12

Simulate Use of Findings

Evaluation's Equivalent of a Dress Rehearsal

The purpose and intended uses of the evaluation have been established with primary intended users. Priority evaluation questions have been identified. Design, measurement, and data collection decisions have been made. Time to gather data? Not quite yet. First, do a final check to make sure that the data to be collected will really answer priority questions and support action. How? Well, one way is to do a utilization-focused dress rehearsal: Engage intended users in a simulation exercise in which the evaluator fabricates some potential results and intended users practice interpreting and using those results. The evaluation facilitator asks, "If the results came out this way, what would you do? If the findings came out this other way, what would that tell you, and what actions would you take? Given what you want the evaluation to accomplish, have we focused on the right outcomes and most useful indicators?"

Athletes, performing artists, astronauts, and entertainers spend hundreds of hours practicing for events that take only a few hours. Firefighters do fire drills. Actors do dress rehearsals. Is it too much to ask intended users to spend a couple of hours practicing to get mentally and analytically ready for the climax of an evaluation? Here's how it works.

Simulated Data Interpretation Scenarios

The stage can be set for analysis, interpretation, and use *before* data are ever collected. Once instruments have been designed—but before data collection—conduct a simulated use session. This involves fabricating possible results and interpreting the action implications

of the made-up data. The evaluator prepares some possible "positive" and "negative" findings on the most important issues. For example, suppose primary users have chosen the job placement rate as the priority outcome variable for a vocational training program. The evaluator might construct data showing a placement rate of 40% for Black participants and 75% for White participants. The evaluator facilitates analysis by asking such questions as the following: "What do these results mean? What actions would you take based on these results? How would you use these data?"

Such a discussion accomplishes four things:

1. The simulated analysis is a check on the design to make sure that all the relevant data for interpretation and use are going to be collected. (Remember this session occurs before actually gathering data.) All too often, at the analysis stage, *after* data collection, evaluators and stakeholders realize that they forgot to ask an important question.

2. The simulated use session trains and prepares stakeholders for the real analysis later. They learn how to interpret data and apply results by practicing doing so.

3. Working through a use scenario prior to data collection helps set realistic expectations about what the results will look like. Strengths and limitations of the design emerge. Methodological and measurement issues can be further clarified and discussed. This helps prepare users for the necessity of interpreting findings in relation to possible actions and likely ambiguities.

4. Use scenarios help build the commitment to use—or reveal the lack of such commitment.

When intended users are unable to deal with how they would use findings prior to data collection, a warning flag goes up that they may be unable, or unwilling, to use findings after data collection. The commitment to use can be cultivated by helping intended users think realistically and concretely about how findings might be applied before data collection gets under way. This relatively safe, even fun, exercise of analyzing simulated data can help strengthen the resolve to use before being confronted with real findings and decisions. This can help overcome resistance to evaluation and remove any remaining barriers to implementing the evaluation data collection.

Quantitative data are fairly easy to fabricate once instruments have been developed. With qualitative data, it's necessary to construct imagined quotations and case examples. This extra work can pay large dividends as decision makers develop a utilization-focused mindset based on an actual experience struggling with data.

An Ancient Example

I often set up the dress rehearsal with fabricated data by taking primary intended users through an ancient evaluation example found in the Old Testament of the Bible.

The example isn't aimed at any religious message, I explain. Instead it shows how the issue of using data to improve programs and inform decisions goes back a long way.

The book of Daniel tells the story of what happened when Nebuchadnezzar, king of Babylon, conquered Jerusalem. He instructed Ashpenaz, the master of his eunuchs, to identify intelligent young children of Israel from respected families who could be trained as civil servants. Throughout history it has been a common practice of conquering nations to establish colonial governments administered by selectively trained and supervised indigenous people. Among those selected by Ashpenaz for service to Nebuchadnezzar were Daniel, Hananiah, Mishael, and Azariah. But when they arrived at the king's palace to begin their training, Daniel and his friends refused to eat the king's meat or drink his wine. They asked to be allowed to eat a vegetarian diet and drink only water. Ashpenaz, the program director, resisted this request, fearing that such a diet would be unhealthy and the king would think he was not feeding the trainees appropriately. Indeed, the stakes were high for Ashpenaz, who responded to Daniel's request by saying: "I fear my lord the king, who hath appointed your meat and your drink. What if he sees you looking worse than the others in the program. *Then shall ye make me endanger my head to the king."*

Daniel wanted to assure Ashpenaz that the diet was healthy and they would not appear sickly if allowed to eat a vegetarian diet, so he proposed an experiment. He asked that he and his friends be allowed to eat the vegetarian diet while others in the program who preferred the meat and wine eat the king's food and see what happens. Daniel agreed that if they started to look unhealthy, they would eat the king's food, but Ashpenaz had to agree that if they appeared healthy at the end of the 10-day experiment, they would be allowed to continue their vegetarian and water diet.

As it turned out, the four vegetarians eating their kosher diet—Daniel, Hananiah, Mishael, and Azariah—not only didn't become sickly looking, "their countenances appeared fairer and fatter in flesh than all the children which did eat the king's meat." So, they were allowed to maintain their vegetarian diet, completed the program after 3 years, passed Nebuchadnezzar's test, and entered into the king's service. In essence, this is an example of a formative evaluation in which a group of program participants asked for a modification in the program, the modification was tried out, and, when the evidence showed it worked, the program was adapted accordingly. Indeed, one of the most common uses of formative evaluation is to adapt a program to the special needs of some subgroup because one size doesn't fit all.

So what does this have to do with a data use simulation?

In the story, all four participants in the vegetarian experimental group came out looking better. That makes the decision to allow them to continue their preferred diet easy. Likewise, if all four had looked worse, it would be clear that they would have to eat the king's diet. But what action should Ashpenaz take if two look better and two look worse? Or if one looks better, two look the same, and one looks worse? Or if three look the same but one looks worse? Discussing these scenarios prepares primary intended users for the possibility of having to deal with different kinds of results, including mixed or ambiguous

findings. It's a simple warm-up exercise before doing the real simulation. And the simulation with fabricated data is a warm-up dress rehearsal before being faced with taking action on real findings.

EARLY SIMULATION EFFORT

"Based on the simulated evaluation findings, who thinks we should ask the donor for more money?"

Standards of Desirability

A simulated use session also offers a prime opportunity to think about and formalize criteria for making judgments—again, *before data collection*. With quantitative data this can be done quite precisely by establishing standards of desirability. I like to have users set at least three levels of attainment:

1. Results level at which the program is considered *highly effective*,

2. Results level at which the program is considered *adequate*, and

3. Results level at which the program is considered *inadequate*.

Such standards can be established for implementation targets (e.g., program participation and completion rates) as well as outcomes (like getting and retaining a job). Suppose you are collecting satisfaction data on a workshop. At what level of satisfaction is the workshop a success? At what level is it merely adequate? At what level of participant satisfaction (or dissatisfaction) is the workshop to be judged ineffective? It's better to establish these kinds of standards of desirability in a calm and deliberative manner *before* actual results are presented.

This exercise may also reveal that satisfaction data alone are an inadequate indicator of effectiveness, an important discovery while there's still time to measure additional outcomes. Let me elaborate.

Suppose the following performance target has been set: *75% of workshop participants will be satisfied*. This doesn't tell us what constitutes an outstanding accomplishment; it doesn't distinguish adequacy from excellence. Nor does it make it clear whether 65% satisfaction is inadequate or merely "lower than we hoped for but acceptable." In a dress rehearsal that targeted a 75% satisfaction rate, I fabricated data that showed only a 73% satisfaction rate. Did that mean the workshop was a failure? The workshop evaluation committee discussed this fabricated finding and decided that just having a single overall measure of satisfaction was not very useful. They wanted to know the characteristics of the people who were satisfied and dissatisfied (e.g., novices versus old-timers), and what were they satisfied or dissatisfied about (which would require an open-ended question). The data interpretation dress rehearsal led to a final revision of the evaluation form before data collection.

Sometimes objectives and performance targets have been established in a proposal or plan a long time before the program is under way or well before an actual evaluation has been designed. Reviewing objectives and targets, and establishing precise standards of desirability just before data collection, increases the likelihood that judgment criteria will be up to date, realistic, meaningful, and *actionable*.

During the early conceptual stage of an evaluation, questions of use are fairly general and responses may be vague. The evaluator asks, "What would you do if you had an answer to your evaluation question? How would you use evaluation findings?" These general questions help focus the evaluation, but once the context has been delineated, the priority questions focused, and methods selected, the evaluator can pose much more specific use questions based on what results might actually look like.

For example, if recidivism in a community corrections program is 55%, is that high or low? Does it mean the program was effective or ineffective? The program had some impact, but what level of impact is desirable? What level spells trouble? Thinking about how to interpret data in advance of getting actual results goes a long way in building the capacity of primary intended users to interpret real findings.

The key point is that if intended users are unwilling or unable to set expectancy levels *before data collection*, there is no reason to believe they can do so afterward. In addition, going through this process ahead of time alerts participants to any additional data they will need to make sense of and act on the results. Many of the most serious conflicts in evaluation are rooted in the failure to clearly specify standards of desirability ahead of data collection. This can lead both to collection of the wrong data and to intense disagreement about criteria for judging effectiveness. Without explicit criteria, data can be interpreted to mean almost anything about a program—or to mean nothing at all.

Preparing for Use

Another way of setting the stage for analysis and use is having stakeholders speculate about results prior to seeing the real data. This can be done prior to data collection or after data collection but prior to actual presentation of findings. Stakeholders are given an analysis table with all the appropriate categories but no actual data (a dummy table). They then fill in the missing data with their guesses of what the results will be. This kind of speculation prepares users for how the results will be formatted and increases interest by building a sense of anticipation. I've even had stakeholders establish a betting pool on the results. Each person puts in a dollar, and the person closest to the actual results on the major outcome wins the pot. That creates interest! And the winner must be present at the unveiling of the findings to win. Strange how attendance at the presentation of findings is increased under these conditions!

A second and more important function of having stakeholders write down their guesses is to provide a concrete basis for determining the extent to which actual results come close to expectations. Program staff members, for example, sometimes argue that they don't need formal evaluations because they know their clients, students, or program participants so well that evaluation findings would just confirm what they already know. I've found that when staff members commit their guesses to paper ahead of seeing actual results, the subsequent comparison often calls into question just how well some staff members know what is happening in the program. At least with written guesses on paper, program staff and other stakeholders can't just say, "That's what I expected." A baseline (in the form of their guesses) exists to document how much something new has been learned.

You can combine establishing standards of desirability and speculating on results. Give stakeholders a page with two columns. The first column asks them to specify what outcomes they consider desirable, and the second column asks them to guess what results they believe will be obtained. Having specified a standard of desirability and guessed at actual

results, users have a greater stake in and a framework for looking at the actual findings. When real results are presented, the evaluator facilitates a discussion on the implications of the data that fall below, at, or above the desired response, and why the actual findings were different from or the same as what they guessed. In facilitating this exercise, the outcomes data presented must be highly focused and limited to major issues. In my experience, animated interactions among users follow as they fully engage and interpret the results. I find that, given sufficient time and encouragement, stakeholders with virtually no methods or statistics training can readily identify the strengths, weaknesses, and implications of the findings. The trick is to move people from passive reception—from audience status—to active involvement and participation.

