in the best interest of the participant? (Noddings, 1995, p. 187) *Care theory* emphasizes the moral interdependence of people: "As [evaluators], we are as dependent on our [participants] as they are on us" (Noddings 1995, p. 196). Reciprocity in mutual care and respect bridges the gap between purposes and needs of the evaluator and the evaluated.

As caring evaluators, we respect the connections among the participants, the program, and ourselves. We want to understand the interactions and the relationships themselves, the interdependencies: How does one person's meaning making interact with and influence another's? A *caring ethic* expects the relationship to be dynamic, symmetrical, and connective; we give respect, and we are respected (Lawrence-Lightfoot, 2000). At the same time, as we honor the participants, we work to create conditions that allow them to respect our efforts to discover and understand their experience.

In the end, the task of reasoning out the ethic to apply in each case falls to the evaluator. Again, I illustrate our ethical reasoning as we conducted the PLBSS evaluation, framing our decisions with the three principles commonly used to guide human subjects research. First, we discussed with everyone we encountered, whatever the relationship to the program or evaluation, the purposes and expected procedures of the study—and to the best of our ability, we made sure these were understood; thus we gained informed consent.

RESPECT

To honor participants' individuality and rights, each interview was a unique conversation. We considered our participants to be capable professionals and treated them as such, engaging in dialogue about their motivations and experiences. Because the policy to change salary allocation to teachers had several motivations—including the need to increase salaries in order to retain quality teachers, the desire to give more autonomy to teachers by allowing them to choose their own professional development, and the belief that increased teacher learning would increase student learning—we explored the reasons behind each personal or group decision and ensured that each was included in the program logic and in the criteria for analysis and interpretation. We were open to concerns about process and content and considered collaboratively which interview and observation strategies would be least intrusive and would respect each participant's right to privacy. We aimed to build mutual relationships that facilitated authentic dialogue about their choices and their use—or nonuse—of professional development. In fact, because we were genuinely interested in what they were doing, teachers and school leaders readily shared their experiences, ideas, and critiques. The conversations (interviews) often lasted longer than scheduled and led to participants' willingness to let us use direct and identifiable quotes.

BENEFICENCE

Given that we could not intuit what different stakeholders would consider benefits or burdens of the evaluation, we sought collaboration in the design: What theory of change guided the program and participation? What might they want to learn from the study? How might they use the information? Then, together, we explored data sources: Which schools would offer the most informative data—and would not suffer from the intrusion of the evaluators for interviews and observations? When might be best times to collect data? What might be best ways to deliver the findings—and to whom? Similarly, to articulate criteria, we discussed their expectations and how they might define quality or success. Regarding interviews and observations, we noted that any promises

for complete confidentiality and anonymity would be false and promised that were we to use a direct quote, we would check with the speaker first. We tried to consider each case individually to meet the needs of each person in that instance. For example, when a person who held a singular role such as coach or principal did not want their words identified, we discussed and reached agreement in how to meet mutual needs, such as disguising the role. These agreements were open to ongoing negotiations and revisions when needed. For example, as the study progressed, more people wanted to own their words and waived anonymity. In short, we were guided by the question, What is right for her, him, or them?

JUSTICE

Our efforts to ensure fairness focused on site selection, data collection strategies, and the analysis-interpretation stage. We reasoned through site selection and strategies to be sure that no group was either exploited or received special reward. Satisfied that no special interests or private agenda were in play, we agreed to select schools that were facing specific and publically acknowledged challenges: one that had been labeled by the U.S. Department of Education as a school requiring corrective action and one whose student population demographic had drastically changed in the past two years. While analyzing and interpreting data, we explicitly considered the following questions: Whose data are these? What voices may be missing? What alternative interpretations are likely? What influence might power relationships play in meaning making? For example, we encouraged meetings that brought together potentially conflicting stakeholders (e.g., administration and union leaders). Also, we were vigilant about speaking to a range of teachers with varying needs, noting (for example) a comment questioning the system's equity of opportunity: "For teachers who coach, have young families, or actually correct student papers, there is not enough time in life to complete the hours of [professional development] and the paperwork to advance [salary]" (personal communication [interview], April 9, 2013).

The PLBSS evaluation met our standards for rigor, probity, and transparency as well as the standards of the district leaders, union personnel, and teachers. But would it pass muster with a larger and external audience?

