From *Utilization-Focused Evaluation*
by Michael Quinn Patton (2008, Sage)

PART 3

Appropriate Methods

Blowhard Evaluation

This is the story of three little pigs who built three little houses for protection from the BIG BAD WOLF.

The first pig worked without a plan, building the simplest and easiest structure possible with whatever materials happened to be laying around, mostly straw and sticks.

When the BIG BAD WOLF appeared, he had scarcely to huff and puff to blow the house down, whereupon the first pig ran for shelter and protection to the second pig's house.

The second pig's house was prefabricated in a most rigorous fashion with highly reliable materials. Architects and engineers had applied the latest techniques and most valid methods to the design and construction of these standardized, prefabricated models. The second pig felt quite confident that his house could withstand any attack.

The BIG BAD WOLF followed the first pig to the house of the second pig and commanded, "Come out! Come out! Or by the hair on my chinny-chin-chin, I'll huff and I'll puff and I'll blow your house down."

The second pig laughed a scornful reply: "Huff and puff all you want. You'll find no weaknesses in this house, for it was designed by experts using the latest and best scientific methods guaranteed not to fall apart under the most strenuous huffing and puffing."

So the BIG BAD WOLF huffed and puffed, and he huffed and puffed some more, but the structure was solid, and gave not an inch.

In catching his breath for a final huffing and puffing, the BIG BAD WOLF noticed that the house, although strong and well built, was simply sitting on top of the ground. It had been purchased and set down on the local site with no attention to establishing a firm connecting foundation that would anchor the house in its setting. Different settings require very different site preparation with appropriately matched foundations, but the prefabricated kit came with no instructions about how to prepare a local foundation.

Understanding all this in an instant, the sly wolf ceased his huffing and puffing. Instead, he simply reached down, got a strong hold on the underside of the house, lifted, and tipped it over. The second pig was shocked to find himself uncovered and vulnerable. He would have been easy prey for the BIG BAD WOLF had not the first pig, being more wary and therefore more alert, dashed out from under the house, pulling his flabbergasted brother with him. Together they sprinted to the house of the third pig, crying "wee wee wee" all the way there.

The house of the third pig was the source of some controversy in the local pig community. Unlike any other house, it was constructed of a hodgepodge of local materials and a few things borrowed from elsewhere. It incorporated some of the ideas seen in the prefabricated houses designed by experts, but those ideas had been altered to fit local conditions and the special interests and needs of the third pig. The house was built on a strong foundation, well anchored in its setting and carefully adapted to the specific conditions of the spot on which the house was built. Although the house was sometimes the object of ridicule because it was unique and different, it was also the object of envy and praise, for it was evident to all that it fit quite beautifully and remarkably in that precise location.

The BIG BAD WOLF approached the house of the third pig confidently. He huffed and puffed his best huffs and puffs. The house gave a little under these strenuous forces, but it did not break. Flexibility was part of its design, so it could sway and give under adverse and changed conditions without breaking and falling apart. Being firmly anchored in a solid foundation, it would not tip over. The BIG BAD WOLF soon knew he would have no pork chops for dinner that night.

Following the defeat of the BIG BAD WOLF, the third pig found his two brother pigs suddenly very interested in how to build houses uniquely adapted to and firmly grounded in a specific location with a structure able to withstand the onslaughts of the most persistent blowhards. They opened a consulting firm to help other pigs. The firm was called "Wee wee wee, all the way home."

—From Halcolm's *Evaluation Fairy Tales*

11

Evaluations Worth Using

Utilization-Focused Methods Decisions

They say there was method to his madness. Perhaps so. It is easier to select a method for madness than a single best method for evaluation, though attempting the latter is an excellent way of achieving the former.

—Halcolm

The three pigs story that precedes this chapter and introduces this part of the book on Appropriate Methods offers an evaluation parable. The first pig built a house that was the equivalent of what is disparagingly called a "quick and dirty evaluation." They are low-budget efforts that give the outward appearance of evaluation, but their value and utility are fleeting. They simply do not stand up under scrutiny. The second pig replicated a high-quality design that met uniform standards of excellence as specified by distant experts. Textbook designs have the advantage of elegance and sophistication, but they don't travel well. As the old proverb cautions, when all you have is a hammer, everything looks like a nail. Prefabricated structures brought in from far away are vulnerable to unanticipated local conditions. Beware the evaluator who offers essentially the same design for every situation.

The third pig, then, exemplifies the utilization-focused evaluator, one who designs an evaluation to fit a specific set of circumstances, needs, and interests. The third pig demonstrated situational adaptability and responsiveness, a strategic stance introduced in Chapter 6. In this chapter, we'll examine how situational responsiveness affects methods decisions.

Methods to Support Intended Uses, Chosen by Intended Users

Methods decisions, like decisions about focus and priority issues, are guided and informed by our evaluation goal: *intended use by intended users*. Attaining this goal is enhanced by having intended users actively involving in methods decisions, an assertion I shall substantiate in depth throughout this chapter. It remains, however, a controversial assertion, evidence about its desirability and effectiveness notwithstanding. The source of the controversy, I'm convinced, is territorial.

For the most part, evaluation professionals have come to accept that use can be enhanced by actively involving intended users in decisions about the evaluation's purpose, scope, and focus to ensure relevance and buy-in. In other words, they can accept playing a consultative and collaborative role during the conceptual phase of the evaluation. Where I often part company with my colleagues is in the role to be played by intended users in making measurement and design decisions. "The evaluator is nothing," they argue, "if not an expert in methods and statistics. Clearly social scientists ought to be left with full responsibility for operationalizing program goals and determining data collection procedures." Edwards and Guttentag (1975) articulated the classic position, one that I find still holds sway today: "The decision makers' values determine on what variables data should be gathered. The researcher then decides how to collect the data" (p. 456).

Utilization-focused evaluation takes a different path.

Beyond Technical Expertise

The common perception of methods decisions among nonresearchers is that such decisions are primarily technical in nature. Sample size, for example, is determined by a mathematical formula. The evaluation methodologist enters the values of certain variables, makes calculations, and out pops the right sample size to achieve the desired level of statistical robustness, significance, power, validity, reliability, generalizability, and so on—all technical terms that dazzle, impress, and intimidate practitioners and nonresearchers. Evaluation researchers have a vested interest in maintaining this technical image of scientific expertise, for it gives us prestige, inspires respect, and, not incidentally, it leads nonresearchers to defer to us, essentially giving us the power to make crucial methods decisions and then interpret the meaning of the resulting data. It is not in our interest, from the perspective of maintaining prestige and power, to reveal to intended users that methods decisions are far from purely technical. But, contrary to public perception, evaluators know that methods decisions are never purely technical. Never. Ways of measuring complex phenomena involve simplifications that are inherently somewhat arbitrary, are always constrained by limited resources and time, inevitably involve competing and conflicting priorities, and rest on a foundation of values preferences that are typically resolved by pragmatic considerations, disciplinary biases, and measurement traditions.

The reason to debunk the myth that methods and measurement decisions are primarily technical is to enhance use. For we know that use is enhanced when practitioners, decision makers, and other users fully understand the strengths and weaknesses of evaluation data, and that such understanding is increased by being involved in making methods decisions. We know that use is enhanced when intended

users participate in making sure that, when trade-offs are considered, as they inevitably are because of limited resources and time, the path chosen is informed by relevance. We know that use is enhanced when users buy into the design and find it credible and valid within the scope of its intended purposes as determined by them. And we know that when evaluation findings are presented, the substance is less likely to be undercut by debates about methods if users have been involved in those debates prior to data collection.

As in all other aspects of the evaluation, then, the utilization-focused evaluator advises intended users about options; points out the consequences of various choices; offers creative possibilities; engages with users actively, reactively, interactively, and adaptively to consider alternatives; and facilitates *their* methods decision. At the stage of choosing methods, the evaluator remains a technical adviser, consultant, teacher, and advocate for quality. The primary intended users remain decision makers about the evaluation. Exhibit 11.1 summarizes reasons why primary intended users should be involved in methods decisions. In the pages that follow, I'll elaborate on these rationales, explore the implications of this approach, and provide examples. Let's begin with an example.

EXHIBIT 11.1

Reasons Primary Users Should Be Involved in Methods Decisions

1. Intended use affects methods choices. Intended users can and should judge the utility of various design options and kinds of data.
2. Limited time and resources necessitate trade-offs: more of this, less of that. Primary users have the greatest stake in such decisions since findings are affected.
3. Methods decisions are never purely technical. Practical considerations constrain technical alternatives. Everything from how to classify participants to how to aggregate data has utility implications that deserve users' consideration.
4. No design is perfect. Intended users need to know the strengths and weaknesses of an evaluation to exercise informed judgment.
5. Different users may have different criteria for judging methodological quality. These should be made explicit and negotiated during methods discussions.
6. Credibility of the evidence and the perceived validity of the overall evaluation are key factors affecting use. These are matters of subjective user judgment that should inform methods decisions.
7. Intended users learn about and become more knowledgeable and sophisticated about methods and using data by being involved in methods decisions. This benefits both the current and future evaluations.
8. Methods debates should take place before data collection, as much as possible, so that findings are not undercut by bringing up concerns that should have been addressed during design. Methods debates among intended users after findings are reported distract from using evaluation results.

The Million Man March

On October 16, 1995, some number of African American men marched on Washington, D.C., as a call to action. The number of men in the march mattered a great deal to both its organizers and critics. Disputes about the number subsequently led to major lawsuits against the National Park Service, which provided the government's official estimates of demonstrations on the Capitol Mall. For weeks after the march, newspaper commentators, television journalists, policymakers, activists, academics, and pundits debated the number. The size of the march overshadowed its substance and intended message. Varying estimates of the number of marchers led to charges and countercharges of racism and bigotry.

Could this controversy have been anticipated, avoided, or at least tempered? Let's consider how the evaluation was conducted and then how a utilization-focused approach would have been different.

First, let's examine what made this march a focus for evaluation. The organizer of the march, Nation of Islam leader Louis Farrakhan, was a controversial figure often accused of being anti-Semitic and fomenting hatred against whites. Some black Congressman and the leadership of the National Association for the Advancement of Colored People refused to join the march. Many other black leaders worked to make it a success. From the moment the march was announced, through the months leading up to it, debate about the legitimacy and purpose of the march received high-visibility media coverage. As the day of the march approached, the central question became: How many will show up?

Why was the number so important? Because the target number became the name of the march: *The Million Man March*. The goal was unusually clear, specific, and measurable. The march's leaders staked their prestige on attaining *that* number. The march's detractors hoped for failure. The number came to symbolize the unity and political mobilization of African American men.

In time for the evening news on the day of the march, the National Park Service released its estimate: 400,000. This ranked the march as one of the largest in the history of the United States, but the number was far short of the 1 million goal. March advocates reacted with disbelief and anger. March critics gloated at Farrakhan's "failure." Who made the estimate? A white man, a career technician, in the National Park Service. He used the method he always used, a sample count from photographs. Leaders of the march immediately denounced the official number as racist. The debate was on. A week later, independent researchers at Boston University, using different counting methods, estimated the number at more than 800,000—double the National Park Service estimate. The leaders of the march continued to insist that more than a million participated. The significance of this historically important event remains clouded by rancorous debate over the seemingly simplest of all evaluation questions: How many people participated in the "program"? (Janofsky 1995).

Suppose, now, for the sake of illustration, that the responsible National Park official—a white male, remember—had taken a utilization-focused approach. In the time leading up the march, as its visibility and potential historical significance became apparent, he could have identified and convened a group of primary stakeholders: one or more representatives of the march's organizers, representatives of the other national black organizations, academics with expertise in crowd estimates, and perhaps police officials from other cities who had experience

Crowd Counting Is an Inexact Science

In estimating the size of the 1995 Million Man March, The Park Service used the same methods it had always used. Officials take pictures from a helicopter that flies along the sides of the Mall and then, using a grid, take into account the number of people per square foot. They also monitor the volume of passengers using local buses and subways. The Park Service said that its standards and methods of crowd measurement were prescribed by Congressional legislation. Farouk el-Baz, director of the Boston University Center for Remote Sensing, said the angle of the Park Service pictures had failed to capture many of those attending. But the center acknowledged that its own estimate of 870,000 people had a 25 percent margin of error, meaning the crowd could have been as small as 655,000 or as large as 1.1 million.

Samuel E. Jordan, acting director of the city's Office for Emergency Preparedness, who served as the city's liaison with march organizers, said that both the Park Service and the Boston estimates had failed to account for the density of people at the march, people standing under trees and those standing on side streets, well within view of giant screens on which the program was televised. "You can go up in a helicopter all you want, but you have to do it right," Mr. Jordan said. "You have to take it all into account." His estimate of one million, he said, had a margin of error of 20 percent.

SOURCE: Janofsky (1995).

estimating the size of large crowds. A couple of respected newsprint and television journalists could have been added to the group. Indeed, and this is surely a radical proposal, a professional evaluator might have been asked to facilitate the group's work.

Once such group was assembled, consider the challenging nontechnical decisions that have to be made to figure out the size of the march. These questions are in addition to technical questions of aerial photography sampling and computer programs designed to count heads in a crowd. To answer these questions requires some combination of common sense, political savvy, appreciation of different perspectives, and pragmatism. Here, then, are some questions that would occur to me if I had been asked to facilitate such a discussion:

1. Who gets counted? It's the million man march aimed at black men. Do women count? Do children count? Do whites count?
2. Do spectators and onlookers get counted as well as "marchers"?
3. When during the daylong event will counts be made? Is there a particular time that counts the most; for example, during Farrakhan's speech? (His speech was 3 hours long, so when or how often during his speech?)
4. Should the final number account for people who came and went over the course of the day or only people present at some single point in time?
5. What geographical boundary gets included in the count? What are the boundaries of the Capitol Mall for purposes of sampling?
6. Sympathy and support marches are scheduled to take place in other cities. Do their numbers count in the 1 million total?
7. Should we report a single number, such as 1 million, or communicate the variability of any such count by reporting a range, for example, 900,000 to 1.1 million?
8. Who are the most credible people to actually engage in or supervise the final analysis?
9. What reviews should the analysis undergo, by whom, before being released officially?

10. Who do we say determined the counting methods and under whose name, or combination of named sponsors, should the results be publicized?

I certainly don't assert that convening a group of primary stakeholders to negotiate answers to these questions would have ended all controversy, but I do believe it could have tempered the rancorous tone of the debate, diffused the racial overtones of the counting process, and permitted more focus on the substantive societal issues raised by the march—issues about family values, community involvement, social responsibility, economic opportunity, and justice. The evaluation task force, once convened to decide how to count from 1 to 1million, might even have decided to prepare methods of following up the march to determine its longer-term impacts on black men, families, and communities—evaluation questions overshadowed by the controversy about the number of participants.

Parallel Evaluation Decisions

I like the Million Man March example because it shows how a seemingly simple question like "how many" can become quite complicated both technically and politically. Parallel challenges can be found in any program evaluation. For example, in most programs the dropout rate is an important indicator of how participants are reacting to a program. But when has someone dropped out? This typically turns out to involve some arbitrary cutoff. School districts vary widely in how they define, count, and report dropouts, as do chemical dependency, adult literacy, parent education, and all kinds of other programs.

No less vague and difficult are concepts such as *in the program* and *finished the program*. Many programs lack clear beginning and ending points. For example, a job-training program aimed at chronically unemployed minority men has a month-long assessment process, including testing for drug use and observing a potential participant's persistence in staying with the process. During this time, the participant, with staff support and coaching, develops a plan. The participant is on probation until he or she completes enough of the program to show seriousness and commitment, but the program is highly individualized so different people are involved in the early assessment and probation processes over very different time periods. There is no clear criterion for when a person has begun probation or completed probation and officially *entered* the program. Yet the decision, in aggregate, will determine the denominator for dropout and completion rates and will be the numerator for the program's "acceptance" rate. Making sure that such categories are meaningful and valid, so that the numbers are credible and useful, involves far more than statistics. Careful thought must be given, with primary intended users, to how the numbers and reported rates will be calculated and used, including whether they can be used for comparisons with similar programs.

Nor are these kinds of categorical decisions only a problem when measuring human behavior. The Minnesota Department of Transportation has categorized road projects as *preservation, replacement,* and *new or expansion*. How these categories are used to allocate funding to regions throughout the state has enormous implications. Now, consider the Lake Street Bridge that connects Minneapolis and Saint Paul. Old and in danger of being condemned, the bridge was torn down and a new one built. The old bridge had only two lanes and no decorative flourishes. The new bridge has four lanes and attractive design features.

Should this project be categorized as replacement or expansion? (In a time of economic optimism and expanding resources, such as the 1960s, new and expansion projects were favored. In a time of downsizing and reduced resources, like the 1990s, replacement projects are more politically viable.) Perhaps, you might argue, the Lake Street Bridge illustrates the need for a new category: part replacement/part expansion. But no replacements are pure replacements when new materials are used and updated codes or standards are followed. And few expansions are done without replacing something. How much mix, then, would have to occur for a project to fall into the new, combined part replacement/ part expansion category? A doctoral degree in research and statistics provides no more guidance in answering this question than thoughtful consideration of how the data will be used, grounded in common sense and pragmatism—a decision that should be made by intended users with intended uses in mind. Such inherently arbitrary measurement decisions determine what data will emerge in findings.

Methods and Measurement Options

> *There cannot be acting or doing of any kind, till it be recognized that there is a thing to be done; the thing once recognized, doing in a thousand shapes becomes possible.*
>
> —Thomas Carlyle, philosopher and historian (1795–1881)

Mail questionnaires, telephone interviews, or personal face-to-face interviews? Individual interviews or focus groups? Even-numbered or odd-numbered scales on survey items? Opinion, knowledge, and/or behavioral questions? All closed questions or some open-ended? If some open-ended, how many? Norm-referenced or criterion-referenced tests? Develop our own instruments or adopt measures already available? Experimental design, quasi-experimental design, or case studies? Participant observation or spectator observation? A few in-depth observations or many shorter observations? Single or multiple observers? Standardized or individualized protocols? Fixed or emergent design? Follow up after 2 weeks, 3 months, 6 months, or a year? Follow up everyone or a sample? What kind of sample: simple random, stratified, and/or purposeful? What size sample? Should interviewers have the same characteristics as program participants: gender, age, race? What comparisons to make: past performance, intended goals, hoped-for goals, other programs? I won't list a thousand such options à la Thomas Carlyle, but I've no doubt it could be done. I would certainly never try the patience of primary stakeholders with a thousand options, but I do expect to work with them to consider the strengths and weaknesses of major design and measurement possibilities.

Christie (2007) found that decision makers could distinguish among the merits and uses of different kinds of designs. Using a set of scenarios derived from actual evaluation studies, she conducted a simulation to examine what decision makers' reported as evaluation design preferences and likely influences. Each scenario described a setting where results from one of three types of evaluation designs would be available: large-scale study data, case study data, or anecdotal accounts. The simulation then specified a particular decision that needed to be made. Decision makers were asked to indicate which type of design would influence their decision making. Results from 131 participants indicated that participants were influenced

by all types of information, yet large-scale and case study data were more influential relative to anecdotal accounts; certain types of evaluation data were more influential among certain groups of decision makers; and choosing to use one type of evaluation data over the other two depended on the independent influence of other types of evaluation data on the decision maker, as well as prior beliefs about program efficacy. In essence, these decision makers had varying design preferences and were quite capable of distinguishing the credibility and utility of various types of evaluation studies—or measurement options. Let me illustrate with a common issue that arises in survey design.

The Odd-Even Question

Should response scales be even numbered (e.g., four or six response choices) or odd numbered (e.g., three or five choices)? It doesn't seem like such a big deal actually, but I've seen evaluators on both sides of the question go at each other with the vehemence of Marxists versus capitalists, osteopaths versus chiropractors, or cat lovers versus dog lovers. What's all the ruckus about? It's about the value and validity of a midpoint on questionnaire items. In conducting workshops on evaluation, one of the most common questions I get is "Should we give people a midpoint?"

An even-numbered scale has no midpoint.

Should the workshop be expanded from 1 day to 2 days?			
Strongly Agree	Agree	Disagree	Strongly Disagree

An odd-numbered scale has a midpoint.

Should the workshop be expanded from 1 day to 2 days?				
Strongly Agree	Agree	No Opinion	Disagree	Strongly Disagree

Even-numbered scales force respondents to lean in one direction or the other (although a few will circle the two middle responses creating their own midpoint if not provided one on the survey). Even-numbered scales allow the respondent to hedge, to be undecided, or, in less kind terms, to cop out of making a decision one way or the other, or yet again, to be genuinely in the middle.

One thing about surveys is clear: If given a midpoint, many respondents will use it. Not given a midpoint, most respondents will answer leaning one way or the other.

Which one is best? Should respondents be given a midpoint? Having carefully considered the arguments on both sides of the issue, having analyzed large number of questionnaires with both kinds of items, and having meditated on the problem at great length, I find that I'm forced to come down firmly and unwaveringly right smack in the middle. *It depends.* Sometimes odd-numbered scales

are best and sometimes even-numbered scales are best. How to decide?

The issue is really not technical, statistical, or methodological. The issue is one of utility. What do intended users want to find out? Will the findings be more useful if respondents are forced to lean in one direction or the other? Or is it more useful to find out what proportion of people are undecided, or "don't know." The evaluator helps the primary intended users determine the value and implications of offering a midpoint. Do they believe that "down deep inside" everyone really leans one way or the other on the issue, or do they believe that some people are genuinely in the middle on the issue and they want to know how many have no opinion?

Not only can nonresearchers make this choice, but they also often enjoy doing so, and engaging them in thinking about such alternatives and their implications teaches evaluative thinking.

Ensuring Methodological Quality and Excellence

I am easily satisfied with the very best.

—Winston Churchill (1874–1965)
British Prime Minister
during World War II

One of the myths believed by nonresearchers is that researchers have agreed among themselves about what constitutes methodological quality and excellence. This belief can make practitioners and other nonacademic stakeholders understandably reluctant to engage in methods discussions. In fact, as the next chapter discusses in depth, researchers disagree with each other vehemently about what constitutes good research and, with a little training and help, I find that nonresearchers can grasp the basic issues involved and make informed choices.

To increase the confidence of nonresearchers that they can and should contribute to methods discussions—for example, to consider the merits of telephone interviews versus face-to-face interviews or mail questionnaires—I'll often share the perspective of journal editors. Eva Baker, Director of the UCLA Center for the Study of Evaluation and former editor of *Educational Evaluation and Policy Analysis (EEPA),* established a strong system of peer review for *EEPA,* requiring three independent reviewers for every article. Eva has told me that in several years as editor, *she never published an article on which all three reviewers agreed the article was good!* I edited the peer-reviewed *Journal of Extension* for 3 years and had the same experience. Robert Donmoyer (1996), features editor of *Educational Researcher,* reported that "peer reviewers' recommendations often conflict and their advice is frequently contradictory. . . . There is little consensus about what research and scholarship are and what research reporting and scholarly discourse should look like" (p. 19).

This kind of inside look at the world of research, like an inside look at how the Supreme Court makes decision (Waldron 2007), can be shocking to people who think that there surely must be consensus regarding what constitutes "good" research or good jurisprudence. The real picture is more chaotic and warlike, what Donmoyer (1996) portrays as "a diverse array of voices speaking from quite different, often contradictory perspectives and value commitments" (p. 19). Perspectives and value commitments? Not just rules and formulas? Perspectives and value commitments imply stakes, which leads to stakeholders, which leads to involving stakeholders to represent their stakes,

Is There Agreement About What Constitutes Quality?

Robin Lin Miller became the distinguished editor of the *American Journal of Evaluation* in 2005, the profession's premier peer-reviewed scholarly journal. After three years experience editing *AJE*, I asked her how much consistency she found among reviewers.

In most cases, by which I mean 75 to 80% of papers submitted, reviewers' judgments follow one of two patterns. In the first, the plurality agrees that a paper requires extensive rewriting to make a contribution of any sort, though the reviewers may not agree on *why* the paper is flawed and how it might be improved to make a contribution. In the second, opinions on the merit of the paper diverge widely, with opinions about it scattered along a continuum; where one reviewer sees novelty and a significant advance another sees flaws that are beyond tolerance or repair. It is only in a minority of cases that consensus does emerge. When it does, it tends to favor the view that the paper makes little in the way of a contribution. Agreement that a paper is good occurs rarely.

—Robin Lin Miller, Editor, *American Journal of Evaluation*

SOURCE: Reprinted with permission of Robin Miller.

even in methods decisions, or should we say, *especially* in methods decisions, then those decisions determine what findings will be available for interpretation and use.

The evidence of disagreements about research standards and criteria for judging quality will not surprise those inside science who understand that a major thrust of methodological training in graduate school is learning how to pick apart and attack any study. There are no perfect studies. And there cannot be, for there is no agreement on what constitutes perfection.

This has important implications for methods decisions in evaluation. There are no universal and absolute standards for judging methods. The consensus that has emerged within evaluation, as articulated by the Omnibus Metaevaluation Checklist (Stufflebeam 2007), the Joint Committee Standards (1994) and the American Evaluation Association's Guiding Principles (Shadish et al. 1995) is that evaluations are to be judged on the basis of appropriateness, utility, practicality, accuracy, propriety, probity, credibility, and relevance. These criteria are necessarily situational and context bound. One cannot judge the adequacy of methods used in a specific evaluation without knowing the purpose of the evaluation, the intended uses of the findings, the resources available, and the trade-offs negotiated. Judgments about validity and reliability, for example, are necessarily and appropriately relative rather than absolute in that the rigor and quality of an evaluation's design and measurement depend on the purpose and *intended use* of the evaluation (Trochim 2006c). The Accuracy Standards of the Joint Committee on Standards (1994) make it clear that validity and reliability of an evaluation depend on the intended use(s) of the evaluation.

> *Valid Information:* The information-gathering procedures should be chosen or developed and then implemented so that they will assure that the interpretation arrived at is valid *for the intended use* [italics added]. (P. A5)

> *Reliable Information:* The information-gathering procedures should be chosen or developed and then implemented so that they will assure that the information obtained is sufficiently reliable *for the intended use* [italics added]. (P. A6)

The Art of Making Methods Decisions

Lee J. Cronbach (1982), an evaluation pioneer and author of several major books on measurement and evaluation, observed that designing an evaluation is as much art as science: "Developing an evaluation is an exercise of the dramatic imagination" (p. 239). This perspective can help free practitioners and other primary users who are nonresearchers to feel they have something important to contribute. It may also open the evaluator to valuing their contributions and facilitating their "dramatic imaginations." The art of evaluation involves creating a design that is appropriate for a specific situation and particular action or policy-making context. In art there is no single, ideal standard. Beauty is in the eye of the beholder, and the evaluation beholders include decision makers, policymakers, program managers, practitioners, participants, and the general public. Thus, any given design is necessarily an interplay of resources, possibilities, creativity, and personal judgments by the people involved.