Preparing for Interpretation: Making Comparisons

In the second quarter (April–June) of 2010, the Dow Jones Industrial Average dropped 9.97%, or some 1,082 points. Reports in the financial news tried a variety of comparisons to place this decline in perspective:

- The worst slide for the April–June quarter since 2002.
- Just 2 percentage points worse than the average for all 17 quarterly losses over the past 10 years.
- One of just five second-quarter losses over the past two decades.
- The first quarterly loss after four rising quarters that showed gains of 4.82% to 15.82%.
- One of only eight quarterly losses versus 12 quarterly gains in five years.
- Only the sixth-worst quarterly loss in 10 years.

All of these comparisons are accurate. Each provides a different perspective. Which ones are more or less useful depends on the situation of the investor. Those who have been investing for 10 years or more may value 10-year comparisons, or even data for two decades. Others may think that 5 years is the more illuminative time line. Some will prefer 1 year at a time. Still others will eschew quarterly results altogether and prefer monthly data or annual data. There is no right comparison. Offering intended users examples of different comparison possibilities before data are collected, analyzed, and presented helps them decide what comparisons will be most useful—and practice thinking comparatively.

Cost-Benefit Analysis Scenario

An interesting and illuminative scenario of decision options for cost-effectiveness analysis is presented in Exhibit 12.1. This table compares the costs and benefits of two programs,

EXHIBIT 12.1

Simulated Comparison of Cost-Benefit Comparisons of Two Program Options, A Versus B

	Program A lower cost than program B	*Program A higher cost than program B*
Program A higher benefit (better outcomes) than program B	Cell 1 Program A has better outcomes and lower costs: easy decision to decide in favor of program A over B.	Cell 2 Program A has higher benefits and higher costs. Decision question: Is the greater benefit worth the greater cost?
Program A lower benefit (worse outcomes) than program B	Cell 3 Program A has lower outcomes and lower costs. Decision issue: Is absolute cost a more important criterion than cost-benefit level?	Cell 4 Program A has worse outcomes and costs more: easy decision to reject program A in favor of B.

A versus B. When one program has both lower costs and higher benefits (better outcomes) as in cell 1, the decision is easy: favor that program. Likewise, when one program has higher costs and worse outcomes as in cell 4, reject that program. Those are the easy decisions and, in my experience, the rarer results. The tougher decision is what to do when one program has both higher costs and better outcomes as in cell 2. Is the greater benefit worth the greater cost? On the other hand, what's the appropriate decision when a program has lower costs and lower outcomes? Many decision makers are more interested in lower costs, even if that means lower outcomes, because, it turns out, cost trumps cost-benefit. Discussing these options and comparisons ahead of time makes it clear that interpretation and judgment will be needed to take action on findings.

Risk Assessment

As long as we're talking about cost-benefit analysis, let's acknowledge that designing an evaluation involves some kind of informal cost-benefit analysis in which potential benefits,

for example, using results to improve the program, are considered in relationship to costs, which include financial resources, evaluator and staff time, and opportunity costs. (What else could have been done with the money spent on evaluation?)

Introducing the notion of risk into evaluation design and intended users' decisions is a way of acknowledging that things seldom turn out exactly the way they are planned. We have many adages to remind us that human endeavors inherently involve risks: "Even the best laid plans . . . ," "Many a slip between cup and lip." And the bumper sticker: "Stuff Happens" (or a more emphatic version that replaces "stuff" with a certain nonacademic 4-letter word).

Explicitly introducing risk into conversations and negotiations between evaluators and primary intended users begins by asking the following kinds of questions:

1. What can go wrong in this evaluation?

2. What is the likelihood that certain things will go wrong?

3. What are the consequences and how bad would they be?

The intent of such front-end risk assessment is *not* to deepen the illusion that one can anticipate and thereby prevent all difficulties. Rather, it is to lay the foundation for contingency thinking as a basis for evaluator–user negotiations and revisions as the evaluation unfolds. Risk analysis should push evaluators and intended users to be prepared for contingencies. Contingency thinking and planning acknowledges the reality that every design will run into execution problems. What distinguishes one evaluation from another is not the absence of problems but the preparation for and ability to solve them. Examining what can go wrong should include thoughtful consideration of what can really be accomplished with available resources. Risk is traditionally defined as the probability of an occurrence multiplied by the severity of the consequences associated with the hazard. Risk analysis requires evaluators and stakeholders to become explicit about different scenarios and how they might behave in each. This can help mentally prepare intended users to be ready to engage around whatever emerges. Examples of common scenarios include:

- Not getting access to promised data because of confidentiality concerns or finding out that the data supposedly available was never actually collected, or is in such bad condition as to be unusable
- Problems getting access to key informants or program participants in a timely manner because of scheduling problems or program administrative difficulties that interfere with access and data collection
- Sudden crisis in a program when a key staff person leaves, or funding is lost, or a critical incident occurs (for example, accidental death of a participant or staff member)
- Partners in a collaborative initiative refusing to participate in an evaluation despite initial promises to do so because of fear about what will be discovered and reported

Bottom line: *Expect the unexpected. Be prepared for contingencies*. And prepare primary intended users to be ready to adapt the evaluation design in the face of what actually unfolds during fieldwork and data collection. As former heavyweight boxing champion Mike Tyson observed: "Every boxer has a plan . . . until he gets hit." STUFF HAPPENS.

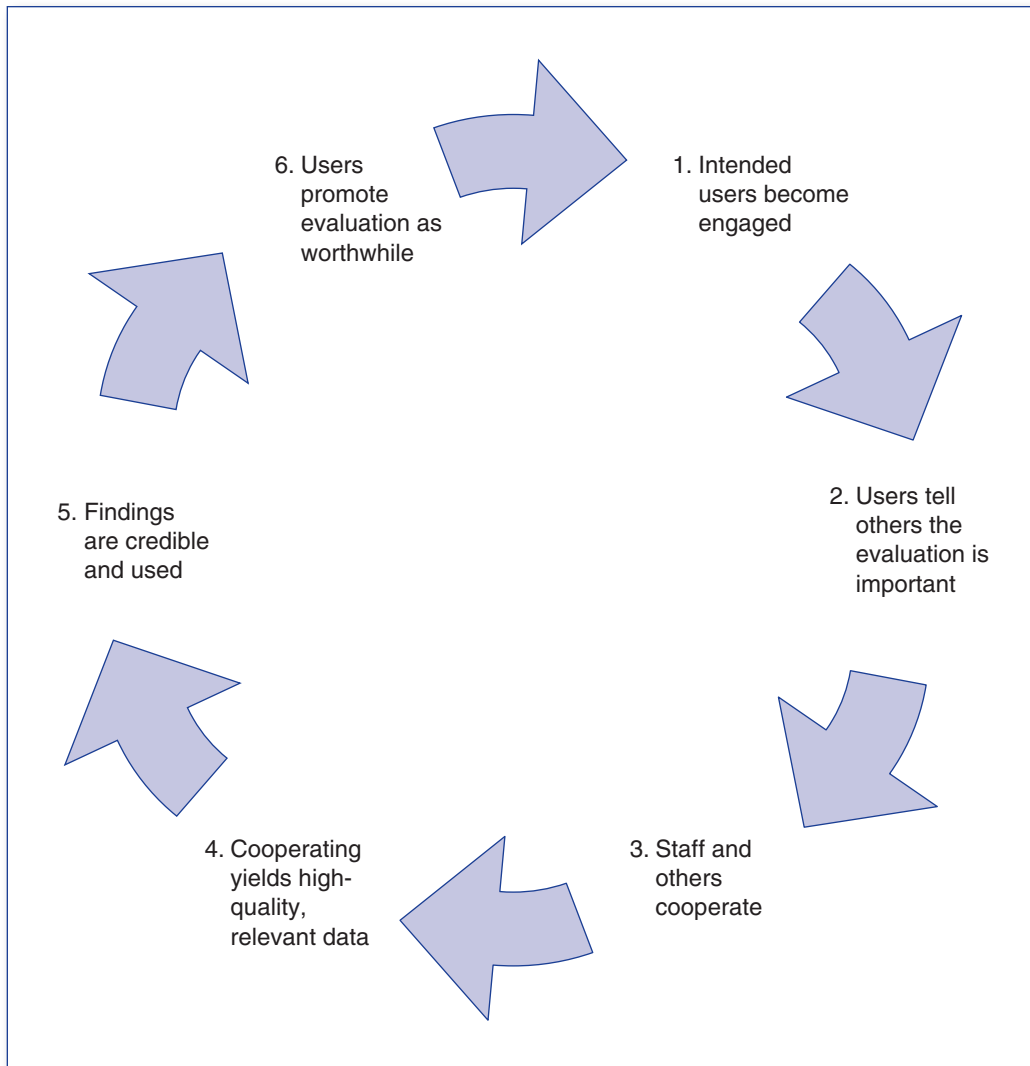## Virtuous and Vicious Utilization Circles

When actions lead to reactions that create still further reactions, a reinforcing systems dynamic is at work in which initial effects interact in what amounts to spiraling ripple effects. A positive spiral is called a virtuous system because it accelerates action in the desired direction. For example, in economics a tax cut for small companies leads to more investment which leads to higher productivity, which leads to higher profits and business expansion, which leads to new jobs, which leads to a healthier economy, which leads to higher tax revenues in the end. In contrast, a downward spiral is a vicious circle in which things get worse and worse: low-income people get hurt or sick, miss work, lose their job, can't pay their rent, get evicted, end up homeless and living on the streets, which makes it harder to get a new job. A virtuous cycle has favorable results and a vicious cycle has deleterious results. These cycles will continue in the direction of their momentum until some external factor intrudes and stops the cycle.

Evaluation use is subject to both vicious and virtuous spirals. A virtuous utilization circle is set in motion when primary intended users become interested and engaged in an evaluation; they communicate to others, like program staff and participants, that the evaluation is genuine and meaningful; staff and participants cooperate with the evaluation; relevant, high-quality data are generated through cooperation; high-quality data increase the credibility of the evaluation, which enhances use; primary intended users feel good about having participated in the evaluation, see and report to others that it was useful, which creates a positive and receptive environment for future evaluations.