Applying Standards for Scientific Inquiry

Ultimately, the users—both intended and accidental—of the evaluation hold the key to its credibility and actionability. Those who believe the evaluation has value will use it in some way commensurate to the perceived value. Still, what

is considered science and the scientific community influence credibility and actionability by legitimizing particular forms or methods of research with its promulgated definition of standards for scientific inquiry. Even groups within the scientific community disagree. In 2004, the U.S. Department of Education in the No Child Left Behind Act labeled randomized controlled trials (RCT) as the gold standard for evaluation, but other groups argue for broader understandings of science. As a science educator who reviewed the Rossman and Rallis (2003) qualitative methods manuscript wrote, "There is more art in science than you recognize." Watson (1968) describes in *The Double Helix* how the discovery of the structure of DNA required rigorous reasoning, but it also involved creativity, politics, mystery, and love. "Science seldom proceeds in the straightforward logical manner imagined by outsiders. Instead, its steps forward (and sometimes backwards) are very often very human events in which personalities and cultural traditions play major roles" (p. ix). Kuhn (1970) reminds us that we accept something as truth until the scientific community accumulates enough evidence that another truth exists and iteratively until more evidence is uncovered to alter this truth.

The point is that scientific knowledge is a social construct, so credible and actionable evidence becomes what the relevant communities of discourse and practice accept as valid, reliable, or trustworthy. Judgments about the quality of inquiry represent "the social construction of knowledge . . . [T]he key issue becomes whether the relevant community of scientists evaluates reported findings as sufficiently trustworthy to rely on them for their own work" (Mishler, 2000, p. 120). A relevant community to inform evaluators' understandings of legitimate scientific inquiry might be the National Research Council, whose Committee on Scientific Principles for Education Research defined *scientific inquiry* as

> [a] continual process of rigorous reasoning supported by a dynamic interplay among methods, theories, and findings. . . . Advances in scientific knowledge are achieved by the self-regulating norms of the scientific community over time, not, as sometimes believed, by the mechanistic application of a particular method to a static set of questions. (Shavelson & Towne, 2002, p. 2)

The committee members declare what scientific inquiry is *not* (static or mechanistic) while recognizing the role of discourse. Science requires the display of findings, along with the reasoning that led to those findings, for others in the community to contest, modify, accept, or reject. Their discourse concludes that scientific research studies do the following:

- pose *significant questions* that can be investigated empirically
- link research to *relevant theory*
- use methods that permit *direct investigation* of the questions

- provide a coherent and explicit *chain of reasoning*
- replicate and *generalize* across studies
- disclose research to encourage *professional scrutiny and critique*

In short, the National Research Council implies that scientific research involves human judgment—reasoning—so applying these principles to use of qualitative methods for evaluation seems appropriate for judging the legitimacy of the method. The PLBSS evaluation offers a tangible case from practice to apply the following principles:

1. Collaboratively chosen, the evaluation posed observable, practical, realistic (all terms used to define *empirical*) questions deemed significant among district leaders, union personnel, teachers, and ourselves (the evaluators).
2. We articulated a strong program theory from which evaluation questions, design, and criteria emerged. As well, we reviewed literature on professional development as grounding for developing the program theory of change with participants.
3. To investigate the questions directly, we conducted real-time on-site observations and interviews and personally reviewed program materials.
4. To ensure that our chain of reasoning was both coherent and explicit, we included program stakeholders in decision making, transparently discussed our methods, revealed data and analyses behind interpretations, and sought feedback and alternative perspectives.
5. While findings apply directly only to the sites studied, we have shared the story of PLBSS with national audiences (with express permission from the program participants); feedback indicates that the insights offer deep understanding of activities that can apply to other turnaround schools in districts elsewhere.
6. Finally, we presented the findings in various forums in the district, the community, and the nation. In these discussions, the study and findings were scrutinized and critiqued. And, as academics, we are writing about the process and findings for publication, thus further disclosing our information to encourage professional scrutiny and critique.

Applying the National Research Council principles is itself an exercise of reasoning with rigor, probity, and transparency. However, missing from the list of principles is a criterion I view as critical to determining the value of evaluation as an applied science: *use*. Therefore, final essential questions I ask of my evaluations include the following: Does the evidence inform the program operation to improve the well-being of participants? In what ways do the findings bridge the gap between theory and practice? We believe that the PLBSS study used question-driven methods that produced a credible story to which audiences responded and that illuminated the initiative,

informing both policy and practice improvements for sustainability and even transferability—that is, action.

Question-Driven Inquiry: Matching Method to Purpose

In designing a study, evaluators first determine the program theory of action (*program/intervention X will produce Y results*) and locate the focus of the evaluation within that logic. When the evaluation questions seek to measure results, quantitative methods are employed. When the evaluation questions what happens between the *X* and *Y* of the program theory—that is, the activities, events, experiences, and perceptions related to program implementation and operation—qualitative methods can access the data that build exploratory or explanatory descriptions that tell the story of the program. Both methods require reasoning with rigor, probity, and transparency to produce credible and actionable evidence. Methods are chosen to match evaluation purposes, not for the method's status in the research world or on evaluators' preferences or expertise. The method appropriate to the evaluation question can yield data to build information that decision makers may use for program improvement. In summary, evaluation is a question-driven inquiry process.