No Perfect Design

"There is no single best plan for an evaluation, not even for an inquiry into a particular program, at a particular time, with a particular budget."

Lee J. Cronbach (1982:231), *Designing Evaluations of Educational and Social Programs*. Cronbach directed the Stanford Evaluation Consortium and was President of the American Educational Research Association, the American Psychological Association, and the Psychometric Society. He was also a member of the National Academy of Sciences and the American Academy of Arts and Sciences.

Instead of one massive experiment or quasi-experiment (the "horse race" model of evaluation), said Cronbach, he favored an eclectic, broad-based, open methodological approach to evaluation; a fleet of smaller studies, each pursuing an important case or component of the policy or program under study. Cronbach encouraged evaluators to design evaluations to understand in some depth the nature of each context and the quality of the intervention in that context. Over time, then, with many such studies, the policy-shaping community could learn in some depth about the social problem and how best to address it. In addition, Cronbach encouraged evaluators to involve members of the setting in the evaluation study and to provide feedback throughout the course of the study (for program improvement purposes) rather than just at the end.

SOURCE: *Encyclopedia of Evaluation* (2005:95).

Still, for nonresearchers, being expected to participate in design decisions can be intimidating. Evaluators reinforce and deepen any nascent inclination toward feeling intimidated by beginning with an emphasis on the importance of establishing "a theoretically sound conceptual framework" for the evaluation design. A theoretically sound conceptual framework sounds like something *The Music Man* would sell to the good people of River City to keep their children out of pool halls.

Conceptualizing an evaluation framework doesn't require some grandiose and theoretical posture or a voluminous and vortiginous vocabulary. That grand old evaluation savant Rudyard Kipling has offered all the conceptual framework one needs

I keep six honest serving men

They taught me all I knew:

Their names are What and Why and When

And How and Where and Who.

I find that the people with whom I'm working warm quickly to the task of designing the evaluation when I recall for them Kipling's "conceptual framework."

What? What do we want to find out?

Why? Why do we want to find that out?

When? When do we need the information?

How? How can we get the information we need?

Where? Where should we gather information?

Who? Who is the information for and from whom should we collect the information we need?

These questions guide the primary focus in making evaluation measurement and methods decisions—getting the best possible data to adequately answer primary users' evaluation questions given available resources and time. The emphasis is on *appropriateness and credibility*—measures, samples, and comparisons that are appropriate and credible to address key evaluation issues. Supplementing Kipling's framework, Exhibit 11.2 presents guidance on the scope and selection of information for an evaluation from Dan Stufflebeam's (2007) very useful *Omnibus Metaevaluation Checklist*.

Hard versus Soft Data

The next chapter will explore in depth the "paradigms debate" involving quantitative/experimental methods versus qualitative/naturalistic approaches. This is sometimes framed as "hard data" versus "soft data" or numbers versus narrative. At this point, it suffices to say that the issue is not hard versus soft but relevant and appropriate versus irrelevant and inappropriate. Participants in the Stanford Evaluation Consortium

EXHIBIT 11.2

Information Scope and Selection

Select and collect a range of information that is sufficient to judge the program's merit, worth, significance, and probity and address key questions of interest to clients and specified stakeholders.

- Determine and document the client's most important evaluation requirements.
- Interview stakeholders to determine their different perspectives on the program.
- Effect evaluator and client agreements on the evaluation's questions and required information.
- Assign priorities to the evaluation's questions and associated information requirements.
- Allocate the evaluation effort in accordance with the priorities assigned to the needed information.
- Allow flexibility for adding questions during the evaluation.
- Obtain sufficient information to address the stakeholders' most important evaluation questions, as appropriate.

SOURCE: Stufflebeam (2007:U3).

(Cronbach et al. 1980) observed that "merit lies not in form of inquiry but in relevance of information" (p. 7). My experience with stakeholders suggests that they would rather have "soft data" about an important question than "hard data" about an issue of less relevance.

Obviously, the ideal is hard data about important questions, *whatever hard data may mean in a particular context*. But in the real world of trade-offs and negotiations, the evaluator too often determines what is evaluated according to his or her own expertise or preference in what to measure, rather than by deciding first what intended users determine is worth evaluating and then doing the best he or she can with methods. Methods are employed in the service of relevance and use, not as their master. Exhibit 11.3 contrasts three pragmatic versus ideal design trade-offs.

EXHIBIT 11.3

Pragmatic Design Principles

Principles offer directional guidelines. They are not recipes, laws, or concrete, absolute prescriptions. Principles help in dealing with trade-offs in the less than perfect real world of evaluation design. Below are three ideals contrasted with three pragmatic options when the ideals cannot be achieved because of real-world constraints. These can be used to generate discussion and get people thinking about which way to lean when faced with tough choices.

Evaluation Ideal	*Pragmatic Principle*
1. Get the best possible data to affect decisions.	1. Less-than-perfect data *available in time to affect decisions* are better than more-perfect data available *after* decisions have been taken.
2. "Hard" data on all questions.	2. Softer data on important questions are better than harder data on less important questions (whatever "softer" and "harder" may mean in a particular context).
3. More and better data	3. To avoid information overload, *less can be more* when data are appropriately focused on priority questions and uses.

One implication of this perspective—that quality and excellence are situational, that design combines the scientific and artistic—is that it is futile to attempt to design studies that are immune from methodological criticism. There simply is no such immunity. Intended users who participate in making methods decisions should be prepared to be criticized regardless of what choices they make. Especially futile is the desire, often articulated by non-researchers, to conduct an evaluation that will be accepted by and respected within the academic community. As we demonstrated above in discussing peer-review research, the academic community does not speak

with one voice. Any particular academics whose blessings are particularly important for evaluation use should be invited to participate in the evaluation design task force and become, explicitly, intended users. Making no pretense of pleasing the entire scientific community (an impossibility), utilization-focused evaluation strives to attain the more modest and attainable goal of pleasing primary intended users. This does not mean that utilization-focused evaluations are less rigorous. It means the criteria for judging rigor must be articulated for each evaluation.

Credibility and Use

Credibility affects use. Credibility includes the perceived accuracy, fairness, and believability of the evaluation *and* the evaluator. In the Joint Committee's (1994) standard on Evaluator Credibility, evaluators are admonished to be "both trustworthy and competent" so that findings achieve "maximum credibility and acceptance" (p. U2). Report clarity, full and frank disclosure of data strengths and weaknesses, balanced reporting, defensible information sources, valid and reliable measurement, justified conclusions, and impartial reporting are all specific standards aimed at credibility as a foundation for use.

For information to be useful and to merit use, it should be as accurate and believable as possible. Limitations on the degree of accuracy should be stated clearly. Decision makers want highly accurate and trustworthy data. This means they want data that are valid and reliable. But in the politically charged environment of evaluation, these traditional scientific concepts have taken on some new and broader meanings.

Overall Evaluation Validity

> *The government ministries are very keen on amassing statistics. They collect them, raise them to the nth power, take the cube root, and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first place from the village watchman, who just puts down what he damn well pleases.*
>
> —Sir Josiah Stamp, 1911,
> English economist (1880–1941)

House (1980:249) has suggested that validity means "worthiness of being recognized": For the typical evaluation, this means being "true, credible, and right" (p. 250). Different approaches to evaluation establish validity in different ways. House applies the notion of validity to *the entire evaluation,* not just the data. An *evaluation* is perceived as valid in a global sense that includes the overall approach used, the stance of the evaluator, the nature of the process, the design, data gathering, and the way in which results are reported. Both the evaluation *and* the evaluator must be perceived as trustworthy for the evaluation to have high validity.

Alkin et al. (1979) studied use and found that "for evaluation to have impact, users must believe what evaluators have to say" (p. 245). The believability of an evaluation depends on much more than the perceived scientific validity of the data and findings. Believability depends on the users' perceptions of and experiences with the program being evaluated, users' prior knowledge and prejudices, the perceived adequacy of evaluation procedures, and the users' trust in the evaluator (Alkin et al. 1979:245–47). Trust, believability, and credibility are the underpinnings of *overall* evaluation validity.

Evaluation Design Checklist

Dan Stufflebeam (2004b) has developed a design checklist that is a generic guide to decisions one typically needs to consider when planning and conducting an evaluation. The checklist presents the logical structure of evaluation design and includes elements that commonly apply to a wide range of evaluation assignments and alternative evaluation approaches. The checklist is intended for use across a broad range of evaluation assignments—both small and large—and for use with a number of different approaches to evaluation. His introduction to the checklist describes the typically iterative and cycling nature of the design process as the evaluation is brought into focus and later adapted to changed understandings and emergent conditions.

When the contemplated evaluation is small in scope and will have only a modest budget, evaluators and their clients can find it useful to consider the full range of evaluation design issues before setting aside those that are not feasible, not particularly relevant to the situation, or especially important. . . . The user will need to exercise good judgment and discretion in determining and applying the most applicable parts of the checklist pursuant to the needs of particular evaluations. This checklist is intended both as an advance organizer and as a reminder of key matters to be considered before and during an evaluation. An ordered list of elements commonly included in evaluation designs is included but these elements are not necessarily intended to be treated in a strict linear sequence. Often, one cycles through the elements repeatedly while planning for and negotiating an evaluation and also during the course of the evaluation. In each such cycle, some elements are addressed, while others typically are set aside for attention later or abandoned because they don't apply to the particular situation. Evaluation design is as much process as product. In using this checklist the objective should be, over time, to evolve an evaluation plan to undergird a sound, responsive, and effective evaluation.

It is emphasized that evaluators and their clients are wise to revisit evaluation design decisions throughout the evaluation, especially as new questions and circumstances emerge. The following, then, is an ordered set of issues to consider when planning, conducting, and reporting an evaluation.

- Focusing the evaluation
- Collecting information
- Organizing information
- Analyzing information
- Reporting information
- Administering the evaluation

SOURCE: Stufflebeam (2004).

It is important to understand how overall evaluation validity differs from the usual, more narrow conception of validity in scientific research. Validity is usually focused entirely on data collection procedures, design, and technical analysis, that is, whether measures were valid or whether the design allows drawing inferences about causality (internal design validity).

A measure is scientifically valid to the extent that it captures or measures the concept it is intended to measure. For example, asking if an IQ test really measures native intelligence (rather than education and socioeconomic advantage) is a validity question. Validity is often difficult to establish, particularly for new instruments. Over time, scientists develop some consensus about the

relative validity of oft-used instruments, such as major norm-referenced standardized educational tests. Rossi, Freeman, and Wright (1979) posited three common criteria for validity of quantitative instruments.

1. *Consistency with Usage:* A valid measurement of a concept must be consistent with past work that used that concept. Hence, a measure of adoption must not be in contradiction to the usual ways in which that term had been used in previous evaluations of interventions.

2. *Consistency with Alternative Measures:* A valid measure must be consistent with alternative measures that have been used effectively by other evaluators. Thus, a measure must produce roughly the same results as other measures that have been proposed, or, if different, have sound conceptual reasons for being different.

3. *Internal Consistency:* A valid measure must be internally consistent. That is, if several questions are used to measure adoption, the answers to those questions should be related to each other as if they were alternative measures of the same thing (pp. 170–71).

Qualitative methods (e.g., such techniques as participant observation and in-depth, open-ended interviewing) pose different validity challenges. In qualitative methods, validity hinges to a greater extent on the skill, competence, and rigor of the researcher because *the observer or interviewer is the instrument*.

> Since as often as not the naturalistic inquirer is himself the instrument, changes resulting from fatigue, shifts in knowledge, and cooptation, as well as variations resulting from differences in training, skill, and experience among different "instruments," easily occur. But this loss in rigor is more than offset by the flexibility, insight, and ability to build on tacit knowledge that is the peculiar province of the human instrument. (Guba and Lincoln 1981:113)

Validity concerns also arise in using official statistics such as health or crime statistics. Joe Hudson (1977) has cautioned about the care that must be taken in using crime statistics because of validity problems:

> First, officially collected information used as measures of program outcomes are, by their very nature, indirect measures of behavior. For example, we have no practical or direct way of measuring the actual extent to which graduates of correctional programs commit new crimes. Second, the measurements provided are commonly open to serious problems. For example, the number of crimes known to authorities in most situations is only a fraction of the number of crimes committed, although that fraction varies from crime to crime. . . . The growing willingness of victims of sexual assault to report their crimes to the police and actively cooperate in prosecution is an example of the manner in which public attitudes can affect officially recorded rates of crime.
>
> Of the various criteria used to measure recidivism, that of arrest appears to be especially problematic. Recidivism rates based on arrest do not tell us whether those arrested have, in fact, returned to criminal behavior but only that they are presumed to have done so. . . . The widespread discretion exercised by the police to arrest is a further source of invalidity. For example, it is probably reasonable to expect that the number of individuals arrested for a particular type of crime within a jurisdiction is to some extent a direct reflection of changing police policies and not totally the function of changing patterns of law-violating behavior. In addition to the power of deciding when to arrest, police also have discretionary authority to determine which of a number of crimes an individual will be arrested for in a particular situation. Thus, if policy

> emphasis is placed upon combating burglary, this may affect decisions as to whether an arrestee is to be arrested for burglary, simple larceny, or criminal damage to property. In short, the discretion of the police to control both the number and types of arrests raises serious validity problems in evaluations which attempt to use this measure of program outcome. (Pp. 88–89)

In summary, then, validity problems, along with the trustworthiness of the evaluator, affect the overall credibility of the evaluation, and this is true for all kinds of data collection—quantitative measures, questionnaires, qualitative observations, government statistics, and social indicators. The precise nature of the validity problem varies from situation to situation, but evaluators must always be concerned about the extent to which the data collected are credible and actually measure what is supposed to be measured; they must also make sure that intended users understand validity issues. In addition, a validity issue of special, though not unique, concern to utilization-focused evaluators is *face validity*.

Face Validity in Utilization-Focused Measurement

Face validity concerns the extent to which an instrument *looks* as if it measures what it is intended to measure (Trochim 2006a). An instrument has face validity if stakeholders can look at the items and understand what is being measured. From a utilization-focused perspective, it is perfectly reasonable for decision makers to want to understand and believe in data they are expected to use. Face validity, however, is generally held in low regard by measurement experts. Predictive validity, concurrent validity, and construct validity—these technical approaches are much preferred by psychometricians. Nunnally (1970), in his classic work on psychometrics, considered face validity to have some possible value when data are gathered for the general public, but he concluded, "Although one could make a case for the involvement of face validity in the measurement of constructs, to do so would probably serve only to confuse the issues" (p. 150). To deepen our understanding of the issue, consider the following case.

Face Validity

In face validity, you look at the operationalization and see whether "on its face" it seems like a good translation of the construct. This is probably the weakest way to try to demonstrate construct validity. For instance, you might look at a measure of math ability, read through the questions, and decide that yep, it seems like this is a good measure of math ability (i.e., the label "math ability" seems appropriate for this measure). Or, you might observe a teenage pregnancy prevention program and conclude, "Yep, this is indeed a teenage pregnancy prevention program." Of course, if this is all you do to assess face validity, it would clearly be weak evidence because it is essentially a subjective judgment call. (Note that just because it is weak evidence doesn't mean that it is wrong. We need to rely on our subjective judgment throughout the research process. It's just that this form of judgment won't be very convincing to others.) We can improve the quality of face validity assessment considerably by making it more systematic. For instance, if you are trying to assess the face validity of a math ability measure, it would be more convincing if you sent the test to a carefully selected sample of experts on math ability testing, and they all reported back with the judgment that your measure appears to be a good measure of math ability.

SOURCE: Trochim (2006a).

The board of directors of a major industrial firm decided to decentralize organizational decision making in hopes of raising worker morale. The president of the company hired an organizational consultant to monitor and evaluate the decentralization program and its effects. From the literature on the sociology of organizations, the evaluator selected a set of research instruments designed to measure decentralization, worker autonomy, communication patterns, and worker satisfaction. The scales had been used by sociologists to measure organizational change in a number of different settings, and the factorial composition of the scales had been validated. The instruments had high predictive and construct validity, but low face validity—that is, a nonresearcher could not look at the items and tell what they were measuring; interpretation depended on understanding factor analysis.

The evaluator found no statistically significant changes between pre- and posttest, so when he met with the board of directors, he dutifully reported that the decentralization program had failed and that worker morale remained low. The president of the company had a considerable stake in the success of the program; he did not have a stake in the evaluation data. He did what decision makers frequently do in such cases—he attacked the data.

President: How can you be so sure that the program failed?

Evaluator: We collected data using the best instruments available. I won't go into all the technical details of factor analysis and Cronbach's alpha. Let me just say that these scales have been shown to be highly valid and reliable. Take this 10-item scale on *individual autonomy*. The best predictor item in this particular scale asks respondents (a) "Do you take coffee breaks on a fixed schedule?" or (b) "Do you go to get coffee whenever you want to?"

President: [visibly reddening and speaking in an angry tone] Am I to understand that your entire evaluation is based on some kind of questionnaire that asks people how often they get coffee, that you never personally talked to any workers or managers, that you never even visited our operations? Am I to understand that we paid you $20,000 to find out how people get their coffee?

Evaluator: Well, there's a lot more to it than that, you see . . .

President: That's it! We don't have time for this nonsense. Our lawyers will be in touch with you about whether we want to press fraud and malpractice charges!

Clearly, the President was predisposed to dismiss any negative findings. But suppose the evaluator had reviewed the instrument and survey design with the president before gathering data. Suppose he had explained what the items were supposed to indicate and then asked,

> Now, if we survey employees with these items measuring these factors, will they tell you what you want to know? Does this make sense to you? Are you prepared to act on this kind of data? Would you believe the results if they came out negative?

Such an exchange might not have made a difference. It's not easy to get busy executives to look carefully at instruments in advance, nor do evaluators want to waste time explaining their trade. Many decision

makers are just as happy not being bothered with technical decisions. After all, that's why they hired an evaluator in the first place, to design and conduct the evaluation! But the costs of such attitudes to use can be high. Utilization-focused evaluators check out the face validity of instruments before data are collected. Subsequent data analysis, interpretation, and use are all facilitated by attention to face validity—making sure users understand and believe in the data.

Useful Designs

Face validity criteria can also be applied to design questions. Do intended users understand the design? Does it make sense to them? Do they appreciate the implications of comparing Program A with Program B? Do they know why the design includes, or does not include, a control group? Is the sample size sufficiently large to be believable? You can be sure that decision makers will have opinions about these issues when results are presented, particularly if findings turn out negative. By asking these questions before data collection, potential credibility problems can be identified and dealt with, and users' insights can help shape the design to increase its relevance. Consider the following case from an evaluation workshop I conducted.

The marketing director for a major retail merchandising company attended to find out how to get more mileage out of his marketing research department. He explained that 2 years earlier he had spent a considerable sum researching the potential for new products for his company's local retail distribution chain. A carefully selected representative sample of 285 respondents had been interviewed in the Minneapolis-Saint Paul greater metropolitan area. The results indicated one promising new line of products for which there appeared to be growing demand. He took this finding to the board of directors with a recommendation that the company make a major capital investment in the new product line. The board, controlled by the views of its aging chairman, vetoed the recommendation. The reason: "If you had presented us with opinions from at least a thousand people, we might be able to move on this item. But we can't make a major capital commitment on the basis of a couple of hundred interviews."

The marketing director tactfully tried to explain that increased sample size would have made only a marginal reduction in possible sampling error. The chairperson remained unconvinced, the findings of an expensive research project were ignored, and the company missed out on a major opportunity. A year later, the item they rejected had become a fast-selling new product for a rival company.

It is easy to laugh at the board's mistake, but the marketing director was not laughing. He wanted to know what to do. I suggested that next time, he check out the research design with the board before collecting data, going to them and saying,

> Our statistical analysis shows that a sample of 285 respondents in the Twin Cities area will give us an accurate picture of market potential. Here are the reasons they recommend this sample size. . . . Does that make sense to you? If we come in with a new product recommendation based on 285 respondents, will you believe the data?

If the board responds positively, the potential for use will have been enhanced, though not guaranteed. If the board says the sample is too small, then the survey might as well include more respondents—or be canceled. There is little point in implementing a design that is known in advance to lack credibility.

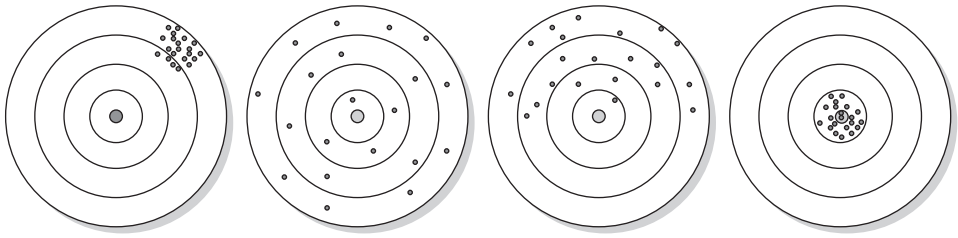
Reliability and Error

Reliability has to do with consistency. A measure is reliable to the extent that essentially the same results can be reproduced repeatedly, as long as the situation does not change. For example, in measuring the height of an adult, one should get the same results from one month to the next. Measuring attitudes and behavior is more complex because one must determine whether measured change means the attitude has changed or the data collection is unreliable. Exhibit 11.4 presents the relationship between reliability and validity.

EXHIBIT 11.4

Reliability and Validity

Bill Trochim, 2008 President of the American Evaluation Association, has a favorite metaphor he used to explain the relationship between reliability and validity. He begins by comparing the center of a target with the concept the evaluator is trying to measure. "Imagine that for each person you are measuring, you are taking a shot at the target. If you measure the concept perfectly for a person, you are hitting the center of the target. If you don't, you are missing the center. The more you are off for that person, the further you are from the center."



Reliable not valid | **Valid not reliable** | **Neither reliable not valid** | **Both reliable and valid**

The figures above show four possible situations. In the first one, the evaluator hits the target consistently, but misses the center of the target—consistently and systematically measuring the wrong value for all respondents. Such a measure is reliable, but not valid (that is, it's consistent but wrong). On the second target, the evaluator measurement efforts are randomly spread across the target, seldom hitting the center of the target but, on average, getting the right answer for the group (though not very accurate for individuals). The result is a valid, but inconsistent group estimate. The third target shows hits spread across the target and consistently missing the center. This measure is neither reliable nor valid. The final target displays what Trochim calls the "Robin Hood" scenario after the infamous medieval archer—the evaluator consistently hits the center of the target making the measure both reliable and valid.

SOURCE: Trochim (2006b). Reprinted with permission of William M. Trochim.

Inconsistent data collection procedures, for example, asking interview questions in different sequence to different respondents, can change results and introduce errors. Nonresearchers will often have unrealistic expectations about evaluation instruments, expecting no errors. For many reasons, all data collection is subject to some measurement error. Henry Dyer, a former president of the highly respected Educational Testing Service, tells of trying to explain to a government official that test scores, even on the most reliable tests, have enough measurement error that they must be used with understanding of their limitations. The high-ranking official responded that test makers should "get on the ball" and start producing tests that "are 100% reliable under all conditions."

Dyer's (1973) reflections on this conversation are relevant to an understanding of error in all kinds of measures. He asked,

> How does one get across the shocking truth that 100% reliability in a test is a fiction that, in the nature of the case, is unrealizable? How does one convey the notion that the test-reliability problem is not one of reducing measurement error to absolute zero, but of minimizing it as far as practicable and doing one's best to estimate whatever amount of error remains, so that one may act cautiously and wisely in a world where all knowledge is approximate and not even death and taxes are any longer certain? (P. 87)

Sources of error are many. For example, consider sources of error in an individual test score. Poor health on the day of the test can affect the score. Whether the student had breakfast can make a difference. Noise in the classroom, a sudden fire drill, whether or not the teacher or a stranger gives the test, a broken pencil, and any number of similar disturbances can change a test score. The mental state of the child—depression, boredom, elation, a conflict at home, a fight with another student, anxiety about the test, low self-confidence—can affect how well the student performs. Simple mechanical errors such as marking the wrong box on the test sheet by accident, inadvertently skipping a question, or missing a word while reading are common problems for all of us. Students who have trouble reading will perform poorly on reading tests, but they are also likely to perform poorly on social studies, science, and math tests.

Some children perform better on tests because they have been taught how to make written tests. Some children are simply better test takers than other children because of their background or personality or because of how seriously they treat the idea of the test. Some schools make children sit all day taking test after test, sometimes for an entire week. Other schools give the test for only a half-day or 2 hours at a time to minimize fatigue and boredom. Some children like to take tests; some don't. Some teachers help children with difficult words, or even read the tests along with the children; others don't. Some schools devote their curriculum to teaching students what is on the tests. Others place little emphasis on test taking and paper-and-pencil skills, thus giving students less experience in the rigor and tricks of test taking.

All these sources of error—and I have scarcely scratched the surface of possibilities—can seriously affect an individual score. Moreover, they have virtually nothing to do with how good the test is, how carefully it was prepared, or how valid its content is for a given child or group. Intrinsic to the nature of testing, these errors are always present to some extent and are largely uncontrollable. These are the reasons why statisticians can never develop a test that is 100 percent reliable.

The errors are more or less serious depending on how a test is used. When looking at test scores for large groups, we can expect that, because of such errors, some students will perform above their true level and other students will perform below their true score. For most groups, statisticians believe that these errors cancel each other. The larger the group tested, the more likely this is to be true.

Different evaluation instruments are subject to different kinds of errors. Whether the evaluation includes data from tests, questionnaires, management information systems, government statistics, or whatever—the analysis should include attention to potential sources of error, and, where possible, calculate and report the degree of error. The point is that evaluators need not be defensive about errors. Rather, they need to explain the nature of errors, help intended users decide what level of precision is needed, consider the costs and benefits of undertaking procedures to reduce error (for instance, a larger sample size), and help users to understand the implications for interpreting findings. Primary intended users can be helpful in identifying potential sources of error. In my experience, their overall confidence in their ability to correctly and appropriately use evaluation data is increased when there has been a frank and full discussion of *both* the data's strengths and weaknesses. In this way, evaluators help make evaluation clients more knowledgeable so they will understand what Dyer's government official did not: The challenge is not reducing measurement error to absolute zero, but rather minimizing it as far as practicable and doing one's best to estimate whatever amount of error remains, so that one may act cautiously and wisely in a world where all knowledge is approximate and not even death and taxes are any longer certain.

Trade-Offs

Different evaluation purposes affect how much error can be tolerated. A summative evaluation to inform a major decision that will affect the future of a program, perhaps touching the lives of thousands of people and involving allocations of millions of dollars, will necessarily and appropriately involve considerable attention to and resources for minimizing error. In contrast, a small-scale, fairly informal, formative evaluation aimed at stimulating staff to think about what they're doing will raise fewer concerns about error. There is a lot of territory between these extremes. How precise and robust findings need to be, given available resources, are matters for discussion and negotiation. The next two sections look at additional concerns that commonly involve negotiation and trade-offs: (1) breadth versus depth and (2) the relative generalizability of findings.