A vicious evaluation circle emerges where primary intended users distrust the evaluation process, bad-mouth the evaluation to others creating resistance from staff and program participants, which undermines cooperation with the evaluation, which leads to poor data and weak findings, which undermines the credibility of the findings and makes the evaluation useless, which confirms and reinforces the intended users' negative suspicions and skepticism about evaluation.
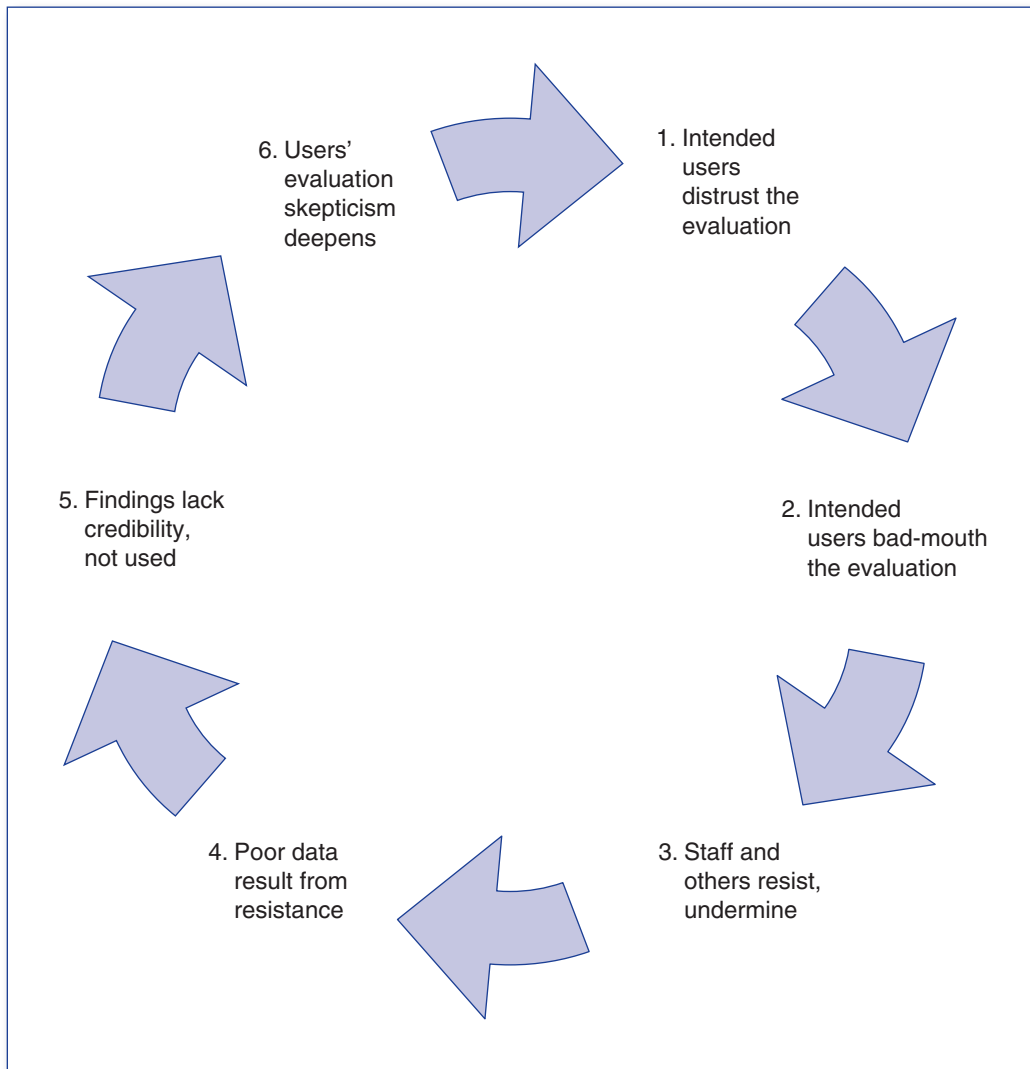
Situation analysis and scenario planning can include creating scenarios with primary intended users about how virtuous use circles can be created and vicious evaluation circles avoided. Exhibit 12.2 depicts a virtuous utilization system dynamic. Exhibit 12.3 depicts a vicious utilization system dynamic. Evaluation risk assessment and stakeholder contingency planning can

EXHIBIT 12.2

A Virtuous Utilization Circle

1. Intended users become engaged

2. Users tell others the evaluation is important

3. Staff and others cooperate

4. Cooperating yields high-quality, relevant data

5. Findings are credible and used

6. Users promote evaluation as worthwhile

EXHIBIT 12.3

Vicious Circle Undermining Evaluation Use

6. Users' evaluation skepticism deepens

1. Intended users distrust the evaluation

2. Intended users bad-mouth the evaluation

5. Findings lack credibility, not used

3. Staff and others resist, undermine

4. Poor data result from resistance

include what can occur within the specific context in which the evaluation will be conducted that would lead to a virtuous or vicious circle—and what would be early indicators of each.

Using Scenarios to Study Stakeholder Engagement and Evaluation Use

This chapter has advocated creating and using scenarios with primary intended users as a final form of engagement before data collection. These scenarios provide a final check that the right data will be collected to address priority issues and help train intended users for interpreting results when the findings have been generated. What does research show about the capacity of nonresearchers to participate in and learn from such scenario exercises?

Christina Christie (2007), a distinguished UCLA evaluation researcher, 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. I reported her findings in an earlier chapter, but they're worth reiterating here. 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 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. Engaging concrete scenarios and considering trade-offs helped decision makers explicitly express their data preferences and interpretive frameworks.

The Ultimate Question

Step 12, simulating use, is the penultimate step before data collection. This next-to-last step before finalizing the evaluation design, methods, and measures sets up the ultimate question before data collection: *Given expected costs and intended uses, is the evaluation worth doing?* The simulation offers a *dress rehearsal* opportunity, a chance to practice use. Are the primary intended users committed to moving forward? Is the evaluation worth doing? I like to formalize this decision with primary intended users as the final step before data collection.

> Details about what is involved in each step are provided in the summary *U-FE Checklist* in the concluding chapter. See page 418 for the checklist items for Step 12 discussed in this chapter.

PRACTICE EXERCISES

1. Using evaluation data or an indicator of any kind, fabricate alternative findings and identify the facilitation questions you would ask for a simulation with primary intended users.

2. Select a metaphor for an evaluation simulation (for example fire drill, dress rehearsal, or something relevant to your world) and discuss the advantages and disadvantages of that metaphor applied to introducing an evaluation simulation to primary intended users.

3. Comedian Mitch Hedberg (1968–2005) did a stand-up routine based on the following scenario:

> I bought a doughnut and they gave me a receipt for the doughnut. I don't need a receipt for the doughnut. I give you money and you give me the doughnut, end of transaction. We don't need to bring ink and paper into this. *I can't imagine a scenario that I would have to prove that I bought a doughnut.* When, perhaps some skeptical friend questions my purchase, I can say, "Don't even act like I didn't get that doughnut, I've got the documentation right here. It's in my file at home, filed under 'D.'"

Questions: Under what scenarios would a customer want such a receipt? What purposes beyond a customer's needs does a receipt serve? How would you use this example with primary intended users? To illustrate what?

13

Gather Data With Ongoing Attention to Use

If there is a 50–50 chance that something can go wrong, then 9 times out of 10 it will.

Paul Harvey,
American radio broadcaster

The previous chapter included a risk assessment of what *might* go wrong during data collection. Now, during data collection is when you find out what *will* go wrong—and what will go as planned. You've identified and engaged primary intended users. You've worked with them to be clear about the evaluation's purpose, prioritizing relevant questions and selecting appropriate methods. You've facilitated a practice session using fabricated findings to deepen the commitment to use and enhance their capacity to move from data to decision and action. Now comes the work of implementing the evaluation plan and gathering data. A marathoner trains for months to get ready for a big race, but the race still has to be run. An acting troupe rehearses for weeks to get ready for live performances with a real audience, then opening night arrives at last. Politicians campaign for months to get elected, then they still have to govern. Utilization-focused evaluators work diligently to engage primary intended users to design a relevant and meaningful evaluation, but the data still have to be collected. This chapter looks at how to gather data with ongoing attention to use. We'll consider four issues that are particularly likely to affect use: (1) effective management of the data collection process, (2) implementing any agreed-on participatory

approaches to data collection that build capacity and support process use, (3) keeping primary intended users informed about how things are going, and (4) providing early feedback by reporting emergent and interim findings.

Effective Management of the Data Collection Process

Management is, above all, a practice where art, science, and craft meet.

Henry Mintzberg,
management scholar and consultant

Step 2 focused on assessing evaluator readiness. In that chapter I reviewed essential competencies for evaluators (Ghere, King, Stevahn, & Minnema, 2006), among which is *project management*—the nuts and bolts of managing an evaluation from beginning to end. But managing data collection is especially important, for, as much as anything, this can determine whether the evaluation findings have integrity and credibility and are delivered in a timely manner. Some brilliant methodologists are lousy managers, and vice versa. Exhibit 13.1 identifies some critical checklist factors that affect competent evaluation management.

Six issues are especially critical in project management:

1. Managing funds to deliver what has been promised within budget

2. Managing time to deliver findings in time to be useful and actually used

3. Managing people, especially members of the evaluation team, so that everyone does his or her job well

4. Managing data collection so that the data have integrity and credibility

5. Managing relationships with people who provide data (program participants, program staff, administrators, program information systems personnel) so that they feel treated with respect, cooperate with the evaluation, and form and communicate a positive impression of how the evaluation is conducted

6. Managing the unexpected so that difficulties get dealt with early before they become a crisis that seriously damages the evaluation. Something will go wrong. Unexpected challenges will emerge. Stay close enough to the action to spot these problems sooner rather than later—and take action accordingly.

The basic message is that it's not enough to have an excellent, well-conceived, and agreed-on evaluation plan. That plan has to be implemented. Just as failures in program implementation often undermine good program ideas, failures in evaluation execution can undermine a well-conceived, utilization-focused evaluation design. "Execution Trumps

EXHIBIT 13.1

Critical Evaluation Management Checklist Factors

❑ Delivery schedule: What is the schedule of evaluation services and products?

❑ Access to data: What existing data may the evaluators use, and what new data may they obtain?

❑ Responsibility and authority: Have the system personnel and evaluators agreed on what persons and groups have both the responsibility and authority to perform the various evaluation tasks?

❑ Budget: What is the structure of the budget? Is it sufficient but reasonable, and how will it be monitored?

❑ Allocation of resources: Have the resources for the evaluation been appropriately distributed across data collection, analysis, and reporting, placing the most effort on the most important information requirements?

❑ Finances: What is the schedule of payments for the evaluation, and who will provide the funds?

❑ Data gathering: How will the data-gathering plan be implemented, and who will gather the data?

❑ Data storage and retrieval: What format, procedures, and facilities will be used to store and retrieve the data?

❑ Attention to trade-offs: How will the evaluation address trade-offs between comprehensiveness and selectivity in collecting, organizing, analyzing, interpreting, and reporting information?

❑ Work management: What oversight and control will be administered to assure that evaluators devote time and effort, as well as their reputations, to the evaluation?

❑ Facilities: What space, equipment, and materials will be available to support the evaluation?

❑ Data-gathering schedule: What instruments will be administered, to what groups, according to what schedule?

❑ Maintaining focus: Are there sufficient safeguards to prevent gathering extraneous information?

❑ Reporting schedule: What reports will be provided, to what audiences, according to what schedule?

❑ Realistic expectations: Have appropriate steps been taken to help stakeholders develop realistic expectations considering available financial, time, and personnel resources?

SOURCE: Based on *Evaluation Plans and Operations Checklist* (Stufflebeam, 1999a). For the full checklist: http://www.wmich.edu/evalctr/wp-content/uploads/2010/05/plans_operations1.pdf

Strategy, Every Time" was the title of a speech given to the philanthropic Evaluation Roundtable by Dr. Steven Schroeder (2002), former president of the Robert Wood Johnson Foundation as he recounted his efforts to eradicate tobacco use in the United States. Implementing good ideas had proved more difficult than generating good ideas, he recounted.