References

Beutler, L. E. (2009, September). Making science matter in clinical practice: Redefining psychotherapy. *Clinical Psychology: Science and Practice, 16*(3), 301–317.

Blumer, H. (1969). *Symbolic interaction*. Englewood Cliffs, NJ: Houghton Mifflin.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Witrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York, NY: MacMillan.

Geertz, C. (1973). *The interpretation of cultures*. New York, NY: Basic Books.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: SAGE.

Hemmings, A. (2006). Great ethical divides: Bridging the gap between institutional review boards and researchers. *Educational Researcher, 35*(4), 12–18.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Lawrence-Lightfoot, S. (2000). *Respect*. Cambridge, MA: Perseus Books.

Learning Forward. (2011). *Standards for professional learning*. Retrieved July 7, 2013, from http://learningforward.org/standards

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: SAGE.

MacIntyre, A. (1981). *After virtue*. Notre Dame, IN: University of Notre Dame Press.

Mishler, E. G. (2000). Validation in inquiry-guided research: The role of exemplars in

narrative studies. In B. M. Brizuela, J. P. Stewart, R. G. Carillo, & J. G. Berger (Eds.), *Acts of inquiry in qualitative research* (pp. 119–145). Cambridge, MA: Harvard Educational Review.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April 18). *The Belmont Report. Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, DC: Office of the Secretary Ethical Principles and Guidelines for the Protection of Human Subjects of Research.

Noddings, N. (1995). *Philosophy of education*. Boulder, CO: Westview.

Patton, M. Q. (2002). *Qualitative research & evaluation method* (3rd ed.). Thousand Oaks, CA: SAGE.

Phillips, D. C. (2007). Adding complexity: Philosophical perspectives on the relationship between evidence and policy. In P. A. Moss (Ed.), *Evidence and decision making: The 106th Yearbook for the National Society for the Study of Education* (pp. 376–402). Malden, MA: Blackwell.

Rallis, S. F. (2010, July–August). "That is NOT what's happening at Horizon!": Ethics and misrepresenting knowledge in text. *International Journal of Qualitative Studies in Education, 23*(4), 435–448.

Rallis, S. F., Churchill, A., & Lawrence, R. B. (2011). *Supporting professional learning: Impacts of the PLBSS in Portland Public Schools.* Amherst, MA: Center for Education Policy.

Rallis, S. F., Keller, L., Lawrence, R., & Soto, A. (2013). *Revisiting PLBSS*. Amherst, MA: Unpublished report.

Rallis, S. F., & Rossman, G. B. (2012). *The research journey: Introduction to inquiry*. New York, NY: Guilford Press.

Rallis, S. F., Rossman, G. B., & Gajda, R. (2007). Trustworthiness in evaluation practice: An emphasis on the relational. *Evaluation and Program Planning, 30,* 404–409.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Reed, G. M., Kihlstrom, J. F., & Messer, S. B. (2006). What qualifies as evidence of effective practice? In J. C Norcross, L. E. Beutler, & R. F. Levant (Eds.), *Evidence-Based practices in mental health: Debate and dialogue on the fundamental questions* (pp. 13–55). Washington, DC: American Psychological Association.

Rossman, G. B., & Rallis, S. F. (2003). *Learning in the field: An introduction to qualitative research* (2nd ed.). Thousand Oaks, CA: SAGE.

Rossman, G. B., & Rallis, S. F. (2010). Everyday ethics: reflections on practice. *International Journal of Qualitative Studies in Education, 23*(4), 379–391.

Rossman, G. B., & Rallis, S. F. (2012). *Learning in the field: An introduction to qualitative research* (3rd ed.). Thousand Oaks, CA: SAGE.

Shavelson, R., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: Committee on Scientific Principles for Education Research, National Research Council.

Strike, K., Haller, E., & Soltis, J. (1998). *The ethics of school administration*. New York, NY: Teachers College Press.

Watson, J. D. (1968). *The double helix: A personal account of the discovery of the structure of DNA*. New York, NY: New American Library.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell, A. C. Kublsh, L. B. Schorr, & C. H. Weiss (Eds.), *Approaches to evaluating community initiatives* (pp. 65–92). New York, NY: Aspen Institute for Humanistic Studies.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.