Breadth versus Depth

Deciding how much data to gather involves trade-offs between depth and breadth. Getting more data usually takes longer and costs more, but getting less data usually reduces confidence in the findings. Studying a narrow question or very specific problem in great depth may produce clear results but leave other important issues and problems unexamined. On the other hand, gathering information on a large variety of issues and problems may leave the evaluation unfocused and result in knowing a little about a lot of things, but not knowing a lot about anything.

During methods deliberations, some boundaries must be set on data collection. Should all parts of the program be studied or only certain parts? Should all participants be studied or only some subset of clients? Should the evaluator aim at

describing all program processes and outcomes or only certain priority areas?

In my experience, determining priorities is challenging. Once a group of primary stakeholders gets turned on to learning from evaluative information, they want to know everything. The evaluator's role is to help them move from a rather extensive list of potential questions to a much shorter list of realistic questions and finally to a focused list of essential and necessary questions. This process moves from divergence to convergence, from generating many possibilities (divergence) to focusing on a few worthwhile priorities (convergence).

This applies to framing overall evaluation questions as well as to narrowing items in a particular instrument, such as a survey or interview. Many questions are interesting, but which are crucial? These end up being choices not between good and bad, but among alternatives, all of which have merit.

Internal and External Validity in Design

Trade-offs between internal and external validity have become a matter of debate in evaluation since Campbell and Stanley (1963) asserted that "internal validity is the sine qua non" (p. 175). Internal validity in its narrowest sense refers to certainty about cause and effect. Did *X* cause *Y*? Did the program intervention cause the observed outcomes? In a broader sense, it refers to the "trust-worthiness of an inference" (Cronbach 1982:106). External validity, on the other hand, refers to the degree of confidence one has in generalizing findings beyond the situation studied.

Internal validity is increased by exercising rigorous control over a limited set of carefully defined variables. However, such rigorous controls create artificialities that limit generalizability. The highly controlled situation is less likely to be relevant to a greater variety of more naturally occurring, less controlled situations. In the narrowest sense, this is the problem of going from the laboratory into real world. In contrast, increasing variability and sampling a greater range of experiences or situations typically reduces control and precision, thereby reducing internal validity. The ideal is high-internal validity and high-external validity. In reality, there are typically trade-offs involved in the relative emphasis placed on one or the other.

Cronbach's (1982) discussion of these issues for evaluation is quite comprehensive and insightful. He emphasized that "both external validity and internal validity are natters of degree and external validity does not depend directly on internal validity" (p. 170). Being able to apply findings to future decisions and new settings is often more important than establishing rigorous causal relations under rigid experimental conditions. He introduced the idea of *extrapolation* rather than generalization. Extrapolation involves logically and creatively thinking about what specific findings mean for other situations, rather than the statistical process of generalizing from a sample to a larger population. He advocated that findings be interpreted in light of stakeholders' and evaluators' experiences and knowledge, and then applied or extrapolated using all available insights, including understanding about quite different situations. This focuses interpretation away from trying to determine truth in some absolute sense (a goal of basic research) to a concern with conclusions that are reasonable, justifiable, plausible, warranted, and useful.

The contrasting perspectives of Campbell (emphasis on internal validity) and Cronbach (emphasis on external validity) have elucidated the trade-offs between designs that give first priority to certainty about casual inference versus those that better support extrapolations to new

settings. These evaluation pioneers formulated fundamentally different theories of practice (Shadish et al. 1991). In working with primary stakeholders to design evaluations that are credible, the evaluator will need to consider the degree to which internal and external validity are of concern, and to emphasize each in accordance with stakeholder priorities. Choices are necessitated by the fact that no single design is likely to attain internal and external validity in equal degrees.

Demand Validity and Consequential Validity

A quite different perspective on validity is "demand validity," a validity that comes from participants and people in communities affirming the value of a program and "demanding" continuation (or expansion). Lois-ellen Datta, one of evaluation's distinguished pioneers and former president of the Evaluation Research Society, originated this concept to describe the Head Start Program when it began in the 1960s: "The program obviously had face validity and demand validity" (Datta 2004:246). By this she meant that in her evaluation fieldwork, she had come to place emphasis on what parents and children reported, tending to "factor in a bit of self-interest when paid staff testify."

When parents tell me of sitting for the first time "at the table," when they speak of how Head Start brought them pride and dignity, as well as hope for their children, when I see (as I did in 1968) a child so handicapped, he was drawn about in a little old red wagon yet integrated joyously into these simple programs—these to me are like a Seurat painting in contrast to the Sherlockian, subtractive approach. The concept perhaps gains a bit of strength because, in my experience, parent/participant stories are not always sunshine and roses. I will hear about the problems, limitations, what should be happening but isn't (Datta, L. 2007, Personal communication).

Consequential validity as a criterion for judging an evaluation design or instrument makes the social consequences of its use a value basis for assessing its credibility and utility. Thus, standardized achievement tests are criticized because of the discriminatory consequences for minority groups of educational decisions made with "culturally biased" tests. Consequential validity asks for assessments of who benefits and who is harmed by an inquiry, measurement, or method (Thomas 2005). Exhibit 11.6 at the end of this chapter presents a discussion about the validity and consequences of various ways from around the world of gathering data on the racial and ethnic backgrounds of program participants.

Truth and Utility

Stakeholders want accurate information; they apply "truth tests" (Weiss and Bucuvalas 1980) in deciding how seriously to pay attention to an evaluation. They also want useful and relevant information. The ideal, then, is both truth and utility. In the real world, however, there are often choices to be made between the extent to which one maximizes truth and the degree to which data are relevant. The simplest example of such a choice is time. The time lines for evaluation are often ridiculously short. A decision maker may need whatever information can be obtained in 3 months, even though researchers insist that a year is necessary to get data of reasonable quality and accuracy. This involves a trade-off between truth and utility. Highly accurate data in a year are less useful to this decision maker than data of less precision and validity obtained in 3 months.

Decision makers regularly face the need to take action with limited and imperfect information. They prefer more accurate information to less accurate information, but they also prefer some information to no information. This is why research quality and rigor are "much less important to utilization than the literature might suggest" (Alkin et al. 1979:24).

The effects of methodological quality on use must be understood in the full context of a study, its political environment, the degree of uncertainty with which the decision maker is faced, and thus his or her relative need for any and all clarifying information. If information is scarce, then new information, even of less-than-ideal quality, may be somewhat helpful.

The scope and importance of an evaluation greatly affect the emphasis that will be placed on technical quality. Eleanor Chelimsky (2006a, 2006b, 1987a, 1987b), former president of the American Evaluation Association and founding Director of the Program Evaluation and Methodology Division of the U.S. Government Accountability Office, has insisted that technical quality is paramount in policy evaluations to Congress. The technical quality of national policy research matters, not only in the short term, when findings first come out, but over the long term as policy battles unfold and evaluators are called on to explain and defend important findings (Chelimsky 1995a).

On the other hand, debates about technical quality are likely to be much more center stage in national policy evaluations than in local efforts to improve programs at the street level, where the policy rubber hits the day-to-day programming road. One evaluator in our study of the use of federal health studies linked the issue of technical quality to the nature of uncertainty in organizational decision making. He acknowledged inadequacies in the data he had collected, but he had still worked with his primary users to apply the findings, fully recognizing their problematic nature:

> You have to make the leap from very limited data. I mean, that's what a decision's like. You make it from a limited data base; and, damn it, when you're trying to use quantitative data and it's inadequate, you supposedly can't make a decision. Only you're not troubled by that. You can use impressionistic stuff. Yeah, your intuition is a lot better. I get a gestalt out of this thing on every program.
>
> This may come as a great shock to you, but that is what you use to make decisions. In Chester Barnard's definition, for example, the function of the executive is to make a decision in the absence of adequate information. [EV148:11]

He went on to express some pride in the cost-benefit ratio of this evaluation, despite admitted methods inadequacies:

> It was a pretty small investment on the part of the government—$47,000 bucks. In the evaluation business that's not a pile of money. The questions I had to ask were pretty narrow and the answers were equally narrow and relatively decisive, and the findings were put to use immediately and in the long term. So, can you beat that? [EV148:8]

Another evaluator expressed similar sentiments about a study that had to be completed in only 3 months.

> There are a million things I'd do differently. We needed more time. . . . At the time, it was probably the best study we could do. . . . I'm satisfied in the sense that some people found it useful. It wasn't just kept on a shelf. People paid attention to that study and it had an impact. Now, I've done other studies that I thought were methodologically really much more elegant that were kind of ignored, just sitting on somebody's shelf.
>
> My opinion is that this really modest study probably has had impact all out of proportion to the quality of the research. It happened to

> be at a certain place at a certain time, where it at least talked about some of the things that people were interested in talking about, so it got some attention. And many other studies that I know of that have been done, that I would consider of higher quality, haven't really gotten used. [EV145:34]

Technical quality (truth tests) may get less attention than researchers desire because many stakeholders are not very sophisticated about methods. Yet they know (almost intuitively) that the methods and measurements used in any study are open to question and attack, a point emphasized earlier in this chapter. They know that experts often disagree among themselves. As a result, experienced decision makers often apply less rigorous standards than academics and, as long as they find the evaluation effort credible and serious, they're more interested in discussing the substance of findings than in debating methods. Credibility involves more than technical quality, though that is an important contributing factor. Credibility, and therefore utility, are affected by "the steps we take to make and explain our evaluative decisions, [and] also intellectually, in the effort we put forth to look at all sides and all stakeholders of an evaluation" (Chelimsky 1995a:219). The perception of impartiality is at least as important as methodological rigor in highly political environments.

Another factor that can reduce the weight decision makers give to technical quality is skepticism about the return on investment of large-scale, elaborately designed, carefully controlled, and expensive studies. Cohen and Weiss (1977) reviewed 20 years of policy research on race and schools, finding progressive improvement in research methods (i.e., increasingly rigorous designs and ever more sophisticated analytical techniques). Sample sizes increased, multiple regression and path analytic techniques were employed, and more valid and reliable data-gathering instruments were developed. After reviewing the findings of studies produced with these more rigorous methods, as well as the uses made of their findings, they concluded that "these changes have led to more studies that disagree, to more qualified conclusions, more arguments, and more arcane reports and unintelligible results" (Cohen and Weiss 1977:78). In light of this finding, simple, understandable, and focused evaluations have great appeal to practitioners and action-oriented evaluation users.

In utilization-focused evaluation, attention to technical quality is tied to and balanced by concern for relevance and timeliness. As one decision maker in our federal health evaluation study put it

> You can get so busy protecting yourself against criticism that you develop such an elaborate methodology that by the time your findings come out, who cares? So, I mean, you get a balance—the validity of the data against its relevance. And that's pretty tough stuff. I mean, that's hard business. [DM111:26]

As no study is ever methodologically perfect, it is important for primary stakeholders to know firsthand what imperfections exist—and to be included in deciding which imperfections they are willing to live with in making the inevitable leaps from limited data to incremental action.

The Dynamics of Measurement and Design Decisions

Research quality and relevance are not set in stone once an evaluation proposal has been accepted. A variety of factors emerge throughout the life of an evaluation that require new decisions about methods.

Case Example of a Flexible Design

In 2006, *Innovation Network* faced the challenge of evaluating an immigration advocacy campaign mounted by the Coalition for Comprehensive Immigration Reform (CCIR). The political environment was especially volatile following May Day marches in cities around the country. The original evaluation design proposed a variety of methods, both quantitative and qualitative, to answer key evaluation questions. The mixed methods design including interviewing key informants, conducting surveys, reviewing documents, and documenting meetings on core strategies. Given the intensity of the immigration campaign, the evaluators needed to be especially sensitive to minimizing the data collection burden for CCIR leadership and coalition members. The design included tracking media coverage, legislation, field activities, and polling studies that did not require primary data collection from campaign staff. As both the campaign and evaluation unfolded, the design changed:

The fast pace of events, and the Coalition's rapid response to them, soon necessitated a greater amount of real-time data collection. The evaluation team began conducting more frequent observation and monitoring of the coalition dynamics that played out in meetings and conference calls. Other challenges inherent to collecting real-time data included massive amounts of data generated through numerous e-mail lists, documents, and field reports.

Two factors were particularly important in convincing the evaluators that they could not rely solely on traditional data collection and needed to redesign the evaluation:

- A legislative policy campaign, like advocacy work generally, involves faster cycles of evolving strategies out of the necessity to react to opportunity windows and respond to external factors.
- The complex interactions among myriad players and stakeholder audiences—who are located along a continuum of connections to and engagement with policymakers—present greater challenges in capturing multiple stories and angles that oftentimes occur simultaneously.

During the most intense periods of the campaign, the evaluators found that "it was unthinkable to conduct interviews with coalition leaders, which resulted in gaps in the data." But, following such intense periods, "there was tangible burnout among everyone in the campaign."

During high-intensity periods, the evaluators continued to monitor numerous meetings, conference calls, and hundreds of e-mails and documents.

The existing methods were not effective in fully capturing the multiple perspectives and many different stories of what happened, especially accounts of interactions with policymakers and their staff. In recognition of the context within which the evaluation was occurring, the evaluation team designed a "Debrief Interview Protocol" specifically for intense periods of advocacy. The intent of this protocol was to engage key players in a focus group shortly after a policy window or intense period occurred, to capture the following information:

- The public mood and political context of the opportunity window
- What happened and how the campaign members responded to events
- What strategies they followed
- Their perspective on the outcome(s) of the period
- How they would change their strategies going forward based on what they learned during that period

(Continued)

(Continued)

By focusing on a specific moment in the campaign and conducting it in a timely manner, this method gathered in-depth and real-time information, while keeping the interaction targeted, practical, and relevant. The idea of the debrief grew out of the need to have a forum that encouraged participation from key groups and individuals engaged in different layers or "spheres of influence" surrounding decision makers.

This emergent design approach proved particularly useful for those involved in the inner workings of the campaign to tell the story of what happened behind the scenes.

The novel aspects of the debrief lie in its systematic application to follow the peaks and valleys of the policy advocacy cycle. It also allows for continued tailoring of the selection of participants and, to some degree, the questions asked based on the nature of the intense period, the parties involved, and the activities that occur.

SOURCE: Stuart (2007:10–11).

Actively involving intended users in making methods decisions about these issues means more than a one-point-in-time acquiescence to a research design.

In every one of the 20 federal health studies we investigated, significant methods revisions and redesigns had to be done after data collection began. While little attention has been devoted in the evaluation literature to the phenomenon of slippage between the original design and methods as actually implemented, the problem is similar to that of program implementation, where original specifications typically differ greatly from what finally gets delivered (see Chapter 9).

In a groundbreaking study, McTavish et al. (1975) studied implementation of 126 research projects funded across seven federal agencies. All 126 projects were rated by independent judges along seven descriptive methodological scales. Both original proposals and final reports were rated; the results showed substantial instability between the two. The researchers concluded,

> Our primary conclusion from the Predictability Study is that the quality of final report methodology is essentially not predictable from proposal or interim report documentation. This appears to be due to a number of factors. First, research is characterized by significant change as it develops over time. Second, unanticipated events force shifts in direction. Third, the character and quality of information available early in a piece of research makes assessment of some features of methodology difficult or impossible. (Pp. 62–63)

Earlier in the report, they had pointed out that

> among the more salient reasons for the low predictability from early to late documentation is the basic change which occurs during the course of most research. It is, after all, a risky pursuit rather than a pre-programmed product. Initial plans usually have to be altered once the realities of data or opportunities and limitations become known. Typically, detailed plans for analysis and reporting are postponed and revised. External events also seem to have taken an expected toll in the studies we examined. . . . Both the context of research and the phenomena being researched are typically subject to great change. (P. 56)

If intended users are involved only at the stage of approving study proposals, they

are likely to be surprised when they see a final report. Even interim reports bear only moderate resemblance to final reports. Thus, making decisions about methods is a continuous process that involves checking out changes with intended users as they are made. While it is impractical to have evaluator-stakeholder discussions about every minor change in methods, utilization-focused evaluators prefer to err in the direction of consultative rather than unilateral decision making, when there is a choice. Stakeholders also carry a responsibility to make sure they remain committed to the evaluation. One internal evaluator interviewed in our federal utilization study, still smarting from critiques of his evaluation as methodologically weak, offered the following advice to decision makers who commission evaluations:

> Very, very often those of us who are doing evaluation studies are criticized for poor methodology, and the people who levy the criticism sometimes are the people who pay for the study. Of course, they do this more often when the study is either late or it doesn't come up with the answers that they were looking for. But I think that a large share of the blame or responsibility belongs to the project monitor, sponsor, or funder for not maintaining enough control, direct hands-on contact with the evaluation as it's going on.
>
> We let contracts out and we keep our hands on these contractors all the time. And when we see them going down a road that we don't think is right, we pull them back and we say, "Hey, you know, we disagree." We don't let them go down the road all the way and then say, "Hey fella, you went down the wrong road." [EV32:15]

I have found this a useful quote to share with primary stakeholders who have expressed reluctance to stay involved with the evaluation as it unfolds. Caveat emptor.

Threats to Data Quality

Evaluators have an obligation to think about, anticipate, and provide guidance about how threats to data quality will affect interpreting and using results. Threats to internal validity, for example, affect any conclusion that a program produced an observed outcome. The observed effect could be due to larger societal changes, as when generally increased societal awareness of the need for exercise and proper nutrition contaminates the effects of specific programs aimed at encouraging exercise and proper nutrition. Maturation is a threat to validity when it is difficult to separate the effects of a program from the effects of growing older; this is a common problem in juvenile delinquency programs, as delinquency has been shown to decline naturally with age. Reactions to gathering data can affect outcomes independent of program effects, as when students perform better on a posttest simply because they are more familiar with the test the second time; or there can be interactions between the pretest and the program when the experience of having taken a pretest increases participants' sensitivity to key aspects of a program. Losing people from a program (experimental mortality) can affect findings since those who drop out, and therefore fail to take a posttest, are likely to be different in important ways from those who stay to the end.

However, it is impossible to anticipate all potential threats to data quality. Even when faced with the reality of particular circumstances and specific evaluation problems, it is impossible to know in advance precisely how a creative design or measurement approach will affect results. For example, having program staff do client interviews in an outcomes evaluation could (1) seriously reduce the validity

and reliability of the data, (2) substantially increase the validity and reliability of the data, or (3) have no measurable effect on data quality. The nature and degree of effect would depend on staff relationships with clients, how staff were assigned to clients for interviewing, the kinds of questions being asked, the training of the staff interviewers, attitudes of clients toward the program, and so on. Program staff might make better or worse interviewers than external evaluation researchers, depending on these and other factors.

An evaluator must grapple with these kinds of data quality questions for all designs. No automatic rules apply. There is no substitute for thoughtful analysis based on the specific circumstances and information needs of a particular evaluation, both initially and as the evaluation unfolds.

Threats to Utility

Whereas traditional evaluation methods texts focus primarily on threats to validity, this chapter has focused primarily on threats to utility. Exhibit 11.5 summarizes common threats to utility. We now have substantial evidence that paying attention to and working to counter these threats to utility will lead to evaluations that are worth using—and are actually used.

EXHIBIT 11.5

Threats to Utility

- Failure to focus on intended use by intended users
- Failure to design the evaluation to fit the context and situation
- Inadequate involvement of primary intended users in making methods decisions
- Focusing on unimportant issues—low relevance
- Inappropriate methods and measures given stakeholders questions and information needs
- Poor stakeholder understanding of the evaluation generally and findings specifically
- Low user belief and trust in the evaluation process and findings
- Low face validity
- Unbalanced data collection and reporting
- Perceptions that the evaluation is unfair or that the evaluator is biased or less than impartial
- Low evaluator credibility
- Political naïveté
- Failure to keep stakeholders adequately informed and involved along the way as design alterations are necessary

Designing Evaluations Worth Using: Reflections on the State of the Art

This chapter has described the challenges evaluators face in working with intended

users to design evaluations worth using. My consulting brings me into contact with hundreds of evaluation colleagues and users. I know from direct observation that many evaluators are meeting these challenges with great skill, dedication, competence, and effectiveness. Much important and creative work is being done by evaluators in all kinds of difficult and demanding situations as they fulfill their commitment to do the most and best they can with the resources available, the short deadlines they face, and the intense political pressures they feel. They share a belief that doing something is better than doing nothing, so long as one is realistic and honest in assessing and presenting the limitations of what is done.

This last caveat is important. I have not attempted to delineate all possible threats to validity, reliability, and utility. This is not a design and measurement text. My purpose has been to stimulate thinking about how attention to intended use for intended users affect all aspects of evaluation practice, including methods decisions.

Pragmatism undergirds the utilitarian emphasis of utilization-focused evaluation. In designing evaluations, it is worth keeping in mind World War II General George S. Patton's Law: *A good plan today is better than a perfect plan tomorrow*.

Then, there is Halcolm's evaluation corollary to Patton's law: *Perfect designs aren't.*

Follow-Up Exercises

1. The chapter opens by asserting that involving primary intended users in making methods decisions is controversial and resisted by many evaluators. What is the controversy? What is the basis for the resistance? Present the essence of the argument against involving nonresearchers in methods decisions. Then, present the essence of the argument in favor of involvement. Finally, present your own philosophy and preference on this issue.

2. Using Rudyard Kipling's poem (below), present the primary design features of an evaluation for an actual program. Describe the program and then describe the evaluation specifying What, Why, When, How, Where, Who.

I keep six honest serving men
They taught me all I knew:
Their names are What and Why and When
And How and Where and Who.

3. Select an evaluation design or measurement issue and to write a script for how you would present and explain the primary options available to nonresearchers who are primary intended users for the evaluation. Include in your explanation the likely consequences for credibility and utility of the results. An example in this chapter is the choice between odd-numbered and even-numbered response options in surveys. Another example would be telephone interviews versus face-to-face interviews. Select your own example and present the options in lay terms.

4. Locate an actual evaluation report for a completed evaluation. Examine the design and methods used in the evaluation. Summarize these design elements on the left-hand side of a page. Next to each design element, on the right side of the page, present two alternatives: (a) an option that would be significantly more expensive and (2) an option that would be significantly less expensive. (You will have to speculate on the level of expense associated with the evaluation's actual design.)

EXHIBIT 11.6

Gathering Background Data: Race, Ethnicity, and Other Demographics

In March 2007, an extensive thread developed on EvalTalk, the American Evaluation Association listserv, concerning the validity and utility of collecting background data on program participants. Below are 20 comments (some edited) to illustrate the diverse perspectives generated from evaluators in different political, cultural, and national contexts. I offer this diversity of views as a way of illustrating why it is important to involve primary intended users in such design and measurement decisions to determine what they consider valid and useful within a particular context for a specific evaluation purpose. How to gather background data is not primarily a technical decision. It is a decision that has significant political, social, cultural, practical, and utility consequences.

Original Questions

What's the best way to ask about a survey respondent's race/ethnicity? Furthermore, what are the correct and most politically sensitive response categories?

As far as evaluation practice is concerned, how often do evaluation consumers use race as a meaningful variable in making decisions? It seems like we always ask race as a standard demographic, and it makes for a great chart or table, but does it always have meaning?

Responses

1. Race is a social construct and the definition of the term has changed over time, and differs from country to country and region to region. For comparative purposes, in 1997 the Office of Budget and Management (OMB) revised the definitions of race for Federal agencies, including the U.S. Census Bureau: (1) American Indian or Alaska Native, (2) Black or African American, (3) Native Hawaiian or Pacific Islander, (4) White, and (5) Some other race. OMB guidelines allow an individual to select more than one race. Additionally, the Feds use two ethnic categories: *Hispanic origin* and *Not of Hispanic origin*. My take is that because we are such a pluralistic society, even these categories may not match an individual's perception of their racial and ethnic identity. As a result, people may increasingly mark the "Some other race" option if offered, making these designations less useful over time.
2. I work in Australia and tend to use a slightly different set of questions (which are based on work by our national bureau of statistics).

They & I have found that in our environment, where we have great diversity in people's countries of origin and languages spoken, asking for ethnicity/race didn't work particularly well. Instead, we ask the following questions:

What country were you born in?
- Australia
- Other (choose from drop-down list)

Does your family come from another country?
- Yes, choose from drop-down list
- No

What is the main language you speak at home?
- English
- Other (choose from drop-down list)

Are you of Aboriginal or Torres Strait Islander descent? (NB. These are our indigenous peoples)
- Yes
- No

This gives us practical information about languages to publish information in and information about cultural heritage. Would this approach work in the States?

3. Given the complexity of race/ethnicity, I have often made this an open-ended sort of question: With what race/ethnicity(ies) do you most strongly identify? The drawback is that the responses don't necessarily align with U.S. Census categories—which can limit quantitative comparisons. The upside is that respondents can tell us as evaluators something about what matters—what has meaning—with regards to racial and ethnic identity.
4. A compromise I have used is to list the categories of the larger data set you intend to use for comparison, if there is one (e.g., federal categories), and a response option labeled "other," and give instructions: Please check all that apply. This allows for people of mixed race to respond with honesty, allows for self-definition of race or ethnicity, and is still more efficient to score than completely open-ended responses.
5. In the U.S., race/ethnicity is often used as a proxy for a host of social, psychological, and even biological variables. We need to focus our attention on measuring those variables directly in such situations, rather than relying on race/ethnicity.
6. I do use race in my evaluations. I use the data as a comparison with census bureau data. One of the issues in both these fields is race disproportionality and as such it's become a focal point in most things that I do. However, I must confess, that the only reason I use ethnicity data is because federal funders request this information. I'm not exactly sure how to use this otherwise.
7. One of the consequences (intended or not?) with the U.S. Government's recent switch to differentiating "Hispanic" ethnicity from racial categories is a blurring of data and obscuring of disparities. For example, I now run across government reports that combine "Hispanic whites" and "non-Hispanic whites" into "white/caucasians." This has the effect of lowering the appearance of racial/ethnic disparities in some reports. For example, while prior reports show a large disparity on some variables between African Americans and non-Hispanic Whites, with Hispanic Whites somewhere in between, the new reports now show a much smaller disparity between African Americans and Whites (Hispanic and non-Hispanic).
8. If your research is focusing on the impact of race and culture on some factor, one should "truly know" the target population(s) and the community of interest. Thus, if you are working in Florida or perhaps New York, it may be very important to allow respondents to share their national origin with you as well as their race, given immigration patterns from the Caribbean. This is particularly true for Hispanics/Latinos but also increasingly for Americans who are of African, Arab, and Asian descent, since they may come from a wide variety of countries, including those of Europe. To illustrate, if you are interacting with Hispanics in the Southwest, you may want to offer a wider variety of choices for respondents:

   Hispanic, Latino, Hispano, Spanish-speaking origin, Latin American,

   Mexican American, Puerto Rican, El Salvadorean, etc.