For many years, the importance of strategic planning and strategy development received primary emphasis in the business world, but recent books have emphasized the importance of execution, as in these best-selling titles:

Execution: The Discipline of Getting Things Done (Bossidy & Charan, 2002)

Execution Premium: Linking Strategy to Operations for Competitive Advantage (Kaplan & Norton, 2008)

Execution Revolution: Solving the One Business Problem That Makes Solving All Other Problems Easier (Harpst, 2008)

The general principle, then, is to manage evaluation data collection thoughtfully and diligently to ensure high-quality findings. The specific utilization-focused principle is to manage the process and solve inevitable problems with an eye to how such problem solving affects use. For example, adapting the design to the realities of fieldwork can change what primary intended users expect the findings to include. There may be problems getting access to certain records or key people. Scheduling problems and delays in administering questionnaires may arise. There may be missing data or even lost data. In all of these cases, decisions will have to be made to adapt the evaluation plan to what actually happens in the field. I hasten to add that such adaptations are par for the course. In 40 years I can't remember any evaluation plan that has unfolded exactly as planned. The point is to monitor how things are unfolding, get on top of problems early—and *keep primary intended users informed of what's happening,* what adjustments are being made, and why.

Managing Participatory Approaches to Data Collection

Managing data collection controlled by an evaluator or research team is one thing. Managing participatory approaches to data collection is a whole other ball game, a completely different kettle of fish, a horse of a different color, a . . .—pick any metaphor that says to you, this is a different matter entirely. Because it is.

Step 6 discussed process use as an evaluation option. Process use involves conducting the evaluation so as to build evaluation capacity among those involved, support active engagement at every stage of the process, including data collection and analysis, and help people learn about evaluation by doing evaluation. When participatory approaches to data collection are involved, managing the process extends beyond focusing on high-quality data collection to facilitating the active involvement of nonresearchers in data collection to support high-quality learning. Here are some examples.

- Participants in a leadership development program do round-robin interviews with each other as part of data collection. Bill interviews Sue, who interviews Abdul, who interviews Kaiya, who interviews Danielle, who interviews Jose, and so forth, back to Bill. The evaluator

works with the participants to construct the interview, offers some interview training, creates a protocol for writing case studies based on the interviews, and monitors and supports the process, but actual data collection and writing the case studies is done by the program participants.

- Program staff in an adult education program for recent immigrants engages in monthly reflective practice sessions in which they report mini-case examples of their experiences with participants. The evaluator is a coach and co-facilitator of the reflective practice process, but participants take primary responsibility for choosing focused topics for reflection, capturing the stories, and identifying key themes. (For details on reflective practice as a participatory evaluation process, see Patton, 2011, pp. 265–275.)

- Young people in a youth engagement program help design a questionnaire to assess needs in their community. They go door-to-door in pairs to administer the questionnaire to residents. The evaluator helps with questionnaire design and trains the young people in how to collect the data. This gives them a direct experience with data collection, reduces the costs of getting data, and deepens their sense of ownership of the findings. (For examples of youth participatory evaluation see Campbell-Patton & Patton, 2010; Flores, 2003, 2007; McCabe & Horsley, 2008; Sherrod, Torney-Purta, & Flanagan, 2010.)

- Villagers in Bogui on the Ouagadougou-Niamey road 10 kilometers east of Fada'Gourma in Burkina Faso keep track of children sick with diarrhea using match sticks. Half of the villagers are using water from a new well while the other half use water as they have done traditionally from a nearby pond. The villagers participate in monitoring differences in illnesses and deciding if any measured differences are significant. This is an example of *participatory learning for responsive development evaluation* (Salmen & Kane, 2006).

- Farmers in India participate in data collection in an agricultural development initiative by keeping *seasonal calendars* (Bamberger, Rugh, & Mabry, 2006, pp. 100, 102). On a monthly basis farmers track and report on indicators like rainfall, what was planted, any chemicals used, what was harvested, labor used, water sources, market prices for agricultural products, and any diseases experienced in crops. A chart is drawn on paper or on the ground marking the months. Participants are then asked to place stones or seeds to indicate the months with, for example, the highest incidence of food shortages, sales of crops, and insect damage in storage. With facilitated discussion led by the evaluator, farmers learn to analyze patterns that affect productivity and use the information with agricultural extension staff to plan future improvements in the development initiative.

- Welfare recipients in a welfare-to-work program keep journals of critical incidents they experience in their job searches. Their journal entries are shared in program support sessions and also provide evaluation data about their experiences and the results of their job searches. The evaluators provide training in journaling and provide encouragement and support to participants to maintain the journals. The evaluators also help facilitate conversations about patterns across the diverse journal entries as part of the participatory evaluation learning experience.

As these examples hopefully make clear, the management and facilitation tasks involved in participatory approaches to data collection are significantly different from those in

which evaluators and researchers collect all the data. Australian action researcher Yoland Wadsworth is one of the world's most experienced authorities on participatory research and evaluation processes. She pioneered participatory evaluation circles of inquiry among residents and staff in a psychiatric hospital in an effort that extended over several years. The mammoth nature of the work involved is merely hinted at in the characteristics of the participatory model that she developed and implemented with the program staff and consumers of mental health services. These elements included:

1. A quality assurance/quality improvement framework that all agreed to use.

2. Two-way staff–consumer communication and dialogue about what was happening in the mental health facility instead of just one-way consumer feedback.

3. Using numerous methods, in many sites, with many elements of resourcing, involving many staff, across the whole organizational network and bureaucracy instead of just producing a satisfaction survey.

4. Using *multiple* consumer feedback and communication methods: for example, exit interviews as patients left and case study narratives of service usage.

5. Doing data collection with consumer and staff participation in three kinds of sites: decision-making sites, consumer-only sites, and staff–consumer dialogue sites.

6. Including data collection and dialogue at the "hardest' spots" (the acute unit and other high-risk situations) instead of focusing only on the "easier" community-based areas.

7. Providing support resources infrastructure (personal peer supports: e.g., pair teaming), check-ins and check-outs, incident debriefing, mentoring, networking, and discretionary funds for additional flexible and responsive ad hoc costs for data collection and engagement (Wadsworth, 2010, pp. 211–212).

The overall participatory effort was aimed at systemic, holistic culture change toward regular staff–consumer collaboration around seeking and acting together on evaluative feedback. Given the comprehensive nature of both the change process and multiple data collection dimensions, effective project management was critical to maintain momentum, demonstrate utility, and keep it all from becoming overwhelming to all involved.

Undergirding this kind of participatory data collection is an astute and realistic assessment of what participants can handle. Engage participants too little and the effort may reek of tokenism. Involve them too much and the data collection can overwhelm them or interfere with program participation. Finding the appropriate and useful level of participation in data collection, and then managing that participation effectively and efficiently, are the central challenges of participatory evaluation approaches. Done well, participatory evaluation enhances use, deepens learning, and builds organizational capacity

for ongoing evaluation. Done poorly, it can annoy all involved. As writer Paul Dickson (2010, p. 1) has advised:

Never try to teach a pig to sing;

it wastes your time and it annoys the pig.

"The sound effects, choreography and costumes were spectacular! I can't wait for the final de-briefing."

Keeping Primary Intended Users Informed About How Things Are Going

Research studies on useful evaluations turn up this finding again and again: *Avoid surprises.*

- The Treasury Board of Canada (2002) reviewed 15 major Canadian evaluations identifying "drivers of effective evaluations" that "were felt, by both the evaluation staff and the program staff, to have contributed significantly to making the evaluations useful and worthwhile." One of their recommendations: *Ensure that there are no last-minute surprises.*
- The World Bank (2004) synthesized lessons from influential evaluations that improved performance and impacts of development programs. One lesson was: *Keep key stakeholders informed*

of the progress of the evaluation. "There should be no surprises when evaluation findings are presented" (p. 22). Related lessons included that "key results must often be communicated informally before the final report is completed," "findings must be delivered in time to affect decisions," "a successful evaluation must adapt to the context within which it will be used," and "the evaluator must understand when and how the findings can most effectively be used" (p. 22).

• The longer the evaluation, the more important it is to work to keep key stakeholders engaged. The 2009 Outstanding Evaluation Award from the American Evaluation Association went to a 10-year experimental abstinence education evaluation. The evaluation included a technical workgroup of key stakeholders. Ongoing attention to stakeholder relations was deemed of critical importance to the evaluation's credibility and use, both because of the length of the evaluation and its controversial subject matter. The evaluators concluded that "making sure that you constantly engage them is important. . . . We gave them briefings and we gave them preliminary results. When we released the report, we used quotations from some of them" (Brandon, Smith, Trenholm, & Devaney, 2010, p. 526).

In our study of the use of federal health evaluations (Patton, 2008), we asked about how surprises affected use. We found that minor surprises on peripheral questions created only minor problems, but major surprises on central questions were unwelcome. One decision maker we interviewed made the point that a "good" evaluation process should build in feedback mechanisms to primary users.

> Evaluation isn't a birthday party, so people aren't looking for surprises. If you're coming up with data that are different than the conventional wisdom, a good evaluation effort, I would suggest, would get those ideas floated during the evaluation process so that when the final report comes out, they aren't a surprise. Now, you could come up with findings contrary to the conventional wisdom, but you ought to be sharing those ideas with the people being evaluated during the evaluation process and working on acceptance. If you present a surprise, it will tend to get rejected. See, we don't want surprises. We don't like surprises around here.

So, what kinds of things do key stakeholders need to be informed about during data collection? Exhibit 13.2 provides 10 examples. The trick is to keep them sufficiently informed to maintain interest and engagement without making the updates a burden.

Providing Feedback and Reporting Interim Findings

It is worth distinguishing feedback from interim findings. Feedback, in this context, is what you tell people from whom you're gathering data and program staff at the end of a site visit. Feedback can be as specific and timely as telling someone you're interviewing that their responses are helpful. Interviewees value such acknowledgement and feedback. They have no way of knowing if their responses are relevant and useful unless they are given feedback during the interview. This doesn't bias their subsequent responses. It makes them more

EXHIBIT 13.2

Ten Examples of Updates to Intended Users During Data Collection

1. Did you get access to important data? Evaluators seeking access to management information system data, government statistics, or program records often encounter unanticipated obstacles. Primary intended users should be informed of both successes and obstacles, and may be able to help with the latter.

2. How is the response rate unfolding? For example, surveys often involve an initial request with two follow-up requests aimed at increasing the response rate. The response rate will be key to the credibility of findings.

3. How is the time line for data collection working out? Evaluation designs typically include schedules for data collection. Delays in data collection can affect time lines and use. Let intended users know about delays sooner rather than later.

4. How are key informants responding to interviews? Key informants are typically knowledgeable and influential people whose opinions matter. In seeking their perspective on key evaluation issues, they will form a perspective about the evaluation itself. Interviewers can usually tell whether key informants are interested, helpful, resistant, suspicious, apathetic, or engaged—or all of these things during a long interview. On large projects with multiple interviewers, I follow up by e-mail with key informants to get their reactions to the interview process. If I'm the interviewer, I give them the e-mail address of the chair of the primary user evaluation task force or group and invite them to share reactions to the interview. This opportunity to play a quality control and feedback role keeps primary intended users engaged—and adds a meaningful layer of direct utility as they play this role.

5. How are focus groups unfolding? Focus groups can involve lots of logistics. Not everyone shows up. Some come late and others leave early. Incentives are often provided that create reactions. Unexpected discussions can occur when focus group participants come together. Key evaluation stakeholders appreciate getting tidbits about how focus groups are working.

6. How is the random assignment of target groups to experimental and control conditions working out? If a randomized control trial is the design of choice, smooth implementation is critical to the credibility of findings. Random assignment can be tough to implement. Once completed, the first test is whether the treatment and control groups are equivalent on important background variables. Share that information with primary intended users. It gives an early heads-up about how the design is unfolding.

7. What's the level of "experimental mortality"? Experimental mortality is the friendly term for losing people during an experiment. There are usually some people who drop out of program interventions as well as the control or comparison group. The issue for credibility is what proportion drop out and how they are distributed between treatment and comparison groups. Even in nonexperimental designs that use pre- and posttests, some people who complete

the pretest fail to complete the posttest. The "mortality" rate among respondents is an early indicator of the quality of the findings.

8. What's going as planned during data collection? Sometimes evaluators only communicate with intended users when there are problems and delays. Don't just communicate bad news. Hopefully—HOPEFULLY!—some things will actually unfold as planned. Successful implementation of the evaluation plan is not just success for the evaluator. It's also success for the primary intended users who were involved in developing the plan.

9. What early findings are emerging? Interim findings must clearly be labeled as interim and preliminary—and not appropriate for wider dissemination. But intended users typically welcome early results on key questions as they emerge. Be sure to follow up quickly as those preliminary findings are confirmed or revised.

10. When will findings be ready for discussion and sharing? Keep critical dates for engagement with the intended users before them. Key stakeholders are typically busy, juggling lots of commitments. Alerting them sufficiently in advance of a critical meeting or conference call about findings will help ensure their involvement in interpreting findings. This prepares them to be involved in supporting and facilitating use.

---

likely to be thoughtful in subsequent responses. Skilled interviewing involves offering appropriate and respectful feedback during the interview (Patton, 2002, chap. 7).

Site Visit Feedback

Feedback to program staff at the end of site visit, at completion of a set of interviews, or after program observations is another matter altogether. Evaluation induces anxiety. When program staff members know that participants are being interviewed or see an evaluator observing them, they are understandably anxious. Some evaluators believe that their independence and objectivity require maintaining distance. They do their site visits and tell program staff that they'll get a copy of the evaluation report when it is done. I consider such behavior disrespectful and insensitive, and it is easily experienced as arrogant. In contrast, aware of process use opportunities, a utilization-focused evaluator wants to create positive regard for evaluation, including among those who provide data and those who may feel they are the object of the evaluation (program staff and leadership). Thus, three kinds of feedback are appropriate, respectful and even useful.

1. Offer a reminder about the purpose of and time lines for the evaluation, with special emphasis on where the site visit just completed fits into the overall data collection of the evaluation; this information will have been communicated in setting up the site visit or interviews, but should be reiterated on site.

2. Explain the data collection details: sampling approach, questions being asked, design being used, and nature of the analysis to be done. Explain these features without jargon. Program participants and staff are not researchers. They won't readily know about how data are collected and analyzed. Be prepared to provide a straightforward explanation. For example, instead of just saying, "We interviewed a random probability sample" and watching eyes glaze over, say:

> We chose the participants we interviewed randomly, like drawing the names out of a hat, so we didn't pick people to interview for any special reason. We wanted to get a balanced picture from a variety of participants, so that's why we chose people randomly. Does that make sense? I'd be happy to answer questions about how we chose who to interview. We're using the same process in all the program sites where we're interviewing people.

3. Provide some specific feedback about the emergent findings, if possible and appropriate. This can be a chance to clarify something observed as well as provide a sense of what the data revealed. For example, after observing an employment program for 3 days and interviewing 20 participants, I met with the program director and senior staff to debrief the site visit. I reported:

> Participants were quite responsive during the interviews and seemed to appreciate the opportunity to tell their stories. Many expressed appreciation that this program exists and that the staff is supportive. They generally reported that they are learning new skills and hopeful about getting jobs, but are anxious about what kind of on-the-job follow-up help they will get from the program, if any. What is your relationship with participants after they graduate? Many of them seem unsure about this.

In this debriefing, I'm both providing a general sense of the findings and using the debrief to clarify an ambiguous issue that arose during the interviews.

Reporting Interim Findings to Primary Intended Users

Feedback to and debriefs with program staff are part of the process of data collection and help them understand the evaluation's purpose and focus as well as providing immediate information about what the evaluation is uncovering. Reporting interim findings to primary intended users has a different purpose: keeping them engaged and preparing for interpretation and use of the eventual findings. Collaborating with primary users means that evaluators should not wait until they have a highly polished final report to share some findings. Evaluators who prefer to work diligently in the solitude of their offices until they can spring a final report on a waiting world may find that the world has passed them by.

Different kinds of interim reports match different evaluation purposes and questions. Formative reporting will focus on potential areas of improvement and is most useful as part of a process of thinking about a program rather than as a one-time information dump. In the more formal environment of a major summative evaluation, the final report will involve high stakes so primary intended users will want early alerts about the likely implication of findings for a program's future. Accountability-focused evaluation will typically focus on whether implementation and outcome targets are being met and whether resources have been allocated and used appropriately. Here again, early alerts about potential problems and advance reports on the direction of findings will be appreciated by intended users. In contrast, developmental evaluation is designed for ongoing, timely feedback, and may not even produce a final report. Thus, interim reporting depends on the evaluation's purpose, time lines, and stakes. That said, the overall themes are keeping intended users interested and engaged—and avoiding surprises. Surprise attacks may make for good war strategy, but in evaluation, the surprise attack does little to add credence to a study. Here are three tips for interim reporting.

1. Emphasize in headlines the *interim* and *confidential* nature of the findings. Let intended users know that you're giving them an early look at some emerging results, but such findings are not ready to be shared widely and are far from definitive.

2. Keep interim reporting limited to one to three findings on important issues. Interim reports should be short and focused, no more than bullet points.

3. The longer the data collection period, the more important it is to find ways to keep intended users updated. Some evaluations are completed in 3 to 6 months; one interim report would suffice. In contrast, the abstinence education evaluation discussed earlier took 10 years to complete. Interim reports on how data collection was proceeding and baseline indicators would gradually yield, over time, to interim reports on overall results.

Gather Data With Ongoing Attention to Use

The basic message of this chapter is that strategizing about use continues throughout data collection. I've offered a few suggestions, examples, and ideas for ongoing engagement with primary intended users during data collection, but these are meant to be illuminative rather than definitive or exhaustive. Each evaluation is different. The overall principle is to analyze your situation and develop ways of maintaining momentum toward use. This includes (1) effective management of the data collection process to enhance the evaluation's credibility, (2) effectively implementing any agreed-on participatory approaches to data collection that

build capacity and support process use, (3) keeping primary intended users informed about how things are going so as to maintain interest and engagement, and (4) providing timely feedback and reporting emergent and interim findings.

Parting Shot

One final point: Be attentive to turnover among primary intended users. The longer data collection takes, the more likely it becomes that one or more intended users may move on to other things. When turnover occurs, don't delay connecting with any replacement intended user or key stakeholder. Bring them up to date. Get their buy-in. Connect them with the continuing intended users. Make this a priority. The primary intended users are the pathway to use.

> Checklist details about what is involved in each step are provided in the summary *U-FE Checklist* in the concluding chapter. See pages 418–420 for the checklist items for Step 13 discussed in this chapter.

PRACTICE EXERCISES

1. Assess your project management skills. What are you good at? What are your management weaknesses? Make a plan to build on your strengths and improve in areas of weakness. Remember: Research design and data collection skills are not enough. Effective, timely project management can make or break an evaluation.

2. Create a classic evaluation scenario in which you are conducting a 3-year evaluation with 1 ½ years of formative evaluation followed by 1 ½ years of summative evaluation. Describe the program and primary intended users. Generate an interim feedback and reporting plan for both the formative and summative evaluation. Discuss and explain the similarities and differences.

3. Describe an evaluation in which you're doing in-depth interviews with participants in a program. Identify several major open-ended interview questions you would ask. Now, write out a script for how you provide appropriate feedback to the interviewee during the interview. Give three examples of what you consider appropriate feedback. Also give three examples of what you would consider inappropriate feedback, comments that might bias the interview, make the interviewee uncomfortable, or otherwise impede the quality of the data collection process.

14

Organize and Present the Data for Interpretation and Use by Primary Intended Users

Analysis, Interpretation, Judgment, and Recommendations

When it comes to evidence, what is believable to one analyst is incredible to another. Evidence may be hard or soft, conflicting or incontrovertible, it may be unpersuasive or convincing, exculpatory or damning, but with whatever qualifier it is presented, the noun evidence is neutral: it means "a means of determining whether an assertion is truthful or an allegation is a fact."

William Safire,
political linguist and New York Times *columnist (2006, p. 18)*

A Framework for Engaging Findings

What? What are the findings? What do the data say?

So what? What do the findings mean? Making interpretations and judgments

Now what? Action implications and recommendations

Four distinct processes are involved in making sense out of evaluation findings:

(1) *Analysis* involves organizing raw data into an understandable form that reveals basic patterns and constitutes the evaluation's empirical findings, thereby answering the *what?* question.

(2) *Interpretation* involves determining the significance of and explanations for the findings; this is part one of the *so what?* question.

(3) *Judgment* brings values to bear to determine merit, worth, and significance, including the extent to which the results are positive or negative; this is part two of the *so what?* question.

(4) *Recommendations* involve determining the action implications of the findings. This means answering the *so what?* question.

Primary intended users should be actively involved in all four of these processes so that they fully understand and buy into the findings and their implications. Facilitating these processes involves helping intended users understand these four fundamental distinctions. Exhibit 14.1 summarizes this framework. We'll now consider each of these processes in greater depth.

EXHIBIT 14.1

A Utilization-Focused Framework for Engaging Findings

Four distinct processes are involved in helping primary intended users make sense out of evaluation findings. This involves answering the basic questions: What? So what? Now what?

1. *What was found?* Present basic findings. This involves description and analysis, essentially organizing raw data, both quantitative and qualitative, into a form that reveals basic patterns so that primary intended users can understand the findings and make sense of the evaluation evidence generated by the data.

2. *So what does it mean?* This involves making *interpretations*. The evaluator facilitates interpreting the findings with primary intended users. Help them ask: What do the results mean? What's the significance of the findings? Why did the findings turn out this way? What are possible explanations of the results? Interpretations *go beyond the data* to add context, determine meaning, and tease out substantive significance.

3. Add *judgments* about what the findings mean: *So what?* (part 2). Values are added to analysis and interpretations to make judgments. Judging merit, worth, and significance means determining the extent to which results are positive or negative. What is good or bad, desirable or undesirable, in the findings? To what extent have standards of desirability been met?

4. *Recommendations: Now what?* The final step (if agreed to be undertaken) adds action to analysis, interpretation, and judgment. What should be done? What are the action implications of the findings? Only recommendations that follow from and are grounded in the data ought to be formulated.

* * * * *

The graphic below depicts the interrelationships among these four dimensions of evaluation sense-making. The three fundamental questions—What? So what? Now what?—are connected to the four evaluation processes of (1) organizing basic findings, (2) making interpretations, (3) rendering judgments, and (4) generating recommendations.

1. Basic findings

2. Interpretations

What?
So what?
Now what?