The same might be wise if the instrument will go to a variety of "Native Americans," who come from different tribes. Responses can also be influenced by the age and perhaps ideology of individuals. Thus, for some groups, often including Hispanics/Latinos, it may be critical to know whether respondents' families have lived in the United States for several generations or they are recent arrivals. The most effective practice, however, is to dialogue with representatives of the target population(s) you are particularly interested in encouraging to respond—and to learn from them what different groups within their community call themselves.

9. What we have here is a failure of communication. It is a clash with Anglo-Saxon/Northwestern European racial categorizations and Hispanic and Semitic ethnic categorizations. In the United States, what chiefly matters is whether you are dark skinned or not, and status and opportunities are accorded on that distinction. In Western Hemispheric Hispanic cultures, what chiefly matters is whether you are an Indian or not. The very concept "race" is defined differently. In the United States, race refers only to biological attributes of a person. In Hispanic culture, "la Raza" refers to which ethnicity one is. So your Raza may be Chicana, but your skin color may be dark or light. Members of this country are struggling with the reconciliation of Mexican American self-identity and traditional American self-identity. I am not contending that that is easy. For my part, I always pick "Other."

(Continued)

(Continued)

10. Here in Florida, we ran into the issue of respondents being offended that they had to select a race after being identified as Hispanic/Latino. That does reraise the question of whether or not Latino is a race as well as ethnicity, but the federal guidelines are pretty explicit:

    —When the survey is being completed by someone other than the respondent, you can use combined (race/ethnicity) categories;

    —When the survey is being completed by respondents themselves, race and ethnicity must be split and the only ethnic options are Hispanic/Latino or Not Hispanic/Latino.

We also ran into an issue where Arab Americans were extremely offended by these categories. It was said to me that Arab is an ethnicity and that they should be able to select a "race" given the various regions from which Arabs may hail. I also began trying to tackle this issue at a local Historically Black College & University. While the majority of the students are black, this particular institution has a strong Caribbean population. The administration pointed out that the services required by Caribbean black students were quite different from African Americans. They felt that, given the definition of ethnicity, they should be able to identify other ethnicities, even when the students are black.

11. I have to agree with those who say that race is an artificial social construct. However, I also have to agree that it is inarguably important in and of itself (aside from the correlations) because white people (primarily) treat others on the basis of their perceived race. Information about race is therefore used by the white-dominated society to treat others differently, and multiple data analyses prove that race is in fact a meaningful contributor to explaining differences in treatment. [On a major long term project] I contributed analyses of racial differences. Our studies showed that in many circumstances there is a separate and significant impact of race over and above the impact of the other variables mentioned. So IMHO race is both an artificial social category with little or no "reality" in biology or genetics AND a crucial piece of information used by individuals and institutions as a basis for classifying and behaving towards individuals.
12. Race will continue to matter in the U.S. until the subtle and overt conception of "American" is no longer "white." So until we address the myth of equality based on flimsy laws that do not account for "pure" and resultant "statistical" racism, race must stay in the picture, and we must work to close the gap between lofty ideas and lived reality by having open and honest discussions which force us all to look at how we maintain an investment in "whiteness."
13. Race comes into play when comparing work force with client population. There are varied opinions on the subject but generally when programs are accredited or funded by an external entity one of the areas that is looked at is the racial composition of staff to the client population. Generally, and this was a big argument a number of years ago in child welfare, the view is that the racial composition of staff should be in close proximity to the racial population of the client population. Some thought this important and believe it can have an effect on the quality of services offered to clients. In the end, there was a belief that staff who have similar backgrounds would be more sensitive to the needs of their client population and thereby provide better services. Thus, this is a reason for including race in an evaluation.
14. In Canada, we don't refer to our French population as another race, or for that matter a minority group. They are Canadians who speak French and possibly English as well. Black refers to a degree of skin pigment and other biological traits that only genetic testing could specify exactly, plus ethnic traits. Whites contain these traits to a degree as well. We wouldn't treat age as categorical—young vs. old. Why should we consider race/ethnicity as one or the other? Native Indian can mean a set of variables—biological, legal, ethnic, linguistic. "Metis" means mixed blood. . . but what percentage makes a person one or the other? There is no answer. Race is a useless variable. I prefer to dispense with the notion of race as antiquated, and I put ethnicity in the same category. Instead, we should measure the dimensions that they are supposed to represent: country of birth, legal status, first language, income, education, et cetera.

15. To be sensitive about how to present the categories on a form is important. We the evaluators have paid scant "professional" attention to recent advancements in human genomics. I got my genetic testing done and found out that I have more close relatives in my deep ancestry in West Asia and Europe than in my native India. In fact, in the current database I am more closely related to a Carter and a Campbell (we might have shared an ancestor as close as just a 100 generations ago) than a Srivastava (an Indian last name) in the same Y-Haplogroup. I faithfully and routinely note myself to be Asian on forms that follow the U.S. Census categories. The Africans have more genetic variations than all other types of humans combined! This is one of the pillars of our recent understanding of human origins in Africa. If you grow up in Asia, you learn that Syrians are Asians as are inhabitants of two-thirds of the Russia's. According to U.S. Census categories, if you are from Russia, you check white. If you are from Iran, you check white. If you are from anywhere in Asia East of Afghanistan, such as neighboring Pakistan, you check Asian. But wait, there's more—if you are from Siberia, which is way East of Afghanistan, you check white—Wow! So race is not only a social construct now, it is so because a particular branch of government told you so.
16. One will notice a significant difference in the race categories used by the National Center for Health Statistics (much finer grained) and those used by the U.S. Census (rather short on fine grains). Why? Because the purposes are different. The point is, one has to consider the purpose, why one is collecting this information. If your evaluation has implications for biologically based drug interventions, you may be better off collecting more genetically detailed categories as we do know genetics play a role in susceptibility to certain diseases and responses to medical interventions. If we are concerned purely about social interventions, we might as well get to the bottom of the socioeconomic construct that we intend to use the race category for.
17. I'm a little startled at the way "race" still appears to be treated as a neutral term by social scientists in government employ in the US. I'm quite sure any social scientist or government employee in Australia would recoil from the idea of using the word in any official document. Not simply that it's politically incorrect, more that it would be regarded as downright insulting—not to mention unscientific. Whatever else may have gone wrong over the last couple of decades, the message seems to have been successfully drummed into a couple of generations here that "There is no such thing as race." You could compare the way "aboriginal" is becoming correct usage to describe the original owners of Canada—and to a lesser extent, of North America generally—at exactly the same time as it seems to have been relegated to ideological unsoundness, in some quarters at least, when applied, as it has been for a couple of centuries, to the original owners of Australia.
18. Given the seemingly increasing complexity in measuring race, do we do more harm than good when we impose rigid categories on our participant subjects?
19. The best background form is one that is customized for the context of the program and the purpose of the evaluation. Off-the-shelf standardized forms (like census bureau categories) may help you in formulating possibilities, but content of any specific evaluation form should flow directly from the evaluation objectives and program's target population.
20. The essential question from a questionnaire construction standpoint is why do you need this information?

5. Using the views and options presented in Exhibit 11.6 at the end of this chapter, identify at least three options for asking program participants about their race or ethnicity, then discuss the likely consequential validity of those options.

6. Explain *demand validity* and discuss the pros and cons of including this concept in an evaluation. Under what evaluation situation would it be appropriate and useful? Under what situation would it possibly reduce the evaluator's credibility. Why?

12

The Paradigms Debate and a Utilization-Focused Synthesis

Lady, I do not make up things. That is lies. Lies is not true. But the truth could be made up if you know how. And that's the truth.

—Lily Tomlin as character "Edith Ann,"
Rolling Stone, October 24, 1974

A former student sent me the following story, which she had received as an e-mail chain letter, a matter of interest only because it suggests widespread distribution.

Once upon a time, not so very long ago, a group of statisticians (hereafter known as quants) and a party of case study aficionados (quals) found themselves together on a train traveling to the same professional meeting. The quals, all of whom had tickets, observed that the quants had only one ticket for their whole group.

"How can you all travel on one ticket?" asked a qual.

"We have our methods," replied a quant.

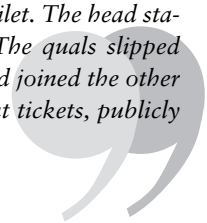
Later, when the conductor came to punch tickets, all the quants slipped quickly behind the door of the toilet. When the conductor knocked on the door, the head quant slipped their one ticket under the door, thoroughly fooling the conductor.

On their return from the conference, the two groups again found themselves on the same train. The qualitative researchers, having learned from the quants, had schemed to share a single ticket. They were chagrined, therefore, to learn that, this time, the statisticians had boarded with no tickets.

"We know how you traveled together with one ticket," revealed a qual, "but how can you possibly get away with no tickets?"

"We have ever more sophisticated methods," replied a quant.

Later, when the conductor approached, all the quals crowded into the toilet. The head statistician followed them and knocked authoritatively on the toilet door. The quals slipped their one and only ticket under the door. The head quant took the ticket and joined the other quants in a different toilet. The quals were subsequently discovered without tickets, publicly humiliated, and tossed off the train at its next stop.



Quants and Quals

Who are *quants?* They're numbers people who, in rabid mode, believe that if you can't measure something, it doesn't exist. They live by Galileo's admonition, "Measure what is measurable, and make measurable what is not so." Their mantra is "What gets measured gets done." And *quals?* They quote management expert W. Edwards Deming: "The most important things cannot be measured." *Quals* find meaning in words and stories, and are ever ready to recite Albert Einstein's observation that "Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted." Relatively speaking, of course.

Quants demand "hard" data: statistics, equations, charts, and formulae. Q*uals,* in contrast, are "softies," enamored with narrative and case studies. *Quants* love experimental designs and believe that the only way to prove that an intervention caused an outcome is with a randomized control trial (RCT). *Quants* are control freaks, say the *quals;* simplistic, even simpleminded, in their naive belief that the world can be reduced to independent and dependent variables. The *qual's* world is complex, dynamic, interdependent, textured, nuanced, unpredictable, and understood through stories, and more stories, and still more stories. *Quals* connect the causal dots through the unfolding patterns that emerge within and across these many stories and case studies. *Quants* aspire to operationalize key predictor variables and generalize across time and space—the holy grail of truth: if *x,* then *y,* and the more of *x,* the more of *y*. *Quals* distrust generalizations and are most comfortable immersed in the details of a specific time and place, understanding a story in the richness of context and the fullness of thick description. For *quals,* patterns they extrapolate from cross-case analyses are possible principles to think about in new situations but are not the generalized, formulaic prescriptions that *quants* admire and aspire to. *Quants* produce *best practices* that assert, "Do this because it's been proven to work in rigorous studies." *Quals* produce themes and suggest, "Think about this and what it might mean in your own context and situation."

Do opposites attract? Indeed, they do. They attract debate, derision, and dialectical differentiation—otherwise known as *the paradigms war*. The story of the *quals* and *quants* offers a window into how the paradigms debate has ebbed and flowed.

This debate about the relative merits of quantitative/experimental methods versus qualitative/case study methods has periodically run out of intellectual steam, but as this edition is being revised, the debate is once again ascendant, this time focused on whether randomized controlled experiments are *the gold standard* for impact evaluations. This chapter will examine the debate and offer a utilization-focused synthesis.

Methodological Debate

The debate has taken different forms over time, including periods of intense rancor as well as times of rapprochement. Thomas D. Cook, one of evaluation's luminaries—the Cook of Shadish, Cook, and Campbell (2001), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference,* the bible of research design—pronounced in his keynote address to the 1995 International Evaluation Conference in Vancouver, "Qualitative researchers have won the qualitative-quantitative debate."

Won in what sense?

Won acceptance. Cook supports use of multiple methods in evaluation and has made it clear that qualitative approaches can be quite valuable for describing what happens in a classroom or program, how the program is implemented, and for deepening our understanding of what outcomes may mean. But to produce strong evidence about causality, he remains convinced of the superiority of experimental designs:

> Since the theoretical warrant for the experimental result is more compelling than the warrant for the non-experimental result, the presumption is that non-experiments are often biased and that, even if they are not, there would be no way to know this in particular instances unless a randomized experiment were also done. . . . The experiment is to be preferred over other potentially bias-free methods because it enjoys greater statistical power and its assumptions are more transparent and better understood when compared to other forms of causal research. (Cook 2006:2, 4)

The validity of experimental methods and quantitative measurement, appropriately used, has never been in doubt. By the 1990s, qualitative methods, appropriately used, had ascended to a level of comfortable respectability, at least as an adjunct to quantitative methods in mixed-methods evaluations. Along the path to valuing mixed methods, evaluation methodologists have engaged in sometimes acrimonious debate, as when Lee Sechrest, American Evaluation Association (AEA) president in 1991, devoted his presidential address to alternatively defending quantitative methods and ridiculing qualitative approaches. He lamented what he perceived as a decline in the training of evaluators, especially in conducting rigorous quantitative studies. He linked this to a more general "decline of numeracy" and increase in "mathematical illiteracy" in the nation. "My opinion," he stated, "is that qualitative evaluation is proving so attractive because it is, superficially, so easy" (Sechrest 1992:4). Partly tongue in cheek, he cited as evidence of qualitative evaluators' mathematical ineptitude a proposal he had reviewed from a qualitative researcher that contained a misplaced decimal point and, as another piece of evidence, an invitation to a meeting of "qualitative research types" that asked for a February 30 reply (p. 5). He concluded,

> If we want to have the maximum likelihood of our results being accepted and used, we will do well to ground them, not in theory and hermeneutics, but in the dependable rigor afforded by our best science and accompanying quantitative analyses. (P. 3)

Beyond the rancor, however, Sechrest joined other eminent researchers in acknowledging a role for qualitative methods, especially in combination with quantitative approaches. He was preceded in this regard by distinguished methodological scholars such as Donald Campbell and Lee J. Cronbach. Ernest House (1977), describing the role of qualitative argument in evaluation, observed that "when two of the leading scholars of measurement and experimental design, Cronbach and Campbell, strongly support qualitative studies, that is strong endorsement indeed" (p. 18). In my own work, I have found increased interest in and acceptance of qualitative methods and, in particular, mixed methods (both quantitative and qualitative in combination).

While a consensus has emerged in the profession that evaluators need to know and use a variety of methods in order to be responsive to the nuances of particular evaluation questions and the idiosyncrasies of specific stakeholder needs, the question of what constitutes *the methodological gold standard* remains hotly contested. There is some contradiction in the assertion that (1) the issue is the appropriateness of methods for a specific evaluation purpose and question, and that where possible, using multiple methods—both quantitative and qualitative—can be valuable, BUT (2) one question is more important than others (the causal attribution question) and one method (RCTs) is superior to all other methods in answering that question. This is what is known colloquially as talking out of both sides of your mouth. Thus, we have a problem. The ideal of evaluators being situationally responsive, methodologically flexible, and sophisticated in using a variety of methods runs headlong into the conflicting ideal that experiments are *the* gold standard and all other methods are, by comparison, inferior, what Scriven (2006b) has called "RCT imperialism" (p. 8). These conflicting ideals play out amidst the realities of limited evaluation resources, political considerations of expediency, and the narrowness of disciplinary training available to most evaluators—training that imbues them with varying degrees of methodological prejudice. Nor is the debate just among evaluation methodologists. Evaluation practitioners are deeply affected, as are users of evaluation—policymakers, program staff, managers, and funders. All can become mired in the debate about whether statistical results from experiments ("hard" data) are more scientific and valid than quasi-experiments and qualitative case studies ("softer" data). Who wants to conduct (or fund) a second-rate evaluation if there is an agreed-on gold standard? What really are the strengths and weaknesses of various methods, including experiments (which, it turns out, also have weaknesses)? What does it mean to match the method to the question?

If evaluators are to involve intended users in methods decisions, as utilization-focused evaluation advocates, evaluators and intended users need to understand the paradigms debate and evaluators need to be able to facilitate choices that are appropriate to a particular evaluation's purpose. This means educating primary stakeholders about the legitimate options available, the potential advantages of multiple methods, and the strengths and weaknesses of various approaches.

The Gold Standard Question

What does it mean for something to be the GOLD STANDARD?

The gold standard is a monetary system in which the standard economic unit of account is a fixed weight of gold. When several nations are using such a fixed unit of account, the rates of exchange among national currencies effectively become fixed to the value of gold. The United States stopped issuing promises to redeem dollars for gold in 1933—part of a policy change for dealing with the Great Depression. As World War II was ending, the international 1944 Bretton Woods system created an obligation for each country to maintain the exchange rate of its currency in terms of gold. The system collapsed in 1971 following the United States' suspension of convertibility from dollars to gold. The system failed, in part, because of its rigidity.

The gold standard question in evaluation is whether one particular method—*randomized control experiments*—should be held up as the best design for conducting impact evaluations and, by being best, should be the standard of excellence toward which evaluators should aspire and against which the quality of evaluation methods are judged. Do randomized control experiments merit the Olympic gold medal for evaluation? That is at the center of the methodological paradigms debate today.

Beyond Methods: The Paradigms Debate

A paradigm is a worldview built on implicit assumptions, accepted definitions, comfortable habits, values defended as truths, and beliefs projected as reality. As such, paradigms are deeply embedded in the socialization of adherents and practitioners: Paradigms tell them what is important, legitimate, and reasonable. Paradigms are also normative, telling the practitioner what to do without the necessity of long existential or epistemological consideration. But it is this aspect of paradigms that constitutes both their strength and their weakness—their strength in that it makes action possible, their weakness in that the very reason for action is hidden in the unquestioned assumptions of the paradigm. In his influential classic, *The Structure of Scientific Revolutions,* Thomas Kuhn (1970) explained how paradigms work.

> Scientists work from models acquired through education and subsequent exposure to the literature, often without quite knowing or needing to know what characteristics have given these models the status of community paradigms. . . . That scientists do not usually ask or debate what makes a particular problem or solution legitimate tempts us to suppose that, at least intuitively, they know the answer. But it may only indicate that neither the question nor the answer is felt to be relevant to their research. Paradigms may be prior to, more binding, and more complete than any set of rules for research that could be unequivocally abstracted from them. (P. 46)

Evaluation was initially dominated by the natural science paradigm of hypothetico-deductive methodology, which values quantitative measures, experimental design, and statistical analysis as the epitome of "good" science. Influenced by philosophical tenets of logical positivism, this model for evaluation came from the tradition of experimentation in agriculture, the archetype of applied research.

> The most common form of agricultural-botany type evaluation is presented as an assessment of the effectiveness of an innovation by examining whether or not it has reached

Paradigm Wars in Other Fields

Particle Physics Experimentalists versus Theorists. "Particle physicists come in two distinct varieties, which, like matter and antimatter, are very much intertwined and, at the same time, agonistic. Experimentalists build machines. Theorists sit around and think" (Kolbert 2007:74). "I am happy to eat Chinese dinners with theorists," the Nobel Prize-winning experimentalist Samuel C. C. Ting once reportedly said. "But to spend your life doing what they tell you to do is a waste of time."

"If I occasionally neglect to cite a theorist, it's not because I've forgotten. It's probably because I hate him," wrote Leon Lederman, another Nobel prize-winning experimentalist.

Arkani-Hamed, a theorist, counters, "It's a general fact about physics that the people you tend to remember are the theorists. At least in the mythology, experiment plays a less central role. And there's a natural reason for that, because the ultimate goal isn't to observe things about nature; the ultimate goal is to understand and explain things about nature. So, for that reason, it's a chicken-and-egg problem. But definitely you want to be a chicken" (Kolbert 2007:74-75).

Financial Analysis Fundamentalists versus Technicians. Technical analysts recommend stocks based entirely on statistical patterns, prediction equations, charts, various benchmarks (e.g., price-earnings ratios, 90-day moving averages, historical support levels, head and should formations, etc.).

Pure technicians don't need to know what the company is or what business it's in; they just need to know the numbers. Fundamentalists, in contrast, visit companies, meet with management, get to know the CEO, examine business strategy, study detailed financial statements and annual reports—and make qualitative judgments about the quality of the company. Fundamentalists call technicians "elves" because they treat their numbers as magic formulae. Technicians accuse fundamental analysts of being soft in the head and numerically impaired.

Religion Literalists versus Interpretativists. Literalists of any religion view their holy scripture as the direct word of their god that must be taken literally: the world was created in 7 days; a virgin birth means a virgin birth; resurrection from the dead means just that; reincarnation means reincarnation; heaven and hell are real places; a ban against eating pork is absolute. Interpretativists view such stories and rules as symbolic and instructive, sources of moral guidance, but not literally true or absolute.

How widespread is this gap? A Gallup poll in 2006 found that nearly half of Americans believe that humans did not evolve over millions of years but were created by God in their present form within the last 10,000 years (Reuters 2007).

Jurisprudence Originalists versus Relativists. U.S. Supreme Court justices who are originalists seek to interpret the Constitution in terms of the original intent of its authors and signers.

Relativists view the Constitution as a living document that must be interpreted in light of changing times, conditions, and understandings.

SOURCE: Reprinted with permission of Suzanne Callahan.

required standards on prespecified criteria. Students—rather like plant crops—are given pretests (the seedlings are weighed or measured) and then submitted to different experiments (treatment conditions). Subsequently, after a period of time, their attainment (growth or yield) is measured to indicate the relative efficiency of the methods (fertilizer) used. Studies of this kind are designed to yield data of one particular type, i.e., "objective" numerical data that permit statistical analyses. (Parlett and Hamilton 1976:142)

By way of contrast, the alternative to the dominant quantitative/experimental paradigm was derived from the tradition of anthropological field studies and undergirded by the philosophical tenets of phenomenology and constructivism. Using in-depth, open-ended interviewing and direct observation, the alternative paradigm relies on qualitative data, naturalistic inquiry, and detailed description derived from close contact with people in the setting under study.

In utilization-focused evaluation, neither of these paradigms is intrinsically better than the other. They represent alternatives from which the utilization-focused evaluator can choose; both contain options for primary stakeholders and information users. *Issues of methodology are issues of strategy, not of morals.* Yet it is not easy to approach the selection of evaluation methods in this adaptive fashion. The paradigmatic biases in each approach are quite fundamental. Great passions have been aroused by advocates on each side. Kuhn (1970) has pointed out that this is the nature of paradigm debates:

> To the extent that two scientific schools disagree about what is a problem and what is a solution, they will inevitably talk through each other when debating the relative merits of their respective paradigms. In the partially circular arguments that regularly result, each paradigm will be shown to satisfy more or less the criteria that it dictates for itself and to fall short of a few of those dictated by its opponent. . . . Since no paradigm ever solves all problems it defines, and since no two paradigms leave all the same problems unanswered, paradigm debates always involve the question: Which problem is it more significant to have solved? (Pp. 109–10)

The contrary positions that sparked the debate in evaluation remain relevant because much social science training is still quite narrow. Evaluators and those who commission or use evaluation will naturally be most comfortable with those methods in which they have been trained and to which they have most often been exposed. A particular way of viewing the world, based on disciplinary training and specialization, becomes so second-nature that it takes on the characteristics of a paradigm. *When all you have is a hammer, everything looks like a nail.* When you are taught that experiments are the gold standard, every evaluation will look like it is an opportunity to conduct an experiment. When all you know is survey research, every evaluation will scream the need for a survey. When all you know is case studies, every evaluation becomes one.

The quantitative-qualitative paradigms debate has been a prominent and persistent topic in evaluation and has generated a substantial literature, only a sample of which is referenced here (Julnes and Rog 2007; Cook 2006; Davidson 2006a; Mark and Henry 2006; Scriven 2006a; Donaldson and Christie 2005; Greene and Henry 2005; Tashakkori and Teddlie 2003; Schwandt 2002; Denzin and Lincoln 2000; Patton 2000, 1978, 1975a; Donmoyer 1996; Cook 1995; Denzin and Lincoln 1994; Guba and Lincoln 1994, 1989, 1981; Eisner 1991; House 1991; Rizo 1991; Cochran-Smith and Lytle 1990; Guba 1990; Owen and Rogers 1999:86–104; Howe 1988; Lincoln and Guba 1985; Cronbach 1982, 1975; Heilman 1980; Reichardt and Cook 1979; Rist 1977; Guttentag and Struening 1975a). Paradigm discussions and debates have also been a regular feature at meetings of professional evaluators worldwide. So let's take a closer look at the two primary paradigm perspectives.

The Quantitative/ Experimental Paradigm

Evidence of the early dominance of the quantitative/experimental (hypotheti-co-deductive) paradigm as the method of

choice in evaluation research can be found in the metaevaluation work of Bernstein and Freeman (1975). The purpose of their study was to assess the quality of evaluative research at the time. What is of interest to us here is the way Bernstein and Freeman defined quality. Exhibit 12.1 shows how they coded their major indicators of quality; a higher number represents higher-quality research. The highest quality rating was reserved for completely quantitative data obtained through an experimental design and analyzed with sophisticated statistical techniques. Bernstein and Freeman did not concern themselves with whether the evaluation findings were important or used, or even whether the methods and measures were appropriate to the problem under study. They judged the quality of evaluation research entirely by its conformance with the dominant quantitative/experimental paradigm. That was the unquestioned gold standard. Such rankings of methods continue today (Schwandt 2007b:119; Petticrew and Roberts 2003).

EXHIBIT 12.1

Experimental Gold Standard Paradigm: Operational Definition of Evaluation Quality

Dimension of Evaluation Quality	Coding Scheme (Higher Number = Higher Quality)
Sampling	1 = Systematic random 0 = Nonrandom, cluster, or nonsystematic
Data analysis	2 = Quantitative 1 = Qualitative and quantitative 0 = Qualitative
Statistical procedures	4 = Multivariate 3 = Descriptive 2 = Ratings from qualitative data 1 = Narrative data only 0 = No systematic material
Impact procedures design	3 = Experimental or quasi-experimental randomization and control groups 2 = Experimental or quasi-experimental without both randomization and control groups 1 = Longitudinal or cross-sectional without control or comparison groups 0 = Descriptive, narrative

SOURCE: Bernstein and Freeman (1975).

Documenting the consensus that existed for how they defined evaluation quality, Bernstein and Freeman cited major evaluation texts of the time (Reicken and Boruch 1974; Rossi and Williams 1972; Caro 1971; Suchman 1967). Representative of the dominant perspective was that of Wholey et al. (1970), "Federal money generally should not be spent on evaluation of individual local projects unless they have been developed as field experiments, with equivalent treatment and control groups" (p. 93). In their widely used methodological primer, Campbell and Stanley (1963) called this paradigm "the only available route to cumulative progress" (p. 3). It was this belief in and commitment to the natural science model on the part of the most prominent academic researchers that made experimental designs and statistical measures dominant. As Kuhn (1970) has explained, "A paradigm governs, in the first instance, not a subject matter but rather a group of practitioners" (p. 80). Those most committed to the dominant paradigm were found in universities, where they employed the scientific method in their own evaluation research and socialized students into the dominant paradigm.