3. Judgments

4. Recommendations
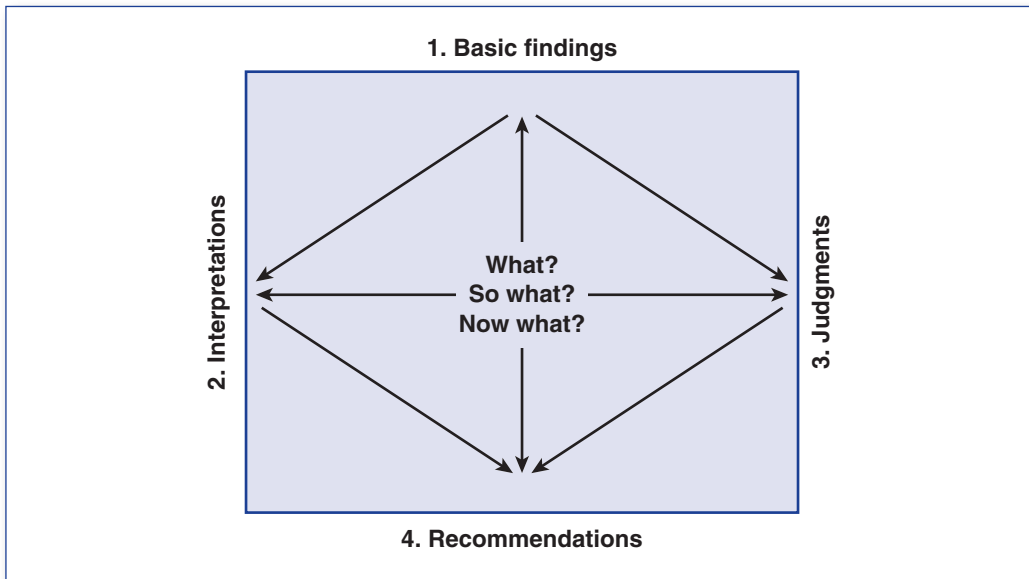
Arranging Data for Ease of Interpretation: Focusing the Analysis

Unless one is a genius, it is best to aim at being intelligible.

Sir Anthony Hope (1863–1933)

The first analytical task in evaluation is assembling and organizing the evidence to answer priority evaluation questions. Once presented, evidence can then be interpreted and a judgment rendered. In working with primary intended users, aim for the simplest presentation that will handle the facts. Evaluators may need and use sophisticated statistical techniques to enhance analytic power or uncover nuances in data, but understandable presentations are needed to give decision makers who are not researchers access to evaluation findings. Certainly, an evaluator can use sophisticated techniques to confirm the strength and meaningfulness of discovered patterns, but the next step is to think creatively about how to organize those findings into a straightforward and understandable format. This means, for example, that the results of a regression analysis might be reduced to nothing more complex than a chi-square table or a set of descriptive statistics (percentages and means). This need not distort the presentation. Quite the contrary, it will usually focus and highlight the most important findings while allowing the evaluator to explain in a footnote or appendix the more sophisticated techniques that were used to confirm the findings.

Our presentations must be like the skilled acrobat who makes the most dazzling moves look easy, the audience being unaware of the long hours of practice and the sophisticated calculations involved in what appear to be simple movements. Likewise, skilled evaluators craft and polish their presentations so that those participating will quickly understand the results, unaware of the long hours of arduous work involved in sifting through the data, organizing it, arranging it, testing relationships, taking the data apart, and creatively putting it back together to arrive at that moment of public unveiling. Here are seven tips for organizing and presenting findings.

1. **Answer the primary evaluation questions.** This would seem like a no-brainer, especially since Step 7 is focusing the priority evaluation questions with primary intended users. But some evaluators seem to get distracted and fail to keep the priority questions as the focus of analysis.

2. **Keep it simple.** This is also known as the KISS principle: *Keep it simple, stupid.* But evaluators being sensitive, diplomatic, and smart, we'll stick to KIS. Simplicity as a virtue means that we are rewarded not for how much we complicate the analysis or impress with our expertise, but for how much we enlighten. It means that we make users feel they can master what is before them, rather than intimidating them with our own knowledge and sophistication. It means distinguishing the complexity of analysis from the clarity of presentation and using the former to inform and guide the latter. Simplicity as a virtue is not simple. It often involves more work and creativity to simplify than blithely rest contented with a presentation of complicated statistics as they originally emerged from analysis.

Providing descriptive statistics in a report means more than simply reproducing the results in raw form. Data need to be arranged, ordered, and organized in some reasonable format that permits decision makers to detect patterns. Consider the three presentations of data shown in Exhibit 14.2. Each presents data from the same survey items, but the focus and degree of complexity are different in each case.

The first presentation reports items in the order in which they appeared on the survey with percentages for every category of response. It is difficult to detect patterns with 40 numbers to examine, so primary intended users will be overwhelmed by the first presentation. The second presentation simplifies the results by dividing the scale at the mid-point and reducing the four categories to two. Sometimes, such an analysis would be very

EXHIBIT 14.2

Three Presentations of the Same Data

Presentation 1: Raw results presented in the same order as items appeared in the survey.				
Expressed Needs of 478 Physically Disabled People	*Great Need for This*	*Much Need*	*Some Need*	*Little Need*
Transportation	35%	36%	13%	16%
Housing	33	38	19	10
Educational opportunities	42	28	9	21
Medical care	26	45	25	4
Employment opportunities	58	13	6	23
Public understanding	47	22	15	16
Architectural changes	33	38	10	19
Direct financial aid	40	31	12	17
Changes in insurance regulations	29	39	16	16
Social opportunities	11	58	17	14

Presentation 2: Results combined into two categories. No priorities emerge.		
	Great or much need	*Some or little need*
Transportation	71%	29%
Housing	71	29
Educational opportunities	70	30
Medical care	71	29
Employment opportunities	71	29
Public understanding	69	31
Architectural changes in buildings	71	29
Direct financial assistance	71	29
Changes in insurance regulations	68	32
Social opportunities	69	31

Presentation 3: Utilization-focused results arranged in rank order by "Great Need" to highlight priorities.	
Rank order	*Great need for this*
Employment opportunities	58%
Public understanding	47
Educational opportunities	42
Direct financial assistance	40
Transportation	35
Housing	33
Architectural changes in buildings	33
Changes in insurance regulations	29
Medical care	26
Social opportunities	11

revealing, but, in this case, no priorities emerge. Since *determining priorities was the purpose of the survey,* decision makers would conclude from the second presentation that the survey had not been useful.

The third presentation arranges the data so that decision makers can immediately see respondents' priorities. Support for employment programs now ranks first as a great need (58%) in contrast to social programs (11%), rated lowest in priority. Users can go down the list and decide where to draw the line on priorities, perhaps after "direct financial assistance" (40%). Failure to arrange the data as displayed in the third presentation places decision makers at an analytical disadvantage. Presentation three is utilization-focused because it facilitates quick understanding of and engagement with the results for their intended purpose: setting priorities for programs supporting people with disabilities.

This same principle applies to qualitative data. A single synthesis case study that captures and communicates major findings will focus the attention of intended users. The more detailed case studies on which the synthesis is based should be available as evidence for the validity of the synthesis, but for purposes of making sense of qualitative data, a summary of major themes and a synthesis case example will make the findings manageable.

3. **Provide balance.** The counterpoint to valuing simplicity is that evaluation findings are seldom really simple. In striving for simplicity, be careful to avoid simplemindedness. This happens most often when results are boiled down, in the name of simplicity, to some single number—a single percentage, a single cost-benefit ratio, a single proportion of the variance explained, or a significance test level. Balance and fairness should not be sacrificed to achieve simplicity. Multiple perspectives and conflicting findings have to be represented through several different numbers, all of them presented in an understandable fashion. Much advertising is based on the deception of picking the one number that puts a product in the best light, for example, gas mileage instead of price. Politicians often pick the statistic that favors their predetermined position.

An example comes from a study of Internal Revenue Service (IRS) audits conducted by the U.S. Government Accountability Office. The cover page of the report carried the sensational headline that IRS audits in five selected districts missed $1 million in errors in four months. The IRS response pointed out that those same audits had uncovered over $26 million in errors that led to adjustments in tax. Thus, the $1 million represented only about 4% of the total amount of money involved. Moreover, the IRS disputed the GAO's $1 million error figure because the GAO included all potential audit items whereas the IRS ignored differences of $100 or less, which are routinely ignored as not worth pursuing. Finally, the $1 million error involves cases of two types: instances in which additional tax would be due to the IRS and instances in which a refund would be due the taxpayer from the IRS. In point of fact, the $1 million in errors would have resulted in virtually no additional revenue to the government had all the errors been detected and corrected.

The gross simplification of the evaluation findings and the headlining of the $1 million error represent considerable distortion of the full picture. *Simplicity at the expense of accuracy is no virtue; complexity in the service of accuracy is no vice.* The point is to present complex matters in ways that achieve both clarity and balance.

4. Be clear about definitions. Confusion or uncertainty about what was actually measured can lead to misinterpretations. In workshops on data analysis I give the participants statistics on farmers, families, and juvenile offender recidivism. In small groups the participants interpret the data. Almost invariably they jump right into analysis without asking how farmer was defined, how family was defined, or what recidivism actually meant in the data at hand. A simple term like "farmer" turns out to be enormously variable in meaning. When does the weekend gardener become a farmer, and when does the large commercial farm become an "agribusiness"? A whole division of the Census Bureau wrestles with these definitional challenges.

Defining "family" is quite complex. There was a time, not so long ago, when Americans may have shared a common definition of family. Now there is real question about who has to be together under what arrangement before they constitute a family. Single-parent families, foster families, same-sex civil unions or marriages, and extended families are just a few of the possible complications. Before interpreting any statistics on families, it would be critical to know how family was defined.

Measuring recidivism is common in evaluation, but the term offers a variety of definitions and measures. Recidivism may mean (1) a new arrest, (2) a new appearance in court, (3) a new conviction, (4) a new sentence, (5) or actually committing a new crime regardless of whether the offender is apprehended. The statistics will vary considerably depending on which definition of recidivism is used.

A magazine cartoon I like shows a group of researchers studying cartoon violence. As they watch a television cartoon, one asks: "When the coyote bounces after falling off the cliff, does the second time he hits the ground count as a second incidence of violence?" Of such decisions are statistics made. But widespread skepticism about statistics ("lies and damn lies") is all the more reason for evaluators to exercise care in making sure that data are useful, accurate, and understandable. Clear definitions provide the foundation for utility, accuracy, and understandability.

5. **Make comparisons carefully and appropriately.**

Noncomparative evaluations are comparatively useless.

Michael Scriven (1993, p. 58)

Virtually all evaluative analysis ends up in some way being comparative. Numbers presented in isolation, without a frame of reference or basis of comparison, seldom make much

sense. A recidivism rate of 40% is a relatively meaningless statistic. Is that high or low? Does that represent improvement or deterioration? An error of $1 million in tax audits is a meaningless number. Some basis of comparison or standard of judgment is needed in order to interpret such statistics. The challenge lies in selecting the appropriate basis of comparison. In the earlier example of the IRS audit, the GAO evaluators believed the appropriate comparison was an error of zero dollars—absolute perfection in auditing. The IRS considered such a standard unrealistic and suggested, instead, comparing errors against the total amount of corrections made in all audits.

Skepticism can undermine evaluation when the basis for the comparison appears arbitrary or unfair. Working with users to select appropriate comparisons involves considering a number of options. Menu 14.1 presents 10 possibilities for making comparisons. Evaluators should work with stakeholders to decide which comparisons are appropriate and relevant to give a full and balanced view of what results are being achieved.

MENU 14.1

Menu of Program Comparisons

The outcomes of a program can be compared to

1. The outcomes of selected "similar" programs

2. The outcomes of the same program the previous year (or any other trend period, e.g, quarterly reports)

3. The outcomes of a representative or random sample of programs in the field

4. The outcomes of special programs of interest, for example, those known to be exemplary models (a purposeful sample comparison, Patton, 2002, pp. 230–234)

5. The stated goals of the program

6. Participants' goals for themselves

7. External standards of desirability as developed by the profession

8. Standards of minimum acceptability: for example, basic licensing or accreditation standards

9. Ideals of program performance

10. Guesses made by staff or other decision makers about what the outcomes would be

Combinations of these comparisons are also possible and usually desirable.

Consider the new jogger or running enthusiast. At the beginning, runners are likely to use as a basis for comparison their previously sedentary lifestyle. By that standard, the initial half-mile run appears pretty good. Then the runner discovers that there are a lot of other people running, many of them covering 3, 5, or 10 miles a week. Compared to seasoned joggers, the runner's half-mile doesn't look so good. On days when new runners want to feel particularly good, they may compare themselves to all the people who don't run at all. On days when they need some incentive to push harder, they may compare themselves to people who run twice as far as they do. Some adopt medical standards for basic conditioning, something on the order of 30 minutes of sustained and intense exercise a least three times a week. Some measure their progress in miles, others in minutes or hours. Some compare themselves to friends; others get involved in official competitions and races. All these comparisons are valid, but each yields a different conclusion because the basis of comparison is different in each case, as is the purpose of each comparison.

In politics it is said that conservatives compare the present to the past and see all the things that have been lost, while liberals compare the present to what could be in the future and see all the things yet to be attained. Each basis of comparison provides a different perspective. Fascination with comparisons undergirds sports, politics, advertising, management, and certainly, evaluation.

6. **Decide what is significant: Identify and focus on important and rigorous claims.** The level of evidence needed in an evaluation involves determining just what level of certainty is required to make findings useful. One way of meeting this challenge is to engage with primary stakeholders, especially program funders, administrators, and staff, about *making claims*. I ask: "Having reviewed the data, what can you claim about the program?" I then ask them to list possible claims; for example, (1) participants like the program, (2) participants get jobs as a result of the program, (3) the drop-out rate is low, (4) changes in participants last over the long term, (5) the program is cost-effective, and (6) the program does not work well with people of color, as examples. Having generated a list of possible claims, I then have them sort the claims into the categories (or cells) shown in Exhibit 14.3. This matrix distinguishes claims by their importance and rigor. Important claims speak to major issues of societal concern. Participants getting and keeping jobs as a result of a training program is a more important claim than that they're satisfied. Rigor concerns the amount and quality of evidence to support claims. The program might have very strong evidence of participant satisfaction, but very weak follow-up data about job retention. The most powerful, useful, and credible claims are those of major importance that have strong empirical support.

This claims framework can also be useful in the design phase to help intended users focus on gathering rigorous data about important issues so that, at the end, the evaluation will be able to report important and strong claims.

EXHIBIT 14.3

Important and Rigorous Claims

		Importance of claims	
		Major	Minor
Rigor of claims	Strong	*	
	Weak		

*GOAL: *Strong claims of major importance.*

The most powerful, useful, and credible claims are those that are of major importance and have strong empirical support.

Characteristics of Claims of *Major Importance*

- Involve making a difference, having an impact, or achieving desirable outcomes
- Deals with a problem of great societal concern
- Affects large numbers of people
- Provides a sustainable solution (claim deals with something that lasts over time)
- Saves money
- Saves time: that is, accomplished something in less time than is usually the case (an efficiency claim)
- Enhances quality
- Claims to be "new" or innovative
- Shows that something can actually be done about a problem: that is, claims the problem is malleable
- Involves a model or approach that could be used by others (meaning the model or approach is clearly specified and adaptable to other situations)

Characteristics of *Strong Claims*

- Valid, believable evidence to support the claim
- Follow-up data over time (longer periods of follow-up provide stronger evidence than shorter periods, and any follow-up is stronger than just end-of-program results)
- The claim is about a clear intervention (model or approach) with solid implementation documentation
- The claim is about clearly specified outcomes and impacts:

  Behavior outcomes are stronger than opinions, feelings, and knowledge

- The evidence for claims includes comparisons:
  —*To program goals*
  —*Over time (pretest, posttest, follow-up)*
  —*With other groups*
  —*With general trends or norms*
- The evidence for claims includes replications:
  —*Done at more than one site*
  —*More than one staff person attained outcomes*
  —*Different cohort groups of participants attained comparable outcomes over time*
  —*Different programs attained comparable results using comparable approaches*
- Claims are based on more than one kind of evidence or data (i.e., triangulation of data):
  —*Quantitative and qualitative data*
  —*Multiple sources (e.g., kids, parents, teachers, and staff corroborate results)*
- There are clear, logical, and/or empirical linkages between the intervention and the claimed outcomes
- The evaluators are independent of the staff (or where internal evaluation data are used, an independent, credible person reviews the results and certifies the results)
- Claims are based on systematic data collection over time
- Claims can be triangulated with and are supported by findings from research and other evaluations

CAVEAT: Importance and rigor are not absolute criteria. Different stakeholders, decision makers, and claims makers will have different definitions of what is important and rigorous. What staff deem to be of major importance may not be so to outside observers. What is deemed important and rigorous changes over time and across contexts. Making public claims is a political action. Importance and rigor are, to some extent, politically defined and dependent on the values of specific stakeholders.

One way to strengthen claims is to connect the evaluation's findings to results from other evaluations and triangulate with research findings. Former AEA President Bill Trochim calls the connection between research and evaluation the Golden Spike (Urban & Trochim, 2009).

7. **Distinguish facts from opinion.**

Everyone is entitled to his own opinion, but not his own facts.

Daniel Patrick Moynihan (1927–2003)
United States senator and distinguished social scientist

Moynihan's observation says it all, which takes us to opinions and interpretations.

"So you're anxious about conflicting positive and negative data. I'm afraid you're suffering from Bipolar Findings Disorder, a common evaluator affliction. It's caused by thinking people are rational and that the world makes sense. I can help."

Summary Principles of Analysis

Before turning to interpretations, let's review the seven principles of solid analysis just presented.

1. Answer the priority evaluation questions.

2. Keep the presentation as simple as the data will allow.

3. Be balanced in presenting alternative perspectives and conflicting data.

4. Be clear about definitions of key concepts.

5. Make comparisons carefully and appropriately.

6. Identify and focus on important and rigorous claims.

7. Distinguish facts from opinion.

Interpretations

We have been discussing how to analyze and organize data so that primary intended users can engage the evaluation findings. We turn now from describing the basic findings—answering the *What?* question—to interpretation—the *So what?* question. *Interpretation involves deciding what the findings mean.* How significant are the findings? What explains the results? Even when those receiving evaluation findings agree on the facts and findings, they can disagree vociferously about what the findings *mean.*

> ### The Importance of Interpretive Frameworks
>
> Management scholars Kathleen Sutcliffe and Klaus Weber (2003) examined the performance of business organizations in relation to the amount and accuracy of information used by senior executives as well as the "interpretive frameworks" they used to make sense of information. They concluded that *the way senior executives interpret their business environment is more important for performance than the accuracy of data they have about their environment.* That is, they concluded that there was less value in spending a lot of money increasing the marginal accuracy of data available to senior executives compared to the value of enhancing *their capacity to interpret* whatever data they have. Executives were more limited by a lack of capacity to make sense of data than by inadequate or inaccurate data. In essence, they found that interpretive capacity, or "mind-sets," distinguish high performance more than data quality and accuracy.
>
> Enhancing the quality and accuracy of our evaluation data through better methods and measures will add little value unless those using the data have the capacity to think evaluatively, think critically, and be able to appropriately interpret findings to reach reasonable and supportable conclusions.

In resisting the temptation to bear alone the burden of interpretation, the utilization-focused evaluator views the interpretive process as a training opportunity through which users can become more sophisticated about data-based decision making. Researchers have internalized the differences between analysis and interpretation, but that distinction will need reinforcement for nonresearchers. In working with stakeholders to understand interpretation, four themes deserve special attention.

1. **Numbers and qualitative data must be interpreted to have meaning.** Numbers are neither bad nor good, they're just numbers. Interpretation means thinking about what the data mean and how they ought to be applied. No magic formulas, not even those for

statistical significance, can infuse meaning into data. Only thinking humans can do that. Interpretation is a human process, not a computer process. Statisticians have no corner on the ability to think and reason. The best guideline may be Einstein's dictum that "the important thing is to keep on questioning."

2. **Data are imperfect indicators or representations of what the world is like.** Just as a map is not the territory it describes, the statistical tables describing a program are not the program. That's why they have to be interpreted.

3. **Statistics and qualitative data contain errors.** Research offers probabilities, not absolutes. The switch from absolute assertions (things either are or are not) to probabilistic thinking (things are more or less likely) is fundamental to empirical reasoning and careful interpretations.

4. **Look for intraocular significance.** Fred Mosteller, an esteemed statistician, cautioned against overreliance on statistically significant differences. He was more interested in "interocular differences, *the differences that hit us between the eyes*" (quoted in Scriven, 1993, p. 71).

Different stakeholders will bring varying perspectives to the evaluation. Those perspectives will affect their interpretations. The first task is get agreement on the basic findings—the "facts." Once there is understanding of the findings, the evaluator facilitates interpretation by having participants consider possible interpretations. Then follows the work of seeking convergence—aiming to reach consensus, if possible, on the most reasonable and useful interpretations supported by the data. Where different perspectives prevail, those varying interpretations should be reported and their implications explored. Judgments (discussed later) follow analysis and interpretations.

An Example of a Utilization-Focused Data-Based Deliberation With Stakeholders

In an evaluation of foster group homes for juvenile offenders, we collected data from natural parents, foster parents, juveniles, and community corrections staff. The primary intended users, the Community Corrections Advisory Board, agreed to a findings review process that involved a large number of stakeholders from both the field and policy levels. We had worked closely with the board in problem identification, research design, and instrumentation. Once the data were collected, we employed a variety of statistical techniques, including alpha factor analysis and stepwise forward regression analysis. We then reduced these findings to a few pages in a simplified form and readable format for use at a half-day meeting that included some 40 of the most powerful elected and appointed officials in the county as well as another 160 field professionals and foster parents.

A major purpose of the evaluation was to describe and conceptualize effective foster group homes for juvenile delinquents so that future selection of homes and training of foster parents could be improved. The evaluation was also intended to provide guidance about how to achieve better matches between juvenile offenders and foster parents. We presented findings on how recidivism, runaway rates, and juvenile attitudes varied by different kinds of group home environments. We had measured variations in foster home environments with a 56-item instrument. Factor analysis of these items uncovered a single major factor that explained 54% of the variance in recidivism. *The critical task in data interpretation was to label that factor in such a way that its relationship to dependent variables would represent something meaningful to primary intended users.* We focused the half-day work session on this issue.