In our mid-1970s study of how federal health evaluations were used, every respondent answered methodological questions with reference to the dominant paradigm. If a particular evaluation being reviewed had departed from what were implicitly understood to be the ideals of "good science," long explanations about practical constraints were offered, usually defensively, under the assumption that since we were from a university, we would be critical of such departures. Studies were described as hard or soft along a continuum in which harder was clearly better and didn't even need explicit definition.

Advocacy of the quantitative/experimental paradigm as the gold standard continues today supported by many examples of the important results yielded by experiments (Boruch 2007). François Bourguignon, Chief Economist of the World Bank was using randomized control trials as the gold standard when asserted that only 2 percent of World Bank programs had been "properly evaluated" (Dugger 2004:A4), ignoring the great variety of World Bank programs and the vast amount of other kinds of excellent evaluation done by scores of World Bank evaluators and contractors (IEG 2006, 2007). The influential and prestigious Poverty Action Lab at the Massachusetts Institute of Technology has been a strong advocate of randomized control trials as evaluation's methodological gold standard (www.povertyactionlab.com). A widely circulated and influential report from the Center for Global Development entitled *When Will We Ever Learn?* advocates experimental designs as the best way to evaluate impact of international development aid (Evaluation Gap Working Group 2006).

The *What Works Clearinghouse* (WWC) was established in 2002 by the U.S. Department of Education's Institute of Education Sciences to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education—and quickly adopted randomized controlled experimentation as its gold standard (Lawrenz and Huffman 2006). WWC (2006) has established standards of evidence for reviewing studies:

> In order for a study to be rated as meeting evidence standards (with or without reservations), it must employ one of the following types of research designs: a randomized controlled trial or a quasi-experiment (including quasi-experiments with regression discontinuity designs, and single-case designs).

The only evaluations that fully meet the evidence standard, then, are randomized

controlled trials (RCTs) or regression discontinuity designs that do not have problems with randomization, attrition, or disruption. What does this mean in practice? Let's use the *What Works Clearinghouse* review of the Middle School *Connected Mathematics Project* as an example. The methods of 22 studies of this curriculum were reviewed. Three met the methodological standard. Those three rigorous evaluations led to the not-very-helpful conclusion that "the curriculum had mixed effects on math achievement." The 19 excluded studies, many published in peer-reviewed journals, represented a variety of other methods but were not reviewed for patterns, learnings, trends, hypotheses, insights, or tendencies that might deepen understanding of the mixed findings. Because those studies did not meet the gold standard, they were dismissed out of hand. That is the epitome of applying paradigm blinders.

In reacting to the Institute of Education's advocacy of experimentation as the gold standard, distinguished evaluator leader Eleanor Chelimsky (2007) welcomed the commitment to more rigorous evaluations that added an illuminative analogy:

> It is as if the Department of Defense were to choose a weapon system without regard for the kind of war being fought; the character, history, and technological advancement of the enemy; or the strategic and tactical givens of the military campaign. (P. 14)

When the U.S. Department of Education's Institute of Education Sciences first published their criteria, the AEA took the unprecedented step of submitting a formal statement of concern opposing such a narrow and rigid view of how to engage in evaluation. The elected leadership of AEA adopted and submitted the position reproduced in Exhibit 12.2. That statement essentially opposes crowning RCTs as the methodological gold standard.

Not all AEA members supported the statement. Intense debate ensued, evoking strong emotions and reactions. For example, distinguished sociologist Peter Rossi, one of the founders of the field of evaluation research, one of the profession's most important textbook authors, and an original member of both the Evaluation Research Society and the AEA, terminated his membership in AEA saying, "Why be a member of the flat earth society?" (Quoted by Lipsey 2007b:202). Mark Lipsey, coauthor with Rossi of the widely used textbook *Evaluation: A Systematic Approach Evaluation* (Rossi, Lipsey, and Freeman 2003) also dropped out of AEA and has refused to attend subsequent national AEA conferences. Lipsey, a strong advocate of RCTs as the best way to conduct impact evaluations, has been active in debating the issues (Lipsey 2007b; Donaldson and Christie 2005), including twice with me at The Evaluators Institute. For the record, I support the AEA position against crowning any single method as the gold standard, though I was not involved in drafting the statement. It is important to understand that this position is not hostile to experiments and supports their use when appropriate and feasible; it is hostile to treating any method as inherently superior to others without regard to context, appropriateness, and feasibility.

This point is well-illustrated by the commentary of one of the evaluation profession's luminaries, Lois-ellin Datta, about the problem of mandating experimental designs. She has provided as powerful an example as I have seen illustrating the importance of taking context into account in deciding whether to conduct an experiment. Yielding to political pressure from advocates of RCTs, the U.S. Congress

EXHIBIT 12. 2

American Evaluation Association Position on "Scientifically Based Evaluation Methods"

Response to the U.S. Department of Education's Institute of Education Sciences proposal, subsequently adopted, to make randomized control experiments the gold standard for evidence in evaluating educational curricula and programs

The American Evaluation Association applauds the effort to promote high quality in the U.S. Secretary of Education's proposed priority for evaluating educational programs using scientifically based methods. We, too, have worked to encourage competent practice through our Guiding Principles for Evaluators, Standards for Program Evaluation, professional training, and annual conferences. However, we believe the proposed priority manifests fundamental misunderstandings about (1) the types of studies capable of determining causality, (2) the methods capable of achieving scientific rigor, and (3) the types of studies that support policy and program decisions. We would like to help avoid the political, ethical, and financial disaster that could well attend implementation of the proposed priority.

1. *Studies Capable of Determining Causality*. Randomized control group trials (RCTs) are not the only studies capable of generating understandings of causality. In medicine, causality has been conclusively shown in some instances without RCTs, for example, in linking smoking to lung cancer and infested rats to bubonic plague. The secretary's proposal would elevate experimental over quasi-experimental, observational, single-subject, and other designs, which are sometimes more feasible and equally valid.

RCTs are not always best for determining causality and can be misleading. RCTs examine a limited number of isolated factors that are neither limited nor isolated in natural settings. The complex nature of causality and the multitude of actual influences on outcomes render RCTs less capable of discovering causality than designs sensitive to local culture and conditions and open to unanticipated causal factors.

RCTs should sometimes be ruled out for reasons of ethics. For example, assigning experimental subjects to educationally inferior or medically unproven treatments, or denying control group subjects access to important instructional opportunities or critical medical intervention, is not ethically acceptable even when RCT results might be enlightening. Such studies would not be approved by Institutional Review Boards overseeing the protection of human subjects in accordance with federal statute.

In some cases, data sources are insufficient for RCTs. Pilot, experimental, and exploratory education, health, and social programs are often small enough in scale to preclude use of RCTs as an evaluation methodology, however important it may be to examine causality prior to wider implementation.

2. *Methods Capable of Demonstrating Scientific Rigor*. For at least a decade, evaluators publicly debated whether newer inquiry methods were sufficiently rigorous. This issue was settled long ago. Actual practice and many published examples demonstrate that alternative and mixed methods are rigorous and scientific. To discourage a repertoire of methods would force evaluators backward. We strongly disagree that the methodological "benefits of the proposed priority justify the costs."
3. *Studies Capable of Supporting Appropriate Policy and Program Decisions*. We also strongly disagree that "this regulatory action does not unduly interfere with state, local, and tribal governments in the exercise of their governmental functions." As provision and support of programs are governmental functions so, too, is determining program effectiveness. Sound policy decisions benefit from data illustrating not only causality but also conditionality. Fettering evaluators with unnecessary and unreasonable constraints would deny information needed by policymakers.

While we agree with the intent of ensuring that federally sponsored programs be "evaluated using scientifically based research . . . to determine the effectiveness of a project intervention," we do not agree that "evaluation methods using an experimental design are best for determining project effectiveness." We believe that the constraints in the proposed priority would deny use of other needed, proven, and scientifically credible evaluation methods, resulting in fruitless expenditures on some large contracts while leaving other public programs unevaluated entirely.

SOURCE: American Evaluation Association (AEA) (2003). http://www.eval.org/doestatement.htm

mandated a randomized experimental evaluation to test the effectiveness of the Head Start program. When Head Start began in 1965, early childhood education for low-income families was rare. By the year 2000, the widespread availability of preschool programs made getting a genuine control group impossible. Still an RCT was mandated. Datta (2007b) commented,

> To my mind, this mandated randomized test is a horrific example of the inappropriate use of what can be, in appropriate circumstances, an excellent design for estimating the value-added of a program and helping establish attribution. The randomized experimental design has no stronger proponent than me when circumstances seem appropriate. A primary reason that the design is inappropriate in the Head Start circumstance is that the control condition for the test is anything but that . . . For such situations, one perhaps thinks more of evaluation designs derived from systems and complexity theories (Pp. 49–50).

The problem of inappropriately mandated experimental designs is by no means limited to the United States. The gold standard debate has global significance. In December 2007, the European Evaluation Society (EES) adopted a statement on "the importance of a methodologically diverse approach to impact evaluation—specifically with respect to development aid and development interventions." As context, the EES noted that

> this statement was prepared in response to strong pressure from some interests advocating for "scientific" and "rigorous" impact of development aid, where this is defined as primarily involving RCTs. This debate has the potential to influence the future direction of evaluation—not only with respect to development but potentially in other areas as well.
>
> EES however deplores one perspective currently being strongly advocated: that the best or only rigorous and scientific way of doing so is through randomised controlled trials (RCTs). In contrast, the EES supports multi-method approaches to IE [impact evaluation] and does not consider any single method such as RCTs as first choice or as the "gold standard". (EES 2007:1)

In 2007, *Network of Networks on Impact Evaluation* (NONIE) was established by international evaluation offices representing more than 100 United Nations, World Bank, and other development organizations, plus representatives from developing countries and various regional and global organizations. That group drafted a document providing guidance for conducting impact evaluations in developing countries. As this book was going to press, NONIE's draft statement had not yet been officially adopted and published, but the near-final draft being circulated for comment emphasized the importance of methodological diversity and appropriateness in support of rigor, and warned against designating any single design as a gold standard. The literature cited in support of this position and the EES statement includes a number of prominent evaluation theorists and methodologists (Scriven 2008; Bamberger and White 2007; Carden 2007; Chatterji 2007; Julnes and George 2007; Picciotto 2007; Pawson 2002a, 2000b; Weiss 2002).

The gold standard debate revolves around "diverse visions for evaluation in the new millennium" (Donaldson and Scriven 2003). In the pages that follow, I will unpack the issues in the debate as I see them, trying to do justice to the competing perspectives while acknowledging that as a utilization-focused evaluator I advocate methodological eclecticism and adapting evaluation methods to the nature of the

evaluation question and the information needs of primary intended users. *Methodological appropriateness is the utilization-focused gold standard.*

In a nutshell, the problem from a utilization-focused perspective is that the very dominance of the quantitative/experimental paradigm has cut off serious consideration of alternative methods and channels millions of dollars of evaluation funds into support for a method that not only has strengths but also has significant weaknesses. The gold standard accolade means that funders and evaluators begin by asking "How can we do an experimental design" rather than asking "Given the evaluation situation and the information needed, what is the appropriate evaluation design?" The prestige of the method determines the evaluation question and design rather than considerations of utility, feasibility, propriety, and accuracy. Under the gold standard label, high-quality impact evaluation is *defined* as testing hypotheses, formulated deductively, through random assignment of program participants to treatment and control groups, and measuring outcomes quantitatively. No other options are worthy of serious consideration—*by definition.*

Yet alternatives exist, as the AEA and EES statements posit. There are ways other than experiments of assessing program processes, outcomes, *and* impacts. In the last quarter century, these alternatives have been used by evaluators and practitioners who found that the dominant paradigm failed to answer—or even ask—their questions. The importance of having an alternative is captured powerfully by the distinguished adult educator Malcolm Knowles (1989) who, in his autobiography, *The Making of an Adult Educator,* listed discovery of an alternative way of evaluating adult learning as one of the eight most important episodes of his life, right there alongside his marriage. Let's find out what he found so illuminating.

The Qualitative/Naturalistic Paradigm

The alternative qualitative/naturalistic methods paradigm was derived most directly from anthropological field methods and more generally from qualitative sociology, phenomenology, and constructionism. It was undergirded by the doctrine of *Verstehen* (understanding):

> Advocates of some version of the *Verstehen* doctrine will claim that human beings can be understood in a manner that other objects of study cannot. Humans have purposes and emotions, they make plans, construct cultures, and hold certain values, and their behavior is influenced by such values, plans, and purposes. In short, a human being lives in a world which has "meaning" to him, and, because his behavior has meaning, human actions are intelligible in ways that the behavior of nonhuman objects is not. (Strike 1972:28)

In essence, the *Verstehen* doctrine asserts that applied social sciences need methods different from those used in agriculture and pharmacology because human beings are different from plants and medicines. The alternative paradigm emphasizes attention to the meaning of human behavior, the context of social interaction, and the connections between subjective states and behavior. The tradition of *Verstehen* places emphasis on the human capacity to know and understand others through empathic introspection and reflection based on detailed description gathered through direct observation, in-depth, open-ended interviewing, and case studies. Evaluation came to have advocates for and users of alternative methods. Robert Stake's (1975) responsive approach was one such early alternative.

> Responsive evaluation is an alternative, an old alternative, based on what people do naturally to evaluate things; they observe and react. The approach is not new. But this alternative has been avoided in district, state, and federal planning documents and regulations because it is subjective and poorly suited to formal contracts. It is also capable of raising embarrassing questions. (P. 14)

Stake recommended responsive evaluation because "it is an approach that trades off some measurement precision in order to increase the usefulness of the findings to persons in and around the program" (p. 14). Stake influenced a new generation of evaluators to think about the connection between methods and use, and his books on *The Art of Case Research* (1995) and *Multiple Case Study Analysis* (2005) have extended that influence.

I became engaged in the paradigms debate when, after being thoroughly indoctrinated into the dominant paradigm in graduate school as a quantitative sociologist, I became involved in evaluating an open education program whose practitioners objected to the narrow and standardized outcomes measured by standardized tests. Because they advocated an educational approach that they considered individualized, personal, humanistic, and nurturing, they wanted evaluation methods with those same characteristics. In attempting to be responsive to my intended users (open educators) and do an evaluation that was credible and useful to them, I discovered qualitative methods. That led me to write a monograph comparing alternative paradigms (Patton 1975a), reactions to which embroiled me directly and personally in the passions and flames of the great paradigms debate. At the time it was exhilarating. Looking back from today's vantage point of methodological eclecticism, the barbs traded by opposing camps would appear silly but for the fact that, in circles not yet touched by the light that eventually emerged from the debate, friction and its attendant heat still burn evaluators who encounter true believers in the old orthodoxies. It is to prepare for such encounters, and be able to rise gently above the acrimony they can inspire, that students of evaluation need to understand the dimensions and passions of the debate.

Dimensions of the Competing Paradigms

By the end of the 1970s, the evaluation profession had before it the broad outlines of two competing research paradigms. Exhibit 12.3 displays the contrasting emphases of the two methodological paradigms. Beyond differences in basic philosophical assumptions about the nature of reality (ontological differences), in its details the paradigms debate included a number of contrasting dimensions, like the relative merits of being close to versus distant from program participants during an evaluation. While reviewing these dimensions will illuminate the nature of the paradigms debate, they also can be thought of as options that might be offered to intended users during methods deliberations and negotiations. We'll begin with the debate about the relative merits of numbers versus narrative—and the mixed-methods approach of valuing both.

Quantitative and Qualitative Data: Different Perspectives on the World

> *In God we trust. All others must have data.*
>
> —W. Edwards Deming

Both *quals* and *quants* agree with Deming. What they disagree about is what constitutes good data.

EXHIBIT 12. 3

Primary Dimensions of the Contrasting Methodological Paradigms

Qualitative/Naturalistic Paradigm	*Quantitative/Experimental Paradigm*
Qualitative data (narratives, description, quotations)	Quantitative data (numbers, statistics)
Naturalistic inquiry (openness)	Experimental designs (control)
In-depth case studies	Treatment and control groups
Inductive analysis	Deductive hypothesis testing
Subjective perspective valued	Objectivity
Close and direct observation of the program	Distant from and independent of the program
Holistic contextual portrayal	Independent and dependent variables
Systems perspective focused on interdependencies	Linear, sequential modeling
Dynamic, continuous view of change	Pre- and postmeasurement of change
Purposeful sampling of relevant cases	Probabilistic, random sampling
Focus on uniqueness and diversity	Standardized, uniform procedures
Emergent, flexible designs	Fixed, controlled design protocols
Thematic content analysis	Statistical analysis
Value uniqueness, particularity	Replication
Extrapolations (lessons and principles)	Generalizations (empirically based external validity

Quantitative measures strive for precision by focusing on things that can be counted. Quantitative data come from questionnaires, tests, standardized observation instruments, information systems, official indicators, and program records. Gathering numerical data requires conceptualizing categories that can be treated as ordinal or interval data and subjected to statistical analysis. The experiences of people in programs and the important variables that describe program outcomes are fit into these standardized categories to which numerical values are attached. The following opinion item is a common example:

How would you rate from the quality of course instruction?			
1. Excellent	2. Good	3. Fair	4. Poor

In contrast, the evaluator using a qualitative approach seeks to capture what a program experience means to participants *in their own words,* through interviews or open-ended questionnaire items, and in day-to-day program settings, through observation. An open-ended course evaluation question would ask

In your own words, how would you describe the quality of the instruction in this course?

Exhibit 12.4 contrasts other examples of quantitative and qualitative questions.

Qualitative data consist of words and narratives: quotations from open-ended questionnaires; detailed descriptions of situations, events, people, interactions, and observed behaviors; interview responses from people about their experiences, attitudes, beliefs, and thoughts; and excerpts or entire passages from documents, correspondence, records, and case histories. The data are collected as open-ended narrative without predetermined, standardized categories such as the response choices that make up typical questionnaires or tests. The evaluation findings are presented as case studies and analysis of patterns across cases (Patton 2002a; Yin 2002).

EXHIBIT 12. 4

Quantitative and Qualitative Questions: Examples from Evaluation Questionnaires

Standardized, Quantitative Items	*Qualitative, Open-Ended Items*
A. The program's goals were clearly communicated to us? 1. strongly agree 2. agree 3. disagree 4. strongly disagree	A. From your perspective, and in your own words, what are the primary goals of this program?
B. How relevant was this training to your job? 1. very relevant 2. somewhat relevant 3. a little relevant 4. not at all relevant	B. How, if at all, does this training relate to your job? *Please be as specific as possible.*
C. How much did you learn from this program? *I learned* 1. a great deal 2. a fair amount 3. a little bit 4. nothing at all	C. What are the most important things you learned from your participation in this program?

Numbers are parsimonious and precise; words provide individualized meanings and nuance. Each way of turning the complexities of the world into data has strengths and weaknesses. Qualitative data capture personal meaning and portray the diversity of ways people express themselves; quantitative data facilitate comparisons because all program participants respond to the same questions on standardized scales within predetermined response categories. Standardized tests and surveys measure the reactions of many respondents in a way that statistical aggregation and analysis are relatively straightforward, following established rules and procedures. In contrast, qualitative methods typically produce a wealth of detailed data about a much smaller number of people and cases; finding patterns and themes in the diverse narratives can be painstaking, time-consuming, and uncertain. But qualitative data in program evaluation is aimed at letting people in programs express their reactions in their own terms rather than impose on them a preconceived set of limited response categories.

So what is there to debate about quantitative versus qualitative when each can contribute in important ways to our understanding of program? And why not just use both approaches, what is called a *mixed-methods design?* Mixed methods are often used, but one kind of data is often valued over the other. The debate about the relative merits of quantitative versus qualitative data stems from underlying assumptions and deeply held values. "If you can't measure it, if you can't quantify it, it doesn't exist," is a refrain many program staff have heard from evaluators insisting on "clear, specific, and measurable goals" (see Chapter 7 on the goals clarification game). Statistics, because they are concrete and precise, seem more factual—and "getting the facts right" is at the heart of good evaluation (Berk 2007). "What gets measured gets done," the mantra of management by objectives and performance measurement, communicates that only what can be quantified is important. Statistical presentations tend to have more credibility, to seem more like "science," whereas qualitative narratives tend to be associated with "mere" journalism. A certain assertiveness, even machismo, often accompanies the demand that outcomes be quantified: hard data connote virility; soft data are flaccid. (Sexual innuendo works in science no less than in advertising, or so it would seem.) But qualitative advocates have their own favorite quotations, among them the famous assertion of nineteenth-century British Prime Minister Benjamin Disraeli "There are three kinds of lies: lies, damned lies, and statistics." Disraeli also observed, "As a general rule the most successful man in life is the man who has the best information." The quantitative-qualitative debate is about what constitutes the "best information."

Kuhn (1970), a philosopher and historian of science, observed that the values scientists hold "most deeply" concern predictions: "quantitative predictions are preferable to qualitative ones" (pp. 184–85). It's a short distance from a preference for quantitative data to the virtual exclusion of other types of data. Bernstein and Freeman (1975) even ranked evaluations that gathered both quantitative and qualitative data as lower in methodological quality than those that gathered only quantitative data (see Exhibit 12.1). The *What Works Clearinghouse* only uses quantitative findings and ignores qualitative data, even in mixed-methods studies. These are examples of what distinguished sociologist C. Wright Mills (1961) classically labeled "abstracted empiricism" (p. 50).

> If the problems upon which one is at work are readily amenable to statistical

> procedures, one should always try them first. . . . No one, however, need accept such procedures, when generalized, as the only procedures available. Certainly no one need accept this model as a total canon. It is not the only empirical manner.
>
> It is a choice made according to the requirements of our problems, not a "necessity" that follows form an epistemological dogma. (Pp. 73–74)

Valuing quantitative measures to the exclusion of other data limits not only what one can find out but also what one is even willing to ask. It is appropriate and easy to count the words a child spells correctly, but what about that same child's ability to use those words in a meaningful way? It is appropriate to measure a student's reading level, but what does reading *mean* to that student? Different kinds of problems require different types of data. If we only want to know the frequency of interactions between children of different races in desegregated schools, then statistics are appropriate. However, if we want to understand the *meanings of interracial interactions,* open-ended, in-depth interviewing will be more appropriate.

One evaluator in our federal utilization study told of struggling with this issue. He was evaluating community mental health programs and reported that statistical measures frequently failed to capture real differences among programs. For example, he found a case in which community mental health staff cooperated closely with the state hospital. On one occasion, he observed a therapist from the community mental health center accompany a seriously disturbed client on the "traumatic, fearful, anxiety-ridden trip to the state hospital." The therapist had been working with the client on an outpatient basis. After commitment to the state facility, the therapist continued to see the client weekly and assisted that person in planning toward and getting out of the state institution and back into the larger community as soon as possible. The evaluator found it very difficult to measure this aspect of the program quantitatively.

> This actually becomes a qualitative aspect of how they were carrying out the mental health program, but there's a problem of measuring the impact of that qualitative change from when the sheriff used to transport the patients from that county in a locked car with a stranger in charge and the paraphernalia of the sheriff's personality and office. The qualitative difference is obvious in the possible effect on a disturbed patient, but the problem of measurement is very, very difficult. So what we get here in the report is a portrayal of some of the qualitative differences and a very limited capacity of the field to measure those qualitative differences. We could describe some of them better than we could measure them. [EV5: 3]

A more extended example will help illustrate the importance of seeking congruence between the phenomenon studied and the data gathered for an evaluation. In a seminal study, Edna Shapiro (1973) found no achievement test differences between (1) children in an enriched Follow Through (FT) program modeled along the lines of open education and (2) children in comparison schools not involved in FT or other enrichment programs. When the children's test scores were compared, no differences of any consequence were found. However, when she observed children in their classrooms, she could see striking differences between the FT and comparison classes. First, the environments were observably different (implementation evaluation). She characterized the FT classrooms as "lively, vibrant, with a diversity of curricular projects and children's

products, and an atmosphere of friendly, cooperative endeavor." In contrast, she described the non-FT classrooms as "relatively uneventful, with a narrow range of curriculum, uniform activity, a great deal of seat work, and less equipment; teachers as well as children were quieter and more concerned with maintaining or submitting to discipline" (Shapiro 1973:529).

Her observations also revealed that the children performed differently in the two environments on important dimensions that standardized achievement tests failed to detect. Shapiro concluded that the narrow nature of the questions asked on standardized tests predetermined nonsignificant statistical results.

> "I assumed," she reflected, "that the internalized effects of different kinds of school experience could be observed and inferred only from responses in test situations, and that the observation of teaching and learning in the classroom should be considered auxiliary information, useful chiefly to document the differences in the children's group learning experiences." (Shapiro 1973:532)

But then she thought about how tests were administered. To assure consistency, each child was removed from the classroom and given the same exact instructions so that differences in scores would show what has been learned that survived outside the familiarity of the classroom. But she came to worry that this imposed an artificiality in the evaluation that actually disguised significant differences in what children had learned and could do.

She observed such marked disparities between children's classroom responses and test responses that she reevaluated the role of classroom data, individual test situation data, and the relation between them. If we minimize the importance of the child's behavior in the classroom, she asked, do we not have to apply the same logic to the child's responses in the test situation, which is also influenced by situational variables? The quantitative test scores provided one, but only one, form of evaluation of what children had learned. Qualitatively observing them answer questions and do school work in their classrooms provided very different findings about what children knew. These intriguing and important differences in learning outcomes under different conditions led her to believe that *both kinds of data should be valued and used.* Fast forward a quarter of a century to the fixation on standardized test scores in federal *No Child Left Behind* accountability standards and it is clear that Shapiro's insights have not been incorporated in any significant way in educational evaluation. Test scores trump all other kinds of data.

It is worth remembering in this regard that one of the functions of scientific paradigms is to provide criteria for choosing problems that can be assumed to have solutions: "Changes in the standards governing permissible problems, concepts, and explanations can transform a science" (Kuhn 1970:106). The problem in education has been defined as raising test scores and reducing disparities in scores. A particular way of measuring learning has come to define the very nature of the problem. Asking a broader question leads to different kinds of evaluation data: What are ways in which children can demonstrate what they have learned? The answer can include test scores, to be sure, but can also include examining the work children do in the classroom, their performance on teacher-made tests, portfolios of students' work, examples of their homework, and their performance on integrated projects where they use what they know. If the educational problem and corresponding

evaluation question is defined at the outset as how to increase test scores, then the curriculum becomes based on that intended outcome (teach to the tests because what gets measured gets done) and the definition of learning becomes entirely quantitative and standardized. Those who value qualitative evaluation data tend to emphasize individualized learning, diverse ways of capturing what students know, and placing what children can do in the context of what opportunities they have to demonstrate what they know. Thus, the methods debate in educational evaluations is integrally interconnected to competing educational paradigms about how children learn and what is important to learn.

Mixed-Methods Designs: Combining Qualitative and Quantitative Data

From a utilization-focused evaluation perspective, both qualitative and quantitative data can contribute to all aspects of evaluative inquiries. In its simplest form in college exams, mixed methods means asking both multiple choice questions and open-ended essay questions. In evaluations it can mean collecting data with both fixed-choice surveys and using statistical indicators of outcomes as well as conducting open-ended interviews and case studies. Evaluators should be able to use a variety of tools if they are to be sophisticated and flexible in matching research methods to the nuances of particular evaluation questions and the idiosyncrasies of specific decision-maker needs. In *Qualitative Research and Evaluation Methods* (Patton 2002a), I have elaborated the conditions under which qualitative methods are particularly appropriate in evaluation research, for example, when program outcomes are highly individualized so case studies are essential to capture variations in outcomes. Sometimes quantitative methods alone are most appropriate as in counting how many graduates of an employment program get and keep jobs. But in many cases, *both qualitative and quantitative methods should be used together* and there are no logical reasons why both kinds of data cannot be used together (Patton 1982a). Mixed methods have been of interest in evaluation for some time, including *Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms* (Greene and Caracelli 1997). As interest and practice have grown, mixed-methods designs are receiving more attention than ever, including a new *Journal of Mixed Methods Research* hailing the "New Era of Mixed Methods" (Tashakkori and Creswell 2007), Jennifer Greene's important book on *Mixing Methods in Social Inquiry* (2007) with its emphasis on *meaningful engagement with difference,* and publication of the *Handbook of Mixed Methods in Social and Behavioral Research* (Tashakkori and Teddlie 2003). A special issue of the journal *Research in the Schools* was devoted to "New Directions in Mixed Methods Research" (Johnson 2006).

All mixed-methods designs combine qualitative and quantitative data in some way. Some mixed designs are primarily quantitative, with qualitative data as supplementary; others are primarily qualitative with quantitative data as ancillary, as when using ethnography in conjunction with statistical data in a mixed-method strategy (Caracelli 2006). "Pure mixed methods designs" give "equal status" to quantitative and qualitative data (Johnson, Onwuegbuzie, and Turner 2007). In whatever combinations multiple methods are used, the contributions of each kind of data should be fairly assessed. In many cases, this means that evaluators working

Mixed and Emergent Methods:
Adapting Both Program and Evaluation to Changing Conditions

In the fall of 2005, a program I run called the National College Choreography Initiative announced a grant to Tulane University to bring two choreographers, Sara Pearson and Patrik Widrig, to its New Orleans campus. Based on 2 years of research and over a decade of visits to New Orleans, they would work with students to create a dance about the environment, in collaboration with the university's Center for Bioenvironmental Research, to be performed on the Mississippi River.

Then Hurricane Katrina hit, the levees broke, and that didn't happen.

On the spur of the moment, with the help of faculty members at the University of Texas at Austin, students were transported to Texas to attend classes there. The choreographers could have abandoned the project, but instead reenvisioned it as *Katrina, Katrina: Love Letters to New Orleans.* Through their art, they expressed their great love for the Big Easy, and students were led through a process that helped them cope with being torn from their homes, friends, and college.

Dancers were costumed with remnants of emergency blue tarps and carried water bottles. Audiences were enthusiastic and others in the arts world took notice. The Kennedy Center booked it twice for its free performance series and then made it part of the center's national outreach program. It then toured the country.

But a logic model would probably have shown the project as an utter failure. After all, the environmental project never happened, the goal changed, most of the original partners didn't participate, and the stated results weren't achieved. Instead, the innovation on the part of artists and faculty members led to a transformative experience: They took stock of the situation and created an experience that would be meaningful to people during a time of dire concern about a national disaster.

Had Dance/USA, the organization that distributed the money, or the National Endowment for the Arts, which provided the funds, insisted on rigid accountability, the grant would not have been made, but Dance/USA released the funds and trusted the faculty members and artists. As stakeholders, they understood the context in which the grantee was operating and honored their extraordinary efforts to fulfill the project in a manner that served the colleges.

How should we evaluate this project? Would its impact be captured by saying that 350 people attended and that the budget balanced? Or should we design a longitudinal study to find out how many kids returned to New Orleans? Or finished college?

When we evaluated the National College Choreography Initiative program—including the Katrina project by Ms. Pearson and Mr. Widrig—we gathered statistics from all 34 projects. We learned that more than 10,000 students participated, 60,000 people attended 174 performances and 226 outreach activities, and more than $665,000 was raised to match the $272,000 in money that was initially provided.

But the numbers don't tell the whole story.

We systematically reviewed the stories that came from faculty members at all 34 colleges. From them, we identified more than 30 indicators of what had changed since the grants were made. Using content analysis, we developed a coding system for areas such as outreach beyond campus, professional networking, new jobs, in-state touring, collaborations with other colleges, and other indicators. We know, for example, that two of the grants led to professional opportunities for students, including a job with a ballet company, and eight involved opportunities for a program to spread beyond the campus grounds, such as the creation of dance curricula for elementary and secondary schools.

Each year, Dance/USA creates a publication that documents not just the statistics, but also the stories showing the successes of each project. And, the National Endowment for the Arts knows of the impact of its grant, and what was to be a one-time effort has now received four rounds of financial support.

SOURCE: Callahan (2007).

Suzanne Callahan is founder of Callahan Consulting for the Arts and the Laboratory for Arts Evaluation. Her book, *Singing Our Praises: Case Studies in the Art of Evaluation*, was awarded Outstanding Publication of the Year from the American Evaluation Association in 2005.

in teams will need to work hard to overcome their tendency to dismiss certain kinds of data without first considering seriously and fairly the merits of those data (Guest and MacQueen 2007). Exhibit 12.5 presents the evaluation standards for including and appropriately analyzing both quantitative and qualitative data in evaluations, giving equal weight to each.

The Gold Standard Debate

While it's not so hard to combine numbers with narratives to create mixed-methods evaluations, it is not so easy to combine experimental designs with naturalistic inquiry designs. The rigor and validity of experiments depend on controlling, standardizing, and precisely measuring the intervention and its effects. Naturalistic inquiry designs eschew control and observe the program as it unfolds naturally including the emergent and diverse effects on participants.

In considering the relative virtues of experimental versus naturalistic designs, the paradigms debate centers on the importance of causal questions in evaluation and how best to conduct impact evaluations. Those evaluation researchers who believe that the most important and central function of evaluation is to measure the effects of programs on participants to make valid causal inferences are strong advocates of randomized experiments as "the standard against which other designs for impact evaluation are judged" (Boruch and Rindskopf 1984:121). This is *the gold standard position* discussed earlier in this chapter. In advocating experimental designs as the gold standard, evaluation researchers such as Boruch (2007); Cook (2006); Rosen, Manor, Engelhard, and Zucker (2006); Lipsey (2007a, 2005, 1990); Schatschneider (2003); Shadish, Cook, and Campbell (2001); and Campbell and Boruch (1975) have demonstrated the power and feasibility of randomized experiments for a variety of programs and interventions. The concerns

EXHIBIT 12.5

Evaluation Standards for Quantitative and Qualitative Data

The evaluation standards give equal attention, weight, and credence to qualitative and quantitative data.

Program Outcomes—Document the full range of program outcomes, so that interested parties can assess the program's success against goals and assessed needs of intended beneficiaries and also assess its positive and negative side effects.

Document the qualitative and quantitative indicators that were employed to assess goal achievement. (S5)

Analysis of Quantitative Information—Appropriately and systematically analyze the evaluation's quantitative information, so that evaluation questions are effectively answered. (A7)

Analysis of Qualitative Information—Appropriately and systematically analyze the evaluation's qualitative information, so that evaluation questions are effectively answered. (A8)

SOURCE: The Omnibus Metaevaluation Checklist (Stufflebeam 2007).

that permeate these writings are concerns about increased rigor, well-controlled interventions, reducing threats to internal validity, precise estimates of program effects, and statistical power—which in combination increase confidence in attributing an outcome to an intervention.

Naturalistic inquiry, in contrast, involves observing ongoing programs as they unfold without attempting to control or manipulate the setting, situation, people, or data. Naturalistic inquiry evaluations look at programs within and in relation to their naturally occurring context. Instead of random assignment, for example, which controls who gets the treatment (program), naturalistic inquiry looks at how staff select participants or how they self-select into a program.

Randomized Control Trials: Heaven or Gold Standard

Randomized controlled trials (RCTs) are positioned as *the methodological gold standard* among advocates of experimental designs (Dugger 2004). Professor David Storey (2006) of Warwick Business School, University of Warwick, has offered a competing metaphor. He has posited "six steps to heaven" in conducting evaluations, where heaven is a randomized experiment. So materially oriented and worldly evaluators are admonished to aspire to the gold standard, while the more spiritually inclined can aspire to follow the path to heaven, where heaven is an RCT.

The metaphors of naturalistic inquiry are more along the lines of *staying grounded*, looking at *the real world* as it unfolds, *going with the flow*, *being adaptable*, and *seeing what emerges*.

Guba and Lincoln (1981) identified two dimensions along which types of scientific inquiry can be described: the extent to which the scientist manipulates some phenomenon in advance in order to study it, and the extent to which constraints are placed on output measures; that is, the extent to which predetermined categories or variables are used to describe the phenomenon under study. They then defined naturalistic inquiry as a "discovery-oriented" approach that minimizes investigator manipulation of the study setting and places no prior constraints on what the outcomes of the research will be. Naturalistic inquiry is thus contrasted to experimental research, in which, ideally, the investigator controls external influences and measures only hypothesized outcome variables.

Debate about whether experimental designs constitute the methodological gold standard revolves, in part, around what level and kind of evidence is needed to determine that an intervention is effective. Consider the challenge of eradicating intestinal worms in children, a widespread problem in developing countries (Bundy and Drake 2004; Drake and Bundy 2001; Brooker et al. 2000; Dickson et al. 2000; Albonico et al. 1997). Suppose we want to evaluate an intervention in which school-age children with diarrhea are given anti-worm medicine to increase their school attendance and performance. To attribute the intervention to the desired outcome, advocates of randomized controlled trials (RCTs) would insist on an evaluation design in which students suffering from diarrhea are randomly divided into a treatment group (those who receive worm medicine) and a control group (those who do not receive the medicine). The school attendance and test performance of the two groups would then be compared. If, after a month on the medicine, those receiving the intervention show higher attendance and

school performance at a statistically significant level compared with the control group (the counterfactual), then the increased outcomes can be attributed to the intervention (the worm medicine).

Advocates of qualitative inquiry question the value of the control group in this case. Suppose that students, parents, teachers, and local health professionals are interviewed about the reasons students miss school and perform poorly on tests. Independently, each of these groups assert that diarrhea is a major cause of the poor school attendance and performance. Gathering data separately from different informant groups (students, parents, teachers, heath professionals) is called *triangulation,* a way of checking the consistency of findings from different data sources. Following the baseline interviews, students are given a regimen of worm medicine. Those taking the medicine show increased school attendance and performance, and in follow-up interviews, the students, parents, teachers, and health professionals independently affirm their belief that the changes can be attributed to taking the worm medicine and being relieved of the symptoms of diarrhea. Is this credible, convincing evidence?

Those who find such a design sufficient argue that the results are both reasonable and empirical, and that the high cost of adding a control group is not needed to establish causality. Nor, they would assert, is it ethical to withhold medicine from students with diarrhea when relieving their symptoms has merit in and of itself. The advocates of RCTs respond that without the control group, other unknown factors may have intervened to affect the outcomes and that *only the existence of a counterfactual* (control group) will establish with certainty the impact of the intervention.

As this example illustrates, those evaluators and methodologists on opposite sides of this debate have different worldviews about what constitutes sufficient evidence for attribution and action in the real world. This is not simply an academic debate. Millions of dollars of evaluation funds are at stake and the results of these evaluations around the world will affect billions of dollars of international development assistance. Consider as another example RCTs evaluating microfinance loans being supported by the International Finance Corporation (IFC) of The World Bank. Microfinance programs give very small loans to people in extreme poverty without any collateral for the loans. With as little as $100, a group of women are able to purchase a sewing machine and make clothes for sale, or a group of men may purchase tools to set up a bicycle repair business. Microfinance loans provide capital to people in poverty when commercial banks are unwilling to take the risk of such loans or when those in poverty are subject to the extremely high interest rates of loan sharks. The differences in income can be quite small because the income levels and loan amounts are quite small. For example, in such a program in Pakistan, the Kashf Foundation reported that 90 percent of its clients were living on less than $1 a day and that, over time, those who had received loans reported 51 percent higher income than new clients applying for loans (Arjumand and Associates 2004).

IFC is funding evaluation of such microfinance programs with RCTs, randomly assigning loan applicants to those who receive the loans and a control group of people who do not receive loans. The financial status of people in both groups are compared over time, sometimes adding additional measures of health, social mobility, nutrition, and children's education. Differences on these indicators, if any, between the treatment and control group,

can confidently be attributed to the loans. One such study conducted by researchers from MIT's Poverty Action Lab, which specializes in conducting randomized experiments in developing countries, found that "those offered credit were more likely to retain wage employment, less likely to experience severe hunger in their households, and less likely to be impoverished" (Karlan and Zinman 2006:1).

In questioning the cost-benefit of such rigorous evaluation designs, advocates of naturalistic inquiry question the added value and expense of the control groups. In contrast to the randomized control group design, the naturalistic inquiry narrative would gather case data on the financial status and lives of people in poverty before they receive the loans. Their longtime history of poverty and lack of access to capital would be documented. Once they receive the loans, they would be periodically interviewed and observed to determine how they had used the loans and what differences the loans have made in their lives as reported by them and by others who know them (triangulation of sources). When the results show that they have used the loans to engage in economic activity that has increased their income and that the increased income has increased their quality of life, these narrative results would support a conclusion that the changes in their lives can reasonably be attributed to the loans. The connection between receiving the loans and enhanced lives is directly observable and measurable, and the attribution reasonable, without the need for a control group. Advocates of RCTs worry that other unknown factors may be at work and that the only way to establish attribution with confidence is to compare the intervention to a counterfactual (control group). The advocates of naturalistic inquiry find no added value in the control group and believe that the costs of monitoring the control group are unjustified and possibly unethical, especially in those designs where randomly denying people small loans puts them at risk of being perceived as being bad credit risks because they have, in fact, been turned down for a loan.

In a utilization-focused evaluation design process, these alternative design scenarios can be presented to primary intended users to help them determine what level of evidence is needed and appropriate given the purposes of and intended audiences for the evaluation. The MIT Poverty Action Lab is conducting a number of such experiments around the world for the IFC and other international donors. My experience with IFC decision makers is that they are treating RCTs as the gold standard because they want to be credible with academics and the Ph.D. economists of The World Bank. Politically, RCTs are the safe way to go to achieve academic respectability.

At a practical policy level, people in the field who implement these programs ask what burning evaluation question about the value of microfinance loans justifies continuing to fund RCTs. Muhammad Yunus won the 2006 Nobel Peace Prize for his work on microfinance that led to his founding the Grameen Bank, which has helped hundreds of thousands of people and has been disseminated as a model throughout the world, although adapted to local cultures and economies as it has been implemented. Yunus did not conduct RCTs to determine the value of small loans to impoverished women and men, but based his judgments on practical observations of the effects on thousands of people's lives. The Grameen Bank is considered a world-changing success story (Westley, Zimmerman, and Patton 2006). What, then, are the RCTs on

microfinance trying to prove? What is the policy question that justifies the large expense of RCT designs to evaluate microfinance programs?

Answering these questions involves decisions about what level of evidence is needed to establish the value of something and the cost-benefit of gathering such evidence. Those convinced by naturalistic inquiry narratives argue that the cause-effect linkage in microfinance is fairly direct and observable. For the cost of administering a control group design, a large number of additional microfinance loans could be given. Control group designs are expensive. The practical question, then, is the following: Given the nature of evidence that can be gathered by following up recipients of small loans to document the effects of those loans on the recipients' lives, what is the added value of a control group? Is the cost of such a control group (and the evidence it would yield) more valuable than making a larger number of loans and following up the effects of those additional loans, thereby substantially increasing the sample size for directly studying the intervention itself. This cost-benefit methodological decision also introduces *ethical considerations* into the trade-off between gathering data from a control group versus giving loans to more people.

There is also the practical question of using results from randomized experiments. One of the field-level development workers I met at an IFC conference on evaluation told me of the difficulties she experienced in administering the microfinance program because of the design rigidities of the experimental design. She was having trouble understanding and explaining the value of randomization beyond its status among academic researchers. She asked, "Can you imagine an agency, government, or bank running a microfinance program based on randomization? What of any practical significance do you learn

from randomization? We have limited resources," she went on. "We have to make selections. Even if randomization was more fair and equitable, giving loans randomly to applicants is not a viable political option. Indeed, the critical questions about microfinance are about how people become eligible, how they are selected to receive a loan, and how the decision is made about how much to give them. *Administering loans randomly is just not a viable policy option,*" she emphasized, shaking her head in frustration.

The randomization process, she felt, made the evaluation results less useful because the design was rigid and artificial. This, she had concluded, was a primary reason why no major businesses conduct such RCTs for their services. North American, European, and Australasian banks do not roll out new services with RCTs. They do, however, engage in thorough and rigorous evaluation. They try out pilot programs before going to scale. They seek customer feedback. They observe carefully what customers respond to and how they behave. They compare one delivery approach with another different delivery approach, with real customers in the real world, and they adjust their services accordingly. Likewise, Microsoft does not introduce and study new software through RCTs. They have a large group of pilot testers (as many as 750,000 worldwide) who provide real-time feedback from real-world uses of their software. One would think that if RCTs were so valuable in determining the effectiveness of services, this field worker speculated, businesses would use them routinely to boost profits. In point of fact, businesses engage in continuous improvement evaluation based on feedback and observing the reactions and behaviors of real customers as well as soliciting feedback from noncustomers. RCTs are more often an academic laboratory-like enterprise for research, not a real-world evaluation exercise to figure out how things work under real-world conditions. Or so goes the critique—and the debate.

Advocates of RCTs respond that you cannot really attribute the increased income and business activity of the loan recipients without a control group of people who did not receive the loans. For certain academics, using RCTs as *the only way* to establish causality is a matter of strong belief. To get results published in many of the most prestigious academic journals, a researcher needs to have conducted an RCT. *But what level of evidence does a typical policymaker need?* If small loans are given to very poor people with a history of poverty and an evaluator documents that they use those loans for businesses and make a small profit, which they use to improve their lives, is that sufficient evidence of the value of such a program? From a policy perspective, the "control condition" is not a viable policy option because *doing nothing is seldom an option.* International agencies and governments are not considering NOT doing microfinance programs. In an RCT design, then, a great deal of money is spent comparing doing something (a treatment) with not doing something (control), when the control condition is not really a policy option. In contrast to the "doing nothing" control condition, policymakers are interested in evaluating different ways of delivering microfinance loans. A comparison design that examines delivering microfinance in different ways, with alternative selection criteria, with alternative support mechanisms, and with alternative sizes of loans, would provide important policy results. An RCT answers one and only one question: Did this one particular approach produce this one particular outcome in this one particular

situation compared with a control group? That question tends to be of much less policy relevance and interest than the question, "What are the costs, benefits, and effects of delivering the intervention in different ways?"

Remember, at the root of paradigm debates are different formulations of the problem, different evaluation questions, and different beliefs about what level of evidence is needed to take action. Exhibit 12.6 provides a summary overview of the logic of experimental designs and 10 common critiques of such designs.

Credibility Issues

From Objectivity versus Subjectivity to Fairness and Balance

Qualitative evaluators are accused frequently of *subjectivity*—a term with the power of an epithet in that it connotes the very antithesis of scientific inquiry. Objectivity has been considered the sine qua non of the scientific method. To be subjective has meant to be biased, unreliable, and nonrational. Subjectivity implies opinion rather than fact, intuition rather than logic, and impression rather than rigor. Evaluators are advised to avoid subjectivity and make their work "objective and value free."

In the paradigms debate, the means advocated by scientists for controlling subjectivity through the scientific method were the techniques of the dominant quantitative/experimental paradigm. Yet quantitative and experimental methods can work in practice to limit and even bias the kinds of questions that are asked and the nature of admissible solutions. Michael Scriven (1972a), evaluation's long-time resident philosopher, has insisted that quantitative methods are no more synonymous with objectivity than qualitative methods are synonymous with subjectivity:

> Errors like this are too simple to be explicit. They are inferred confusions in the ideological foundations of research, its interpretations, its application. . . . It is increasingly clear that the influence of ideology on methodology and of the latter on the training and behavior of researchers and on the identification and disbursement of support is staggeringly powerful. Ideology is to research what Marx suggested the economic factor was to politics and what Freud took sex to be for psychology. (P. 94)

The possibility that "ideological" preconceptions can lead to dual perspectives about a single phenomenon goes to the very heart of the contrasts between paradigms. Two scientists may look at the same thing, but because of different theoretical perspectives, assumptions, or ideology-based methodologies, they may literally not see the same thing (Petrie 1972:48). Indeed, Kuhn (1970) has pointed out,

> Something like a paradigm is prerequisite to perception itself. When a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see. In the absence of such training there can only be, in William James's phrase, "a bloomin' buggin' confusion." (P. 113)

A child's parable, the story of Han and the Dragon, illustrates this point at another level of simplicity. Han, a small boy, lived in a city threatened by wild horsemen from the North. The Mandarin ruler and his advisers decided that only the Great Cloud Dragon could save the city, so they prayed for the Dragon's intervention. As he prayed, the Mandarin envisioned a dragon that looked like a proud lord—a Mandarin. The captain

EXHIBIT 12. 6

The Logic of Experimental Designs and 10 Common Criticisms

When experiments are advocated as the gold standard in evaluations it is because of how they assess *cause* and *effect* relationships. Experimental designs are viewed as the ideal because they control the hypothesized cause (the intervention or program) to ensure that the cause precedes the effect, that the cause is related to the effect, and that extraneous factors that could produce the effect can be ruled out. One of the few ways of establishing and validating causality a priori is to carefully design the experiment to test an explicit hypothesis, namely, *this* intervention will produce *this* outcome. Hypothesis testing avoids the problems and weaknesses of post hoc, after-the-fact, retrospective speculations on causality. As historian Lee Simonson observed, "Any event, once it has occurred, can be made to appear inevitable by a competent historian" (Forbes 2007:196).

Random Assignment and the Hypothetical Counterfactual

The ideal way to control extraneous influences is to randomly assign people (or other units of analysis, e.g., classrooms, programs, communities) to two groups: those that experience the hypothesized cause (the intervention) and a control group that does not receive the intervention. Random assignment controls selection bias and creates a *hypothetical counterfactual* that represents what would have happened to the treatment group had they not received the treatment. By statistically comparing the outcomes of the control group (hypothetical counterfactual) with the treatment group, the evaluator can assess the extent to which the outcomes can be attributed to the treatment (in other words, to judge whether the intervention caused the outcomes). Random assignment distributes potential extraneous or unknown causes across both groups so that the only difference between the two groups is the intervention. This increases confidence in determining causality because any other influence on the observed outcome would only occur by chance.

Replicability

Well-designed experiments allow for replication and contribute to meta-analyses. Two carefully conducted randomized experiments undertaken with subjects from the same population and using the same protocols should arrive at essentially the same results. The possibility of replication increases the credibility of any particular findings and meta-analyses of results from multiple experiments generate especially rigorous results (Lipsey and Wilson 2000).

Ten Common Criticisms of Experimental Designs for Evaluation

1. *Experiments Have Limited Applicability*. Experiments work for only quite specific, standardized, highly controlled and high-fidelity interventions, like an immunization or a standardized curriculum. A good example is an evaluation in 178 Kenyan schools conducted by researchers at MIT's Poverty Action Lab. Large, poster-sized flip charts were provided to one-half of the schools in classrooms covering science, mathematics, geography, and health. With 2 years of follow-up data, the evaluation concluded that the impact of flip charts on student test scores was nearly zero and statistically insignificant (Glewwe, Kremer, Moulin, and Zitzewitz 2004). Those skeptical of such narrow studies wonder why one would even conduct such a large-scale, expensive study to test the hypothesis that flip charts, by themselves, would raise test scores.

In the real world, program interventions are seldom as controlled and standardized as flip charts because staff adapts what they do to the needs of particular participants and changing circumstances, using multiple interventions together. Complex community interventions and programs that unfold over longer periods of time are especially hard to control and standardize during the experimental period. The simple, linear cause-effect models on which experimental designs are based cannot capture the complexities of complex, dynamic, nonlinear systems (see Chapter 10).

(Continued)

(Continued)

What if interventions that change only one thing at a time fail . . . because they change only one thing at a time? Then, the evaluators have defined out of evaluation consideration precisely the interventions most likely to have an impact. The multipronged, interactive, custom-tailored, evolving interventions that draw on many disciplines and systems to impact not only individuals but also neighborhoods, institutions, and systems are anathema to the traditional evaluator (Schorr 1998:144).

2. *Experiments Interfere with Adaptive Management and Continuous Improvement*. The requirement for standardization and control of the intervention can actually interfere with the program and reduce effectiveness because staff are constrained from adapting the intervention and individualizing the treatment. A rigorous experiment requires ensuring that the intervention has high fidelity (is rigorously implemented in a standardized manner), but this reduces flexibility and prohibits ongoing improvements in the program.
3. *The Black Box Critique*. RCTs may establish that an intervention caused an outcome but not *why* it did so. Moreover, unless there is very good implementation data about the intervention, the classic pre-post RCT may not be able to report details about what the intervention actually was, which severely limits interpreting the results. To understand how an intervention works, like, for example, international aid, evaluators have to "open the black box" (Bourguignon and Sundberg 2007).
4. *Failing to Learn from Natural Variation*. Within an intervention (or program), some people gain a lot, some a little, and some nothing. Because the primary (and often only) comparison is between the aggregate treatment group outcomes and the aggregate control group outcomes, significant within-group variation is not sufficiently analyzed and understood. What factors contribute to within-group variations? RCTs don't answer this question. Indeed, most experimental designs yield findings of no significant difference between the treatment and control, what Peter Rossi called "the iron law of evaluation" (Chen 2007; Rossi 1987). RCTs, by their very design, may fail to capture and understand important differences within the treatment group itself. With limited time and resources, as is always the case, which question would produce more useful results for policymaking and program improvement: (a) the RCT question that asks *how does the average outcome in the treatment group compare with the average outcome in the control group*, versus (b) the natural variation question that asks *what factors explain different levels of outcomes within the program and what are the implications of those different levels of outcome for setting policy and improving the program.*
5. *The Control Condition Is an Irrelevant Policy Option*. Comparing the costs and benefits of alternative interventions is typically more important to policymakers than a standard RCT design that compares a single intervention to a control group (no intervention). Doing nothing is seldom a policy option. Policymakers want data to choose among competing interventions or to establish the level at which an intervention must be implemented to achieve desired outcomes. Pragmatically, comparison of a treatment with a control is less interesting and meaningful because not doing anything is seldom the realistic policy alternative. This distinguishes the comparative method in program evaluation from RCTs (Scriven 1991b:112).

Moreover, finding meaningful control conditions for national programs such as Head Start can be especially problematic. One of evaluation's pioneers, Lois-ellin Datta, has thoughtfully reviewed the mandate by Congress to conduct a RCT of Head Start. Bringing her considerable experience and expertise to bear on this mandate, she concluded, "The randomized control design has no stronger proponent than me when the circumstances are appropriate. A primary reason that the design is inappropriate in the Head Start circumstance is that the control condition for the test is likely to be anything but that . . . For such situations, one perhaps thinks more of evaluation designs derived from sytems and complexity theories" (Datta 2007b:49–50).

6. *Limited Generalizability of Experimental Results*. RCTs aim for high "internal validity," that is, confidence that the outcome can be attributed to the intervention. To achieve this, RCTs control variation and extraneous factors. Controls are necessary to achieve a confident level of attribution and thereby make the specific results of a single RCT valid for that situation. *But those very same controls reduce the generalizability of the results*. The distinguished psychometrician and evaluation pioneer Lee J. Cronbach (1982) wrote an extensive treatise on the trade-offs between internal validity and external validity (generalizability) and concluded that policymakers are more often interested in extrapolating findings to other places than in ensuring the cause-effect relationship in one highly controlled setting. High internal validity typically reduces external validity, while designs that look at patterns across a number of variations can increase the likelihood of finding meaningful extrapolations. RCTs tend to pay little or no attention to contextual factors such as culture, societal context, and politics. Indeed, RCTs attempt to control for such factors through random assignment. But contextual factors are enormously important for understanding generalizability. Suppose one tests an HIV/AIDS education effort in one part of South Africa using an RCT. Would you have confidence that those results could be generalized to Niger? to Mexico? to China? Even to other parts of South Africa? Cronbach emphasized the importance of cumulative learning about the effectiveness of interventions as we go from one situation to the next, "refining our understanding as we go, as well as extrapolating what is learned in one setting to others."

Cronbach argued against using evaluation simply to answer the question, "Did this program cause the desired outcomes in participants?" both because the question is very difficult to answer in the diverse settings of the real world and because the question misses the point. The point being that because the program will look somewhat different in each context, evaluation should endeavor to understand in rich detail the challenges and the potentialities of a given social or educational intervention in this context and in that one and in that one over there, toward important insights into how to best address our persistent social problems" (Greene 2004:175).

7. *Randomization Is Artificial.* In the real world, people don't get into programs randomly. Self-selection, staff selection, and active recruitment, the bane of experimentalists, are the way participants get into real programs. Randomization reduces generalizability to real-world conditions. People tend to come to programs in social groups or to be selected through staff assessments that involve some degree of judgment. It is more useful to evaluate programs as they occur in the real rather than under the artificial, nongeneralizable, and nonsustainable conditions of randomization. A related critique is that the behavior of participants in the treatment group may be affected by the experiment (Hawthorne effect) since double-blind experiments typically are not possible in program evaluation. It is also often quite difficult to keep people in the control and treatment groups from having contact. These issues don't mean that experiments can't be done, but they are not easy to administer, and the more controls introduced to make sure the experiment is well implemented, the more artificial the results may become.

8. *High Costs of Control Group Designs*. Designs involve cost-benefit calculations: What is the value of likely findings given the costs of getting those findings? For the high cost of getting data from a control group (which as Item 5 asserts is typically not a viable policy option, and as the previous item asserts, has limited generalizability), an evaluator could gather more in-depth data comparing implementation factors, contextual variables, variations in outcomes, and comparing various real intervention alternatives.

(Continued)

(Continued)

9. *Ethical Concerns*. Control groups involve withholding an intervention from those in need. This is usually justified because there aren't enough resources to serve all those in need and/or because the intervention is unproven so it's by no means certain that something of value is being withheld. Many creative solutions to ethical concerns have been developed, but ethical gray areas remain and it can be politically difficult to explain why a group of people in need is being randomly denied a service.
10. *The experimental goal standard creates distorted incentives in making methods decisions*. The most basic wisdom in research and evaluation is that you begin by assessing the situation, figure out what information is needed, and determine the appropriate and relevant questions. The methods are then selected to answer those questions. However, when RCTs are treated as the *gold standard*, evaluators begin by asking, "How do I do an RCT?" This puts the method before the question. It also creates perverse incentives. For example, in some agencies, project managers are getting positive performance reviews and even bonuses for supporting and conducting RCTs. Under such incentive conditions, project managers will seek to do RCTs whether they are appropriate or not. No one wants to do a second-rate evaluation, but if RCTs are really the gold standard, anything else is second-rate. As distinguished evaluator Nick Smith (2007) has asked

When the federal anoints a particular method as the gold standard for conducting evaluation, is it not likely that many other groups . . . may infer that alternative methods are thus inferior? Might this not result in the overgeneralization and inappropriate use of the gold standard method and diminished use of alternative approaches that may be more effective in those contexts? (p. 120).

This also leads to rushing into RCT designs before the program is ready. For example, a widely circulated and influential report from the Center for Global Development advocates RCTs for impact evaluation of international development aid. The report posits that RCTs "must be considered from the start—the design phase—rather than after the program has been operating for many years, when stakeholders may ask, "So what is the program really accomplishing?" (Evaluation Gap Working Group 2006:13). That sounds reasonable, but for an RCT to work, an intervention (a program) must be clearly identified, standardized, and carefully controlled. This means you would never begin a new effort, program, or innovation with an RCT. The most well substantiated finding in a quarter century of evaluation may be that new efforts need a period of time to work out bugs, overcome initial implementation problems, and stabilize the intervention. Not even drug studies begin with RCTs. They begin with basic efficacy studies and dosage studies to find out if there is initial evidence that the drug produces the desired outcome without unacceptable side effects. Only then are RCTs undertaken. Beginning new projects with RTC designs shows a fundamental lack of understanding about how programs unfold in the real world. It fails to understand the role of formative evaluation in getting ready for summative evaluation. It also increases the likelihood of finding no impact because the intervention wasn't yet ready for summative RCT testing.

Nor do RCTs work well for complex interventions, like comprehensive, multifaceted community initiatives. Carol Weiss (2002) has observed: "Random assignment has a spare beauty all its own, but the sprawling changeable world of community programs is inhospitable to it" (p. 222). In explaining why an alternative was needed to RCTs for such initiatives, Lisbeth Schorr (1998)wrote

> The new approaches to the evaluation of complex interventions share at least four attributes: They are built on a strong theoretical and conceptual base, emphasize shared interests rather than adversarial relationships between evaluators and program people, employ multiple methods and perspectives, and offer both rigor and relevance. (P. 147)

of the army imagined and prayed to a dragon that looked like a warrior. The merchant thought that a dragon would appear rich and splendid, as he was. The chief workman was convinced that a dragon would be tough and strong. The wise man conceived of the dragon as "the wisest of all creatures," which meant it must look like a wise man. In the midst of the crisis, a small fat man with long beard and bald head arrived and announced that he was the Great Cloud Dragon. The Mandarin and his advisers ridiculed the old man and dismissed him rudely. Only because of Han's kindness did the old man save the city, transforming himself into a magnificent dragon the color of sunset shining through rain, scales scattering the light, claws and teeth glittering like diamonds, beautiful and frightening at the same time, and most important, beyond any possibility of preconception because the dragon was beyond prior human experience. But only Han saw the dragon, because only he was open to seeing it (Williams 1976).

Qualitative researchers prefer to describe themselves as open rather than subjective. They enter a setting without prejudgment, including no preconceived hypotheses to test. Scriven (1991b) has defined objectivity as being "unbiased or unprejudiced," literally, not having "prejudged." This definition

> misleads people into thinking that anyone who comes into a discussion with strong views about an issue can't be unprejudiced. The key question is whether the views are justified. The fact that we all have strong views about the sexual abuse of small children and the importance of education does not show prejudice, only rationality. (P. 248)

The debate about objectivity versus subjectivity includes different assumptions about whether it is possible for us to view the complexities of the real world without somehow filtering and simplifying those complexities. The qualitative assumption is that, at even the most basic level of sensory data, we are always dealing with perceptions, not "facts" in some absolute sense. "The very categories of things which comprise the 'facts' are theory dependent" (Petrie 1972:49) or, in this case, paradigm dependent. It was this recognition that led the distinguished qualitative sociologist Howard Becker (1970) to argue that "the question is not whether we should take sides, since we inevitably will, but rather whose side we are on" (p. 15).

Distinguished evaluation theorist and methodologist Robert Stake (2004) answered this question in an important article on advocacies in evaluation (see sidebar). He began by noting that we often care about the thing being evaluated—*and should care.* We don't have to pretend neutrality about the problems programs are attacking to do fair, balanced, and neutral evaluations of those programs. Who wants an uncaring evaluator who professes neutrality about homelessness, hunger, child abuse, community violence, or HIV/AIDS? My younger brother died of AIDS early in the epidemic. My entire family has been involved actively in AIDS Walks and other activities. When I am engaged with HIV/AIDS monitoring and evaluation systems (Patton 2004), I do not pretend neutrality. I want to see prevention programs work. That means I am motivated to hold staff feet to the fire of evaluation to assure that the program works—because I know from personal experience that lives are at risk.

As a utilization-focused evaluator, I find it helpful to replace the traditional scientific search for objective truth with a search for useful and balanced information. For the classic mandate to be objective, I substitute the mandate to be fair and conscientious in taking account of multiple perspectives, multiple interests, and multiple realities.

Beyond Neutrality: What Evaluators Care About

1. We often care about the thing being evaluated.
2. We, as evaluation professionals, care about evaluation.
3. We advocate rationality.
4. We care to be heard. We are troubled if our studies are not used.
5. We are distressed by underprivilege. We see gaps among privileged patrons and managers and staff and underprivileged participants and communities.
6. We are advocates of a democratic society.

SOURCE: Robert Stake (2004:103–107).

The Program Evaluation Standards reflect this change in emphasis:

Propriety Standard on Complete and Fair Assessment: The evaluation should be complete and fair in its examination and recording of strengths and weaknesses of the program being evaluated, so that strengths can be built on and problem areas addressed. (Joint Committee 1994:P5)

Accuracy Standard on Impartial Reporting: Reporting procedures should guard against distortion caused by personal feelings and biases for any party to the evaluation, so that evaluation reports fairly reflect the evaluation findings. (Joint Committee 1994:A11)

Words such as fairness, neutrality, and impartiality carry less baggage than objectivity and subjectivity. To stay out of arguments about objectivity, I talk with intended users about balance, fairness, and being explicit about what perspectives, values, and priorities have shaped the evaluation, both the design and findings. Others choose to use the term *objective* because of its political power. At the national policy level, former AEA President Eleanor Chelimsky recommended

> Although all of us realize that we can never be entirely objective, that is hardly an excuse for skewed samples, or grandiloquent conclusions or generalizations that go beyond the evaluator's data, or for any of 101 indications to a careful reader that a particular result is more desired than documented.
>
> There are, in fact, a great many things that we can do to foster objectivity and its appearance, not just technically, in the steps we take to make and explain our evaluative decisions, but also intellectually, in the effort we put forth to look at all sides and all stakeholders of an evaluations. (1995a:219)

The Continuum of Distance from versus Closeness to the Program

Here are the opposing paradigm positions: Too much closeness may compromise objectivity. Too much distance may diminish insight and understanding.

Quantitative researchers depend on distance to guarantee neutrality and academic integrity. Scholarly comportment connotes calm and detached analysis without personal involvement or emotion. The qualitative paradigm, in contrast, assumes that without empathy and sympathetic introspection derived from direct experience, one cannot fully understand a program. Understanding comes from trying to put oneself in the other

person's shoes, thereby discerning how others think, act, and feel. Methodologically, this means getting close to the action, observing people in the realities of program life, and attending to detail by observing program participants over time.

Qualitative evaluators strive to capture participants' experiences in their own terms, learn how they think about and experience the program. In the Shapiro study of FT open classrooms, her presence in classrooms over an extended period of time and her closeness to the children allowed her to see things that were not captured by standardized tests. She could see what they were learning. She could feel their tension in the testing situation and their spontaneity in the more natural classroom setting. Had she worked solely with data collected by others or only at a distance, she would never have discovered the crucial differences she uncovered between FT and non-FT classrooms—differences that allowed her to evaluate the innovative program in a meaningful and relevant way.

In a similar vein, one evaluator in our utilization of federal health evaluations expressed frustration at trying to make sense out of data from more than 80 projects when site visit funds were cut out of the evaluation: "There's no way to understand something that's just data, you know. You have to go look" [EV111: 3]. Qualitative methodologist John Lofland (1971) concluded likewise,

> In everyday life, statistical sociologists, like everyone else, assume that they do not know or understand very well people they do not see or associate with very much. They assume that knowing and understanding other people require that one see them reasonably often and in a variety of situations relative to a variety of issues. Moreover, statistical sociologists, like other people, assume that in order to know or understand others, one is well-advised to give some conscious attention to that effort in face-to-face contracts. They assume, too, that the internal world of sociology—or any other social world—is not understandable unless one has been part of it in a face-to-face fashion for quite a period of time. How utterly paradoxical, then, for these same persons to turn around and make, by implication, precisely the opposite claim about people they have never encountered face-to-face—those people appearing as numbers in their tables and as correlations in their matrices! (P. 3)

It is instructive to remember that many major contributions to our understanding of the world have come from scientists' personal experiences—Piaget's closeness to his children, Freud's proximity to and empathy with his patients, Darwin's closeness to nature, and even Newton's intimate encounter with an apple.

On the other hand, closeness is not the only way to understand human behavior. For certain questions and for situations involving large groups, distance is inevitable. But, where possible, face-to-face interaction can deepen insight, especially in program evaluation. This returns us to the recurrent themes of using mixed methods and matching evaluation methods to intended use by intended users.

The issue of distance from versus closeness to the program supersedes methods in that it concerns the basic relationship between evaluators and those being evaluated, which affects how evaluations are used. VanLandingham (2007) has contended that, in the U.S. federal government, too much independence "can restrict evaluators' role to that of a voice crying in the wilderness rather than speaking truth to power" (p. 25). In a similar vein, experienced auditors Perry, Thomas, DuBois, and McGowan (2007) presented a case study involving county jails in which they

concluded that "the traditional focus on ensuring independence has led agencies conducting legislative audits to avoid utilization-focused strategies and overlook the benefits of engaging stakeholders" (p. 69). They found that they could conduct a utilization-focused evaluation that

> not only provided independently verified information and analysis of current jail operations and costs, but also developed objective tools that proved useful for prospective analysis by stakeholders. . . . Working closely with agency management and developing tools and methodologies for their future use did not compromise the independence of the auditing function. (P. 76)

Of Variables and Wholes

The quantitative/experimental paradigm operationalizes independent and dependent variables, then measures their relationships statistically. Outcomes must be identified and measured as specific variables. Treatments and programs must also be conceptualized as discrete, independent variables. Program participants are also described along standardized, quantified dimensions. Sometimes a program's goals are measured directly, for example, student achievement test scores, recidivism statistics for a group of juvenile delinquents, or sobriety rates for participants in chemical dependency treatment programs. Evaluation measures can also be indicators of a larger construct, for example, "community well-being" as a general construct measured by indicators such as crime rates, fetal deaths, divorce, unemployment, suicide, and poverty.

Adherents of the qualitative paradigm argue that the variables-based approach (1) oversimplifies the interconnected complexities of real-world experiences, (2) misses major factors of importance that are not easily quantified, and (3) fails to capture a sense of the program and its impacts as a "whole." The qualitative/naturalistic paradigm strives to be holistic in orientation. It assumes that the whole is greater than the sum of its parts; that the parts cannot be understood without a sense of the whole; and that a description and understanding of a program's context is essential to an understanding of program processes and outcomes. This, of course, follows the wisdom of the fable about the blind children and the elephant. As long as each felt only a part—a fanlike ear, the ropelike tail, a treelike leg, the snakelike trunk—they could not make sense of the whole elephant. The qualitative, systems-oriented paradigm goes even further. Unless they could see the elephant at home in the African wilderness, they would not understand the elephant's ears, legs, trunk, and skin in relation to how the elephant has evolved in the context of its ecological niche.

Philosopher and educator John Dewey (1956) advocated a holistic approach to both teaching and research, if one was to reach into and understand the world of the child.

> The child's life is an integral, a total one. He passes quickly and readily from one topic to another, as from one spot to another, but is not conscious of transition or break. There is no conscious isolation, hardly conscious distinction. The things that occupy him are held together by the unity of the personal and social interests which his life carries along. . . . [His] universe is fluid and fluent; its contents dissolve and re-form with amazing rapidity. But after all, it is the child's own world. It has the unity and completeness of his own life. (Pp. 5–6)

Again, Shapiro's (1973) work in evaluating innovative FT classrooms is instructive. She found that test results could not be interpreted without understanding the

larger cultural and institutional context in which the individual child was situated. Nor is this only true for children. Beyer and Gillmore (2007) have made the case for more holistic, longitudinal, and multidimensional assessment of student learning in higher education because "simplistic measures aren't enough" (p. 43). Years ago Deutscher (1970) cautioned that despite our personal experience as living, working human beings, we have focused in our research on parts to the virtual exclusion of wholes:

> We knew that human behavior was rarely if ever directly influenced or explained by an isolated variable; we knew that it was impossible to assume that any set of such variables was additive (with or without weighting); we knew that the complex mathematics of the interaction among any set of variables was incomprehensible to us. In effect, although we knew they did not exist, we defined them into being. (P. 33)

Although most scientists would view this radical critique of variable analysis as too extreme, I find that teachers and practitioners often voice the same criticisms. Innovative teachers complain that experimental results lack relevance for them because they have to deal with the whole in their classrooms; they can't manipulate just a couple of factors in isolation from everything else going on. The reaction of many program staff to scientific research is like the reaction of Copernicus to the astronomers of his day: "With them," he observed,

> it is as though an artist were to gather the hands, feet, head, and other members for his images from diverse models, each part excellently drawn, but not related to a single body, and since they in no way match each other, the results would be monster rather than man. (Quoted in Kuhn 1970:83)

How many program staff have complained of the evaluation research monster?

Yet it is no simple task to undertake holistic evaluation, to search for the Gestalt in programs. The challenge for the participant observer is "to seek the essence of the life of the observed, to sum up, to find a central unifying principle" (Bruyn 1966:316).

The advantages of using variables and indicators are parsimony, precision, and ease of analysis. Where key program elements can be quantified with validity, reliability, and credibility, and where necessary statistical, normality, and independence of measurement, statistical portrayals can be quite powerful and succinct. The advantage of in-depth case studies and qualitative portrayals of holistic settings and impacts is that attention can be given to nuance, setting, interdependencies, complexities, idiosyncrasies, and context. In combination, the two approaches can be powerful and comprehensive; they can also be contradictory and divisive.

Two Views of Change

The paradigms debate is in part about how best to understand and study change. The quantitative/experimental paradigm typically involves gathering data at two points in time, pretest and posttest, then comparing the treatment group with the control group statistically. Ideally, participants are assigned to treatment and control groups randomly, or, less ideally, are matched on critical background variables. Such designs assume an identifiable, coherent, and consistent treatment. Moreover, they assume that, once introduced, the treatment remains relatively constant and unchanging. In some designs, time series data are gathered at several predetermined points rather than just at pretest and posttest. The

purpose of these designs is to determine the extent to which the program (treatment) accounts for measurable changes in participants to make a summative decision about the value and effectiveness of the program in producing desired change (Lipsey 1990; Boruch and Rindskopf 1984; Mark and Cook 1984).

In contrast, the qualitative/naturalistic paradigm conceives programs as dynamic and ever developing, with "treatments" changing in subtle but important ways as staff members learn, as clients move in and out, and as conditions of delivery are altered. Qualitative/naturalistic evaluators seek to describe these dynamic program processes and understand their holistic effects on participants. Thus, part of the paradigms debate has been about the relative utility, desirability, and possibility of understanding programs from these quite different perspectives for different purposes.

The quantitative/experimental/summative approach is most relevant for fairly established programs with stable, consistent, and identifiable treatments and clearly quantifiable outcomes, in which a major decision is to be made about the effectiveness of one treatment in comparison with another (or no) treatment.

The qualitative/naturalistic/formative approach is especially appropriate for developing, innovating, or changing programs in which the focus is improving the program, facilitating more effective implementation, and exploring a variety of effects on participants. This can be particularly important early in the life of a program or at major points of transition. As an innovation or program change is implemented, it frequently unfolds in a manner quite different from what was planned or conceptualized in a proposal. Once in operation, innovative programs are often changed as practitioners learn what works and what does not, and as they experiment, grow, and change their priorities. Developmental evaluation, which tracks incremental changes and forks-in-the-road over time, takes a dynamic view of programs.

Changing developmental programs can frustrate evaluators whose design approach depends on specifiable unchanging treatments to relate to specifiable predetermined outcomes. Evaluators have been known to do everything in their power to stop program adaptation and improvement so as to maintain the rigor of their research design. The deleterious effect this may have on the program itself, discouraging as it does new developments and redefinitions in midstream, is considered a small sacrifice made in pursuit of higher-level scientific knowledge. But there is a distinct possibility that such artificial evaluation constraints will contaminate the program treatment by affecting staff morale and participant response.

Were some science of planning and policy or program development so highly evolved that initial proposals were perfect, one might be able to sympathize with these evaluators' desire to keep the initial program implementation intact. In the real world, however, people and unforeseen circumstances shape programs, and initial implementations are modified in ways that are rarely trivial.

Under conditions in which programs are subject to change and redirection, the naturalistic evaluation paradigm replaces the static underpinnings of the experimental paradigm with a dynamic orientation. A dynamic evaluation is not tied to a single treatment or to predetermined outcomes but rather focuses on the actual operations of a program over a period of time, taking as given the complexity of a changing reality and variations in participants' experiences over the course of program participation.

Again, the issue is one of matching the evaluation design to the program, of meshing evaluation methods with decision-maker information needs. The point of contrasting fixed experimental designs with dynamic process designs in the paradigms debate is to release evaluators "from unwitting captivity to a format of inquiry that is taken for granted as the naturally proper way in which to conduct scientific inquiry" (Blumer 1969:47).

Nowhere is this unwitting captivity better illustrated than in those agencies that insist, in the name of science, that all evaluations must employ experimental designs. Two examples will illustrate this problem. In Minnesota, the Governor's Commission on Crime Prevention and Control required experimental evaluation designs of all funded projects. A small Native American alternative school was granted funds to run an innovative crime prevention project with parents and students. The program was highly flexible; participation was irregular and based on self-selection. The program was designed to be sensitive to Native American culture and values. It would have been a perfect situation for formative responsive evaluation. Instead, program staff was forced to create the illusion of an experimental pretest and posttest design. The evaluation design interfered with the program, alienated staff, wasted resources, and collected worthless information, unrelated to evolving program operations, under the guise of maintaining scientific consistency. The evaluators refused to alter or adapt the design and data collection in the face of a program dramatically different from the preconceptions on which they had based the design.

The second example is quite similar but concerns the Minnesota Department of Education. The state monitor for an innovative arts program in a free school for at-risk students insisted on quantitative, standardized test measures collected in pretest and posttest situations; a control group was also required. The arts program was being tried out in a free school as an attempt to integrate art and basic skills. Students were self-selected and participation was irregular; the program had multiple goals, all of them vague; even the target population was fuzzy; and the treatment depended on who was in attendance on a given day. The free school was a highly fluid environment for which nothing close to a reasonable control or comparison group existed. The teaching approach was highly individualized, with students designing much of their program of study. Both staff and students resented the imposition of rigid, standardized criteria that gave the appearance of a structure that was not there. Yet the Department of Education insisted on a static, hypothetico-deductive evaluation approach because "it's departmental evaluation policy."

On the other hand, the direction of the design error is not always the imposition of overly rigid experimental formats. Boruch (2007), Cook (2006), and Campbell and Boruch (1975) have shown that many evaluations suffer from an underutilization of experimental designs, which may do a disservice to program by underestimating outcomes and removing uncertainty about attribution. Eminent evaluation methodologist Peter Rossi emphasized that rigorous experimental designs increased credibility by permitting replication; because statistical tools can be implemented in systematic ways, it is

> both possible and desirable for any quantitative analysis to be replicated. Two randomized experiments undertaken with subjects from the same population and using the same protocols should arrive at the same results, save for a bit of noise. If

> one program evaluator analyzed the data set, it should be possible for another program evaluator to retrace the steps undertaken and arrive at the same results. The threat of replication helped keep all parties honest and, when results were reproduced, helped bolster the credibility of any findings. (Berk 2007:204)

Matching methods to programs and decision-maker needs is a creative process that emerges from a thorough knowledge of the organizational dynamics and information uncertainties of a particular context. Regulations to the effect that all evaluations must be of a certain type serve neither the cause of increased scientific knowledge nor that of greater program effectiveness, which was the central message of the AEA statement on experimental designs discussed earlier in this chapter (see Exhibit 12.2). Julnes and Rog (2007) edited an important volume of *New Directions for Evaluation* on "Informing Federal Policies on Evaluation Methodology: Building the Evidence Base for Method Choice in Government Sponsored Evaluation." It provides important insights into how the paradigms debate translates into practical issues of "actionable evidence" (pp. 4–5).

Alternative Sampling Logics

The quantitative paradigm employs random samples sufficient in size to permit valid generalizations and appropriate tests of statistical significance. Qualitative inquiry involves small "purposeful samples" of information-rich cases (Patton 2002a:230–47). Differences in logic, assumptions, and purposes distinguish these sampling strategies. When the evaluation is aimed at generalization, some form of random probabilistic sampling is the design of choice. A needs assessment, for example, aimed at determining how many residents in a county have some particular problem would be strongest if based on a random sample of county residents.

Case studies, on the other hand, become particularly useful when intended users need to understand a problem, situation, or program in great depth, and they can identify cases rich in needed information—"rich" in the sense that a great deal can be learned from a few exemplars of the phenomenon of interest. For example, much can be learned about how to improve a program by studying dropouts or successes *within the context* of a particular program. Case studies are context specific.

But what about generalizations? Paradigm differences emerge in the relative value attached to generalizing.

Cronbach (1975) observed that generalizations decay over time; that is, they have a half-life much like radioactive materials. Guba and Lincoln (1981) were particularly critical of the dependence on generalizations in quantitative methods because, they asked, "What can a generalization be except an assertion that is context free? . . . [Yet] *it is virtually impossible to imagine any human behavior that is not heavily mediated by the context in which it occurs*" (p. 62).

Cronbach and colleagues in the Stanford Evaluation Consortium (1980) offered a middle ground in the paradigms debate with regard to the problem of generalizability and the relevance of evaluations. They criticized experimental designs that were so focused on controlling cause and effect that the results were largely irrelevant beyond the experimental situation. On the other hand, they were equally concerned that entirely idiosyncratic case studies yield little of use beyond the case study setting. They suggested, instead, that designs balance depth and breadth, realism and control, so as to permit reasonable *extrapolation* (pp. 231–35).

Unlike the usual meaning of the term *generalization*, an *extrapolation* connotes that one has gone beyond the narrow confines of the data to think about other applications of the findings. Extrapolations are modest speculations on the likely applicability of findings to other situations under similar, but not identical, conditions. Extrapolations are logical, thoughtful, and problem oriented rather than purely empirical, statistical, and probabilistic. Evaluation users often expect evaluators to thoughtfully extrapolate from their findings in the sense of pointing out lessons learned and potential applications to future efforts.

Designs that combine probabilistic and purposeful sampling (mixed methods designs) have the advantage of extrapolations supported by quantitative and qualitative data. Larger samples of statistically meaningful data can address questions of incidence and prevalence (generalizations to a known population), while case studies add depth and detail to make interpretations more meaningful and grounded. Such designs can also introduce a balance between concerns about individualization and standardization, the distinction in the next section.

Standardization or Diversity: Different Emphases

The quantitative paradigm requires the variety of human experience to be captured along standardized scales. Individuals and groups are described as exhibiting more or less of some trait (self-esteem, satisfaction, competence, knowledge), but everyone is rated or ranked on a limited set of predetermined dimensions. Statistical analyses of these dimensions present central tendencies (averages and deviations from those averages). Critics of standardized instrumentation and measurement are concerned that such an approach only captures quantitative differences thereby missing significant qualitative differences and important idiosyncrasies. Critics of statistics are fond of telling about the person who drowned in a creek with an average depth of 6 inches; what was needed was some in-depth information about the 6-foot pool in the middle of the creek.

The qualitative paradigm pays particular attention to uniqueness, whether this be an individual's uniqueness or the uniqueness of a program, community, home, or other unit of analysis. When comparing programs, the qualitative evaluator begins by trying to capture the unique, holistic character of each program with special attention to context and setting. Patterns across individuals or programs are sought only after the uniqueness of each case has been described.

For program staff in innovative programs aimed at individualizing treatments, the central issue is how to identify and deal with individual differences among participants. Where the emphasis is on individualization of teaching or on meeting the needs of individual clients in social action programs, an evaluation strategy of case studies is needed that focuses on the individual, one that is sensitive both to unique characteristics in people and programs and to similarities among people and commonalities across treatments. Case studies can and do accumulate. Anthropologists have built up an invaluable wealth of case study data that includes both idiosyncratic information and patterns of culture (Human Relations Area Files 2007).

Using both quantitative and qualitative approaches can permit the evaluator to address questions about quantitative differences on standardized variables and qualitative differences reflecting individuals and program uniquenesses. The more a program aims at individualized outcomes, the greater the appropriateness of qualitative methods. The more a program emphasizes

common outcomes for all participants, the greater the appropriateness of standardized measures of performance and change.

Whither the Evaluation Methods Paradigms Debate?

Evaluation is much too important to be left to the methodologists.

—Halcolm

Early in the development of evaluation, the paradigms debate became characterized and labeled as the *qualitative-quantitative debate*. Overall, in the last quarter century, evaluation has become more methodologically eclectic with an increased emphasis on methodological appropriateness—matching the data collection and design to the nature of the evaluation situation and questions, and the information priorities of primary stakeholders. This makes *methodological pluralism and appropriateness the new gold standard* (e.g., Lawrenz and Huffman 2006). This is even true of advocates of experimental designs, who focus their advocacy primarily on summative evaluations for attributing impact, recognizing that other kinds of evaluation (e.g., formative and developmental) benefit from other methods. Many evaluation theorists and methodologists have worked to resolve conflict as an artist might, creating swirls and strokes to connect ideas and approaches where there was once the void of misunderstanding and mistrust. Eight trends support and illuminate this movement toward methodological appropriateness as the true gold standard for evaluation as posited in Exhibit 12.7.

EXHIBIT 12.7

Gold Standard Question Revisited: Methodological Appropriateness Trumps Experimental Design Orthodoxy

This chapter opened by observing that the *gold standard question in evaluation* is whether one particular method—*randomized control experiments*—should be held up as the best design for conducting impact evaluations and, by being best, should be the standard of excellence toward which evaluators should aspire and against which the quality of evaluation methods are judged. Do randomized control experiments merit the Olympic gold medal for evaluation? That is at the center of the methodological paradigms debate today.

The Utilization-Focused Evaluation Gold Standard Is *Methodological Appropriateness*

Methodological appropriateness means matching the evaluation design to the evaluation situation taking into account the priority questions and intended uses of primary intended users, the costs and benefits of alternative designs, the decisions that are to be made, the level of evidence necessary to support those decisions, ethical considerations, and utility. No design should be lauded as a gold standard without regard to context and situation. To do so is to create incentives to do randomized control experiments regardless of their appropriateness or meaningfulness.

1. *Evaluation has matured as a genuinely interdisciplinary and multimethod field of professional practice*. A balanced approach to methods has become commonplace with increasing emphasis on using mixed methods whenever possible to overcome the inherent and inevitable weaknesses and limitations of any single method. Methodological tolerance, flexibility, and concern for appropriateness rather than orthodoxy now characterize the practice, literature, and discussions of evaluation as evidenced by the AEA statement in Exhibit 12.1. The sense of tolerance and emphasis on appropriateness is nicely captured in the title of the volume of *New Directions for Evaluation* edited by Julnes and Rog (2007): "Informing Federal Policies on Evaluation Methodology: Building the Evidence Base for Method Choice in Government Sponsored Evaluation." Note the emphasis on *choice*. In their introduction, they emphasize that different kinds of evidence can inform different kinds of actions.

2. *Increasing attention to evaluation use has contributed to this methodological diversity*. When the utilization crisis emerged in the 1960s, two major recommendations for solving the problem were offered. The first focused on upgrading methodological rigor to increase the accuracy, reliability, and validity of evaluation data, and thereby increasing use. The second set of recommendations focused on evaluation processes: increasing attention to stakeholder needs, acting with greater political savvy, championing findings among intended users, and matching methods to questions. Methodological rigor alone has not proven an effective strategy for increasing use. Direct attention to issues of use, as in utilization-focused evaluation, has proven effective. High-quality evaluations manifest both technical adequacy and utility.

3. *Professional standards adopted by evaluation associations around the world have emphasized methodological appropriateness rather than paradigm orthodoxy*. These standards (e.g., Stufflebeam 2007) provide criteria in addition to methodological quality for judging the excellence of evaluations. This has made it possible to employ a variety of methods and still do an evaluation judged of high quality.

4. *Attention to general evaluation competencies and the accumulation of practical evaluation experience during the last two decades has reduced paradigms polarization*. The practical experience of evaluators working to improve program effectiveness has led them to become pragmatic in their approaches to methods issues. In that *pragmatism* (Morgan 2007) has emerged a commitment to do what works rather than a commitment to methodological rigor as an end in itself. This also means having more than methodological competence. The important and influential work of King et al. (2001) on a "Taxonomy of Essential Program Evaluator Competencies" shows that professional evaluators not only need more than "Systematic Inquiry" skills (which include knowledge of quantitative, qualitative, and mixed methods) but also skills in Professional Practice, Situational Analysis, Project Management, Reflective Practice, and Interpersonal Competence (e.g., communication skills).

5. *The strengths and weaknesses of both quantitative/experimental methods and qualitative/naturalistic methods are now better understood*. In the original debate, quantitative methodologists tended to attack some of the worst examples of qualitative evaluations while the qualitative evaluators tended to hold up for critique the worst examples of quantitative/experimental

approaches. With exemplars of both qualitative and quantitative evaluations, analyses of the strengths and weaknesses of each, and experience in how to combine methods, the meaning and utility of methodological appropriateness has become clearer.

6. *Advances in methodological sophistication and diversity within both paradigms, and in mixed methods, have strengthened diverse applications to evaluation problems*. The proliferation of books and journals in evaluation, including but not limited to methods contributions, has converted the field into a rich mosaic that cannot be reduced to quantitative versus qualitative in primary orientation. This is especially true of qualitative methods, which had more catching up to do, in which a great deal of important work has been published addressing questions of validity, reliability, and systematic analysis (Stake 2005; Patton 2002a; Yin 2002; Denzin and Lincoln 2000). The paradigms debate, in part, increased the amount of qualitative and mixed methods work being done, created additional opportunities for training in qualitative methods, and brought attention by methodologists to problems of increasing the quality of qualitative data and mixed methods designs. As the quality of qualitative methods has increased and the utility of qualitative approaches has been demonstrated, the attacks on qualitative methods have become both less strident and less common. The same can be said of developments in quantitative/experimental methods, as methodologists have focused on fine-tuning and adapting social science methods to a variety of evaluation and public policy situations (Patton 2008; Scriven 2008; Julnes and Rog 2007; Mohan and Sullivan 2007; Hudley and Parker 2006; Durland and Fredericks 2005; Braverman et al. 2004; Greene and Caracelli 1997; Sechrest and Scott 1993; Smith 1992; Lipsey 1990; and Trochim 1986). Lipsey (1988), whose quantitative/experimental credentials are impeccable, epitomized the emergent commitment to matching methods to problems and situations when he concluded

> Much less evaluation research in the quantitative-comparative mode should be done. Though it is difficult to ignore the attractiveness of assessing treatment effects via formal measurement and controlled design, it is increasingly clear that doing research of this sort well is quite difficult and should be undertaken only under methodologically favorable circumstances, and only then with extensive prior pilot-testing regarding measures, treatment theory, and so forth. The field of evaluation research and the individual treatments evaluated would generally be better served by a thorough descriptive, perhaps qualitative, study as a basis for forming better concepts about treatment, or a good management information system that provides feedback for program improvement, or a variety of other approaches rather than by a superficially impressive but largely invalid experimental study. (Pp. 22–23)

7. *Support for methodological eclecticism from major figures and institutions in evaluation has increased methodological tolerance*. Early in this chapter, I noted that when eminent measurement and methods scholars such as Donald Campbell and Lee J. Cronbach, their commitment to rigor never being in doubt, began publicly recognizing the contributions that qualitative methods could make, the acceptability of qualitative/naturalistic approaches was greatly enhanced. Another important endorsement of multiple methods came from the Program Evaluation and Methodology Division of the United States General Accounting Office (GAO), which arguably did the most important and influential evaluation work at the national level

(until it was disbanded in 1996). Under the leadership of Assistant Comptroller General and Former AEA President (1995) Eleanor Chelimsky, GAO published a series of methods manuals, including *Quantitative Data Analysis* (GAO 1992d), *Case Study Evaluations* (GAO 1990a), *Prospective Evaluation Methods* (GAO 1990b), and *The Evaluation Synthesis* (GAO 1992c). The GAO manual on *Designing Evaluations* (1991) puts the paradigms debate to rest as it describes what constitutes a strong evaluation. Strength is not judged by adherence to a particular paradigm. It is determined by use and technical adequacy, whatever the method, within the context of purpose, time, and resources.

Strong Evaluations

Strong evaluations employ methods of analysis that are appropriate to the question; support the answer with evidence; document the assumptions, procedures, and modes of analysis; and rule out the competing evidence. Strong studies pose questions clearly, address them appropriately, and draw inferences commensurate with the power of the design and the availability, validity, and reliability of the data. Strength should not be equated with complexity. Nor should strength be equated with the degree of statistical manipulation of data. Neither infatuation with complexity nor statistical incantation makes an evaluation stronger.

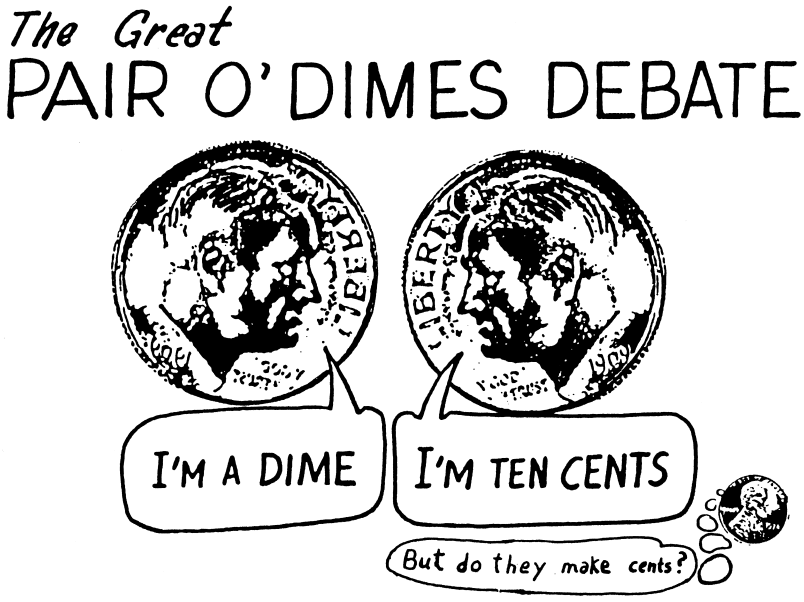
The strength of an evaluation is not defined by a particular method. Longitudinal, experimental, quasi-experimental, before-and-after, and case study evaluations can be either strong or weak. . . . That is, the strength of an evaluation has to be judged within the context of the question, the time and cost constraints, the design, the technical adequacy of the data collection and analysis, and the presentation of the findings. A strong study is technically adequate and useful—in short, it is high in quality.

SOURCE: From *Designing Evaluations*, Government Accountability Office (1991:15–16).

8. *Evaluation professional societies have supported exchanges of views and high-quality professional practice in an environment of tolerance and eclecticism.* The evaluation professional societies and journals serve a variety of people from different disciplines who operate in different kinds of organizations at different levels, in and out of the public sector, and in and out of universities. This diversity, and opportunities to exchange views and perspectives, has contributed to the emergent pragmatism, eclecticism, and tolerance in the field. A good example is the volume of New Directions for Program Evaluation on "The Qualitative-Quantitative Debate: New Perspectives" (Reichardt and Rallis 1994a). The tone of the eight distinguished contributions in that volume is captured by phrases such as "peaceful coexistence," "each tradition can learn from the other," "compromise solution," "important shared characteristics," and "a call for a new partnership" (Datta 1994; Reichardt and Rallis 1994b, 1994c; Rossi 1994; Yin 1994). That volume also emphasized mixed methods and included these themes: "blended approaches," "integrating the qualitative and quantitative," "possibilities for integration," "qualitative plus quantitative," and "working together" (Datta 1994; Hedrick 1994; House 1994; Reichardt and Rallis 1994c; Smith 1994; see also Mark and Shotland 1987).

Pragmatism Ascendant

Over the years of the debate, philosophical paradigms debate that focuses on fundamental differences in epistemology and ontology has been distinguished from the more narrow methodological paradigms debate. For example, Guba and Lincoln (1981) have argued that the experimentalist (scientific) and naturalistic paradigms contain incompatible assumptions about the inquirer/subject relationship and the nature of truth. The experimental/scientific paradigm assumes that reality is "singular, convergent, and fragmentable," while the naturalistic paradigm holds a view of reality that is "multiple, divergent, and inter-related" (Guba and Lincoln 1981:57). These opposite assumptions are not about methods alternatives; they are fundamental assumptions about the nature of reality. Pragmatically, an evaluator can conduct interviews and observations under either set of assumptions—and the data will stand on their own. Let me illustrate.



An evaluator is working with a group of educators, some of whom are "progressive, open education" adherents and some of whom are "back-to-basics" fundamentalists. The open education group wants to frame the evaluation of a particular program within a qualitative/naturalistic framework. The basic skills people want a rigorous, quantitative/experimental approach. Must the evaluator make an either/or choice to frame the evaluation within either one or the other paradigm? Must an either/or choice be made about the kind of data to be collected? Are the views of each group so incompatible that each must have its own evaluation?

I've been in precisely this situation a number of times. I do not try to resolve their paradigms debate but, rather, to inform their dialogue. I try to establish an environment of tolerance and respect for different, competing viewpoints, and then

focus the discussion on the actual information that is needed by each group: Test scores? Interviews? Observations? The design and measures must be negotiated. Multiple methods and multiple measures will give each group some of what they want. The naturalistic paradigm educators will want to be sure that test scores are interpreted within a larger context of classroom activities, observations, and outcomes. The quantitative paradigm educators will likely use interview and observational data as background to explain and justify test score interpretations. My experience suggests that both groups can agree on an evaluation design that includes multiple types of data and that each group will ultimately pay attention to and use "the other group's data." In short, a particular group of people can arrive at agreement on an evaluation design that includes both qualitative and quantitative data without resolving ultimate paradigmatic issues (e.g., whether reality is absolute or socially constructed). Such agreement is not likely, however, if the evaluator begins with the premise that the paradigms are incompatible and that the evaluation must be conducted within the framework of either one or the other.

Perhaps an analogy will help here. A sensitive, practical evaluator can work with a group to design a meaningful evaluation that integrates concerns from both paradigms in the same way that a skillful teacher can work with a group of Buddhists, Christians, Jews, and Muslims on issues of common empirical concern without resolving which religion has the "correct" worldview.

Another example, an agricultural project in the Caribbean that included social scientists and government officials of varying political persuasions. Despite their fundamental policy and philosophical differences, the Marxist and Keynesian economists and sociologists had little difficulty agreeing on what data were needed to understand agricultural extension needs in each country. Their interpretations of those data also differed less than I expected.

Thus, the point I'm making about the paradigms debate extends beyond methodological issues to embrace a host of potential theoretical, philosophical, religious, and political perspectives that can separate the participants in an evaluation process. I am arguing that, from a practical perspective, the evaluator need not even attempt to resolve such differences. By focusing on and negotiating data collection alternatives in an atmosphere of respect and tolerance, the participants can come together around a commitment to an empirical perspective, that is, bringing data to bear on important program issues. As long as the empirical commitment is there, the other differences can be negotiated in most instances. This is what David Morgan (2007) has called "Paradigms Lost and Pragmatism Regained."

Debating paradigms with one's clients, and taking sides in that debate, is different from debating one's colleagues about the nature of reality. I doubt that evaluators will ever reach consensus on the ultimate nature of reality. But the methodological paradigms debate can go on among evaluators without paralyzing the practice of practical evaluators who are trying to work responsively with primary stakeholders to get answers to relevant empirical questions. The belief that evaluators must be true to only one paradigm in any given situation underestimates the human capacity for handling ambiguity and duality, shifting flexibly between perspectives. In short, I'm suggesting that evaluators would do better to worry about understanding and being sensitive to the worldviews and evaluation needs of their clients than to

maintain allegiance to or work within only one perspective.

Beyond Paradigm Orthodoxies: A Paradigm of Choices

The paradigms debate elucidates the complexity of choices available in evaluation. It also demonstrates the difficulty of moving beyond narrow disciplinary training to make decisions based on utility. It is premature to characterize the practice of evaluation as completely flexible and focused on methodological appropriateness rather than disciplinary orthodoxy, but it is fair to say that the goals have shifted dramatically in that direction. The debate over which paradigm was the right path to truth has been replaced, at the level of methods, by *a paradigm of choices*.

Exhibit 12.8 summarizes the contrasting themes of the paradigms debate and describes the synthesis that is emerging with the shift in emphasis from methodological orthodoxy to methodological appropriateness and utility. *Utilization-focused evaluation offers a paradigm of choices.* Today's evaluator must be sophisticated about matching research methods to the nuances of particular evaluation questions and the idiosyncrasies of specific decision-maker needs. The evaluator must have a large repertoire of research methods and techniques available to use on a variety of problems.

The utilization-focused evaluator works with intended users to include any and all data that will help shed light on evaluation questions, given constraints of resources and time. Such an evaluator is committed to research designs that are relevant, rigorous, understandable, and able to produce useful results that are valid, reliable, and believable. The *paradigm of choices* recognizes that different methods are appropriate for different situations and purposes.

Classic Advice on Methodological Pluralism

The evaluator will be wise *not* to declare allegiance to either a quantitative-scientific-summative methodology or a qualitative-naturalistic-descriptive methodology.

SOURCE: *Toward Reform of Program Evaluation,* Stanford Evaluation Consortium Cronbach et al. (1980:7).

What Are Appropriate Standards of Evidence? Different Methods Yield Different Findings

It would be easy to conclude from this review of the paradigms debate that mixed methods are the solution. But the fact is that different methods often yield conflicting results. When that happens, which findings take priority? Which pass muster? Relaxing the gold standard on the front-end of design by incorporating mixed methods may just mean that it reappears on the back end when divergent and conflicting results have to be interpreted. Consider this cautionary tale.

For over a decade, the David and Lucile Packard Foundation made grants to test the effectiveness of a home visitation approach to parents with young children. Educators and health professionals visited parents in their homes to educate them about appropriately interacting with their children to enhance learning. While in the home they did developmental screening for children to look for problems that might need attention. Home visitation is an early intervention program to prevent child abuse or neglect, identify potential developmental needs of children in high-risk groups, and enhance school readiness. Through the Foundation's Center for the

EXHIBIT 12.8

The Evaluation Methods Paradigms Debate
Summary of Emphases: Thesis, Antithesis, Synthesis

	Thesis: Originally Dominant Social Science Research Paradigm	*Antithesis: Original Alternative Paradigm*	*Synthesis: Utilization-Focused Evaluation Paradigm of Choices*
Purpose	Summative	Formative	Intended use for intended users
Measurement	Quantitative data	Qualitative data	Appropriate, credible, useful data
Design	Experimental designs	Naturalistic inquiry	Creative, practical, situationally responsive designs
Researcher stance	Objectivity	Subjectivity	Fairness and balance
Conceptualization	Independent and dependent variables	Holistic interdependent system	Stakeholder questions and issues
Relationships	Distance, detachment	Closeness, engagement	Collaboration, consultative
Approach to study of change	Pre-post measures, time series, static portrayals at discrete points in time	Process-oriented, evolving, capturing ongoing dynamism	Developmental, action oriented. What needs to be known to get program from where it is to where it wants to be?
Relationship to prior knowledge	Confirmatory, hypothesis testing	Exploratory, hypothesis generating	Either or both
Sampling	Random, probabilistic	Purposeful, key informants	Combinations, depending on what information is needed
Primary approach to variations	Quantitative differences on uniform, standardized variables	Qualitative differences, uniquenesses	Flexible: focus on comparison most relevant to intended users and evaluation questions
Analysis	Descriptive and inferential statistics	Case studies, content and pattern analysis	Answer to stakeholders' questions
Types of statements	Generalizations	Context bound	Extrapolations, lessons learned
Contribution to theory	Validating theoretical propositions from scientific literature	Grounded theory derived from the situation	Describing, exploring, and testing stakeholders' and program's theory of action
Goals	Truth, scientific acceptance	Understanding, perspective	Utility, relevance: meaningful and useful to intended users

Future of Children publishes a journal called *The Future of Children* with a distribution of 40,000 to 50,000 copies and a $1 million cost per issue. It can take 18 months to develop a single issue. The journal has become prestigious as a credible source of information (Sherwood 2005).

In 1999 a special issue of *The Future of Children* was planned entitled "Home Visiting: Recent Program Evaluations." Both the Packard Foundation's evaluations using randomized trials and evaluations of others were screened for inclusion. As findings from several home visiting experiments were reported from 1996 to 1998 period, a pattern of mixed or no significant effects became evident. The Foundation brought together a group of evaluators, program directors, and independent experts to review findings for the special publication. Controversy centered on what kind of evidence would be reported since there were evaluation examples that employed experimental designs, quasi-experimental designs, mixed methods, and qualitative methods—and *evaluations with different methods were yielding different results*. What were appropriate standards of evidence?

The view of the Foundation staff was that *only* the main effects of randomized trials should qualify for publication. Those findings were largely negative, showing no or very small effects. Analysis of effects for subgroups was more positive, as was qualitative evidence, case studies, and some quasi-experimental evaluations. Indeed, the results from the randomized trials were consistently less positive than the results from the quasi-experimental studies (e.g., matched samples instead of random samples).

The implications of these findings would be far-reaching. Considerable contentiousness developed. Should *only* the results of randomized controlled experiments be considered credible and published? Should a variety of evaluations be published showing mixed and conflicting results? How should such different findings from varying methods be interpreted? Answering that question takes us from the paradigm debate about design into interpreting and reporting findings, the subject of the next chapter. I'll open that chapter by telling you what the journal published and the reactions to the decision they made about how to handle the varying evidence that came from different methods—a quite common result of mixed methods evaluations.

Live by the Evidence, Die by the Evidence

As you ponder the preceding cliffhanger, let me close this chapter with one more cautionary tale, a true story. A major philanthropic foundation invited a distinguished and well-known evaluation methodologist to conduct a day of training on methods. He was a powerful and insistent advocate of randomized controlled experiments as the only evidence worth having if the Foundation was going to have an impact on policy, which was its aspiration. He emphasized that only the findings from experiments were sufficiently credible to be useful and technically respectable.

The day after the training, the staff gathered to debrief what they had learned. The senior staff member who had arranged the training opened by apologizing. This surprised the group.

"Well, he was not the greatest presenter, it's true," replied a junior staff member, "but he's an academic. At least he knew his stuff."

"But he doesn't live by his own advice. As far as I'm concerned, he had no credibility after the first break, and it got worse throughout the day."

The whole group looked at her in stunned silence, completely surprised by this strong reaction. Finally someone asked, "What'd he say that so turned you off?"

"It wasn't what he said. It's what he did. On every break he dashed outside to smoke a couple of cigarettes. Live by the evidence, die by the evidence. We won't be having him back again."

Follow-Up Exercises

1. Locate a Web site for an organization that funds evaluations. This can be an international agency, federal or state government, philanthropic foundation, or independent research institution. Find where it discusses its approach to evaluation. Use the paradigm dimensions discussed in this chapter to characterize the evaluation methods being advocated. What is the paradigm perspective, either explicit or implicit, in this approach? Give concrete examples to support your judgment.

2. Locate an evaluation that used mixed methods, both quantitative and qualitative data. To what extent and in what ways were the data synthesized and integrated, or did they involve separate and distinct evaluation questions? Discuss the differences and complementarities of the two kinds of data.

3. What is your opinion about the methodological gold standard issue? Should there be a methodological gold standard? If so, what—and why? If not, why not? What is illuminating and what is distorting about the "gold standard" metaphor applied to evaluation methods?

4. Identify a program and one or more evaluation questions for that program. Provide an overview of a design that is (a) entirely quantitative/experimental, (b) entirely qualitative/naturalistic, and (c) mixed methods. Offer these three alternatives in the form of a memo written to primary intended users of the proposed evaluation. Identify strengths and weaknesses of each approach.

5. Assess your own methodological strengths and weaknesses. What methods are you most knowledgeable about and comfortable with? Why? In what evaluation methods do you lack training and expertise? Discuss how your competences and training affect your capability to match methods to the nature of the evaluation questions. To what extent can you be methodologically flexible and eclectic? Do a capacity assessment. For assistance, see "A Professional Development Unit for Reflecting on Program Evaluation Competencies" (Ghere, King, Stevahn, and Minnema 2006).