The session began with a brief description of the evaluation's methods, then the results were distributed. In randomly assigned groups of four, these diverse stakeholders were asked to look at the factor analysis items and *label the factor or theme represented by those items in their own words.* After the groups reported their distinct labels, discussion followed. Consensus emerged around distinguishing *supportive and caring environments* from *authoritarian and nonsupportive foster home environments.*

The groups then studied tables showing the relationships between this treatment environment factor and program outcome variables (runaway and recidivism rates). The relationships were not only statistically significant but intraocularly so. Juveniles who reported experiencing more supportive foster home environments had markedly lower recidivism rates, lower runaway rates, and more positive attitudes. Having established the direction of the data, we discussed the limitations of the findings, methodological weaknesses, and the impossibility of making firm causal inferences. Key decision makers were already well aware of these problems. Then, given those constraints, the group was asked for recommendations. The basic thrust of the discussion concerned ways to increase the supportive experiences of juvenile offenders. The people carrying on that discussion were the people who fund, set policy for, operate, and control juvenile offender programs. The final written evaluation report included the recommendations that emerged from that meeting as well as our own independent conclusions and recommendations as evaluators. But, the final written report took another 4 weeks to prepare and print; the use process was already well under way as the meeting ended (both *findings use* and *process use*).

Four main points are illustrated by this example. First, nonresearchers can understand and interpret data when presented with clear, readable, and focused statistical tables. Second, as experienced data analysts know, the only way to really understand a data set is to spend some time getting inside it; busy decision makers are unwilling and unable to spend days at such a task, but a couple of hours of structured time spent in facilitated analysis and interpretation can pay off in greater understanding of and commitment to using results. Third, evaluators can learn a great deal from stakeholders' interpretations of data if they are open and listen to what people knowledgeable about the program have to say. Just as decision makers do not spend as much time in data analysis as do evaluators, so evaluators

do not spend as much time in program analysis, operations, and planning as do decision makers. Each can learn from the other in the overall effort to make sense out of the data and provide future direction for the program. Fourth, the transition from analysis to action is facilitated by having key actors involved in analysis. Use does not then depend on having to wait for a written report.

Making Causal Inferences

No discussion of evaluation nuts and bolts is complete without some mention of the **causation** *issue. . . . [C]ausation is both one of the most difficult and one of the most important issues in evaluation.*

E. Jane Davidson,
Evaluation Methodology Basics (2005, p. 67)

Because almost every evaluation involves some claim about causation—for example, that the program being evaluated had certain outcomes—this issue is of crucial importance in evaluation.

Michael Scriven, Causation (2005, p. 43)

Step 8 discussed attribution questions in designing an evaluation. Step 9 discussed conceptualizing the program's theory of change as partly a challenge of mapping causal hypotheses. Step 11 discussed making sure intended users understand potential methods controversies and their implications, especially the debate about whether experimental designs with randomized control groups as counterfactuals is the gold standard for establishing causality. Attribution and causality as prominent themes in question formulation and evaluation design highlight why these are also prominent themes when interpreting findings. The extent to which an intervention can be said to have caused observed outcomes is one of the crucial interpretation issues in evaluation.

A central question is: *Given the evaluation's purpose and intended use, what level of evidence is needed?* What degree of certainty is needed by primary intended users to use the evaluation findings? I discussed this earlier in considering the claims that are made in an evaluation's findings and making sure that major claims are supported by rigorous evidence.

An example from the chapter on Step 11 is worth reconsidering here. Suppose decision makers need to determine whether to give worm medicine to school-age children with diarrhea to increase their school attendance and performance. First, some context: 600 million people have hookworms. In Congo, one study found that 82% of children have worms making 70% anemic and affecting school attendance. Worms, elephantiasis, and trachoma kill 500,000 people annually; ordinary worms kill 130,000 people a year, through anemia and intestinal obstruction. Citing these statistics, advocates argue: "The cheapest way to

increase school attendance in poor countries isn't to build more schools, but to deworm children. Yet almost no government aid goes to deworming. In Africa, you can deworm a child for 50 cents" (Kristof, 2007, p. A19). So, what kind of evaluation evidence is needed to take action? Does one need a randomized controlled experiment to establish the linkage between deworming and school attendance—and the cost benefit of spending 50 cents per child per year? Or, if students, parents, teachers, and health professionals all affirm in interviews that diarrhea is a major cause of the poor school attendance and performance, and we follow up with those given a regimen of worm medicine, can we infer causation at a reasonable enough level to recommend action? If 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? *Primary intended users ultimately must answer these questions with discussion and implications facilitated by the evaluator.*

Direct inquiry into the relationship between worm medicine and school attendance, without an experimental design, involves tracing the causal chain and looking for *reasonable evidence* of linkages along the causal chain. This is how coroners determine cause of death, how arson investigators determine the cause of a fire, and how accident investigators determine the cause of an airplane crash. Epidemiologists follow backward the chain of events and contacts to establish the source of a disease or explain the outbreak of an epidemic. In all these cases, those carrying out the investigation examine the evidence and determine *the most probable cause.* Often they apply the principle of Occam's razor in choosing among alternative explanations:

All things being equal, the simplest solution tends to be the best one.

Occam's Razor: Valuing Straightforward Explanations

In the 14th century, an English logician, William of Ockham, postulated the principle that the explanation of any phenomenon should make as few assumptions as possible—eliminating or "shaving off" unnecessary complications. The simplest explanation *compatible with the data* is most valued. This principle, sometimes called the "law of parsimony," is popularly known as **Occam's razor:**

All things being equal, the simplest solution tends to be the best one.

Occam's razor is a heuristic guide to interpretation that emphasizes economy, parsimony, and simplicity—useful attributes for evaluators to aspire to in working with primary intended users.

Michael Scriven has called a related form of causal tracing the *modus operandi method*. This language comes from detective work in which a criminal's MO (*modus operandi* or method of operating) is established as a *signature trace* that connects the same criminal to different crimes. "The modus operandi method works best for evaluands that have highly distinctive patterns of effects" (Davidson, 2005, p. 75). I evaluated an employment training program aimed at chronically unemployed, poorly educated men of color. Prior to the program they blamed society for their problems and expressed overt anger. After the program, which included an intense empowerment component, they described themselves as taking control of their lives, abandoning anger, no longer indulging in a "victim mentality," and taking responsibility for their actions and the consequences of those actions. This language was the "signature" of the program. When graduates who had attained jobs attributed their success to being "empowered" and continued to express themselves in this way a year after leaving the program, it seemed reasonable to attribute this change in outlook to the program. Connecting the dots along the causal chain means looking at the participants' baseline attitudes and behaviors, looking at what they experienced in the program, and examining their subsequent attitudes, behaviors, and job status. The connections in this case were direct and reasonable.

Direct observation and logic are a powerful source of attribution. We don't need a randomized controlled trial to understand why parachutes work as they do (see sidebar).

A study in the *British Medical Journal* by Smith and Pell (2003) concluded:

No randomized control trials of parachute use have been undertaken.

As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evaluation by using randomized controlled trials. Advocates of evidence-based medicine have criticized the adoption of interventions evaluated by using only observational data. We think that everyone might benefit if the most radical protagonists of evidence-based medicine organized and participated in a double blind, randomized, placebo-controlled, crossover trial of the parachute.

SO

Only two options exist. The first is that we accept that, under exceptional circumstances, common sense might be applied when considering the potential risks and benefits of intervention.

OR

Those who criticize interventions that lack an evidence base will not hesitate to demonstrate their commitment by volunteering for a double-blind, randomized, placebo-controlled, crossover trial.

Engineers design machines, bridges, and buildings based on meeting specific criteria about what works. You don't need a counterfactual to determine if a bridge will get people across a river—or if using solar cookers in Africa reduces wood use (and deforestation). The evidence is direct and observable.

Contribution Analysis

In working with primary intended users, it can be quite useful to distinguish between attribution analysis and contribution analysis. John Mayne (2008, 2011) distinguishes attribution questions from contribution questions as follows:

Traditional causality questions (attribution)

- Has the program caused the outcome?
- To what extent has the program caused the outcome?
- How much of the outcome is caused by the program?

Contribution questions

- Has the program made a difference? That is, has the program made an important contribution to the observed result? Has the program influenced the observed result?
- How much of a difference has the program made? How much of a contribution?

Contribution analysis is especially appropriate where there are multiple projects and partners working toward the same outcomes, and where the ultimate impacts occur over long time periods influenced by several cumulative outputs and outcomes over time. Exhibit 14.4 elaborates *contribution analysis*.

A utilization-focused evaluator can support and facilitate primary intended users, including program staff and substantive experts, interpreting the data in search of explanations. Since the question of "Why did these results occur?" will inevitably arise, the evaluator can help primary intended users anticipate what level of evidence they will need to credibly answer that question to their own satisfaction, including understanding the challenges of establishing causality, and what expertise will be needed to generate explanations if so doing is deemed important.

Rendering Judgment

The four-part framework of this chapter on elucidating the meanings of evaluation findings consists of (1) analyzing and organizing the data so that primary intended users can

EXHIBIT 14.4

Contribution Analysis

Contribution analysis (Mayne, 2008, 2011) examines a causal hypothesis (theory of change) against logic and evidence to examine what factors could explain evaluation findings. The result of a contribution analysis is not definitive proof that the program has made an important contribution, but rather evidence and argumentation from which it is reasonable to draw conclusions about the degree and importance of the contribution, within some level of confidence. The aim is to get *plausible association* based on a preponderance of evidence, as in the judicial tradition. The question is whether a reasonable person would agree from the evidence and argument that the program has made an important contribution to the observed result. In utilization-focused evaluation, the "reasonable" persons making this assessment are the primary intended users.

A contribution analysis produces a *contribution story* that presents the evidence and other influences on program outcomes. A major part of that story may tell about behavioral changes that intended beneficiaries have made as a result of the intervention.

Attributes of a credible contribution story

A credible statement of contribution would entail:

- A well-articulated context of the program, discussing other influencing factors
- A plausible theory of change (no obvious flaws) which is not disproven
- A description of implemented activities and resulting outputs of the program
- A description of the observed results
- The results of contribution analysis
  - The evidence in support of the assumptions behind the key links in the theory of change
  - Discussion of the roles of the other influencing factors
- A discussion of the quality of the evidence provided, noting weaknesses

understand and engage the findings, (2) facilitating interpretation, (3) rendering judgment, and (4) generating recommendations. Having covered the first two, we arrive at the third, *the essence of the evaluative function.* At the center of the word *evaluation* is "valu[e]." Rendering a judgment involves applying values to the data and interpretation of the findings. Data are data. Findings alone do not determine whether a result is good or bad. Values and standards are needed for that determination. Data may show that gender equity or racial integration has increased as a result of a project intervention. Whether that increase is "good" or "enough to demonstrate merit or worth" depends

on what values inform that judgment. Those who support gender equity or racial integration will render a judgment of *good*. Those who oppose gender equity or racial integration will judge increases as *bad*. Regardless, the findings remain the findings. It is the judgment that varies depending on the values brought to bear.

Who makes this judgment? One perspective is that the evaluator must independently render judgment (Scriven, 1980, 1993). Others have argued that the evaluator's job can be limited to supplying the data and that the stakeholders alone make the final judgment (e.g., Stake, 1996). Utilization-focused evaluation treats these opposing views as options to be negotiated with primary users. The evaluator's job can include offering interpretations, making judgments, and generating recommendations if, as is typical, that is what the evaluation users want. Even so, in order to facilitate direct engagement and increase users' ownership, prior to offering *my* interpretations, judgments, and recommendations, I first give decision makers and intended users an opportunity to arrive at their own conclusions unencumbered by my perspective, but facilitated by me. That puts me in the role of *evaluation facilitator*—facilitating others' interpretation, judgments, and recommendations. In doing so, I find that I have to keep returning, sensitively and diplomatically, to the distinctions among analysis, interpretation, judgment, and recommendations.

Having facilitated the engagement of primary intended users, I can also render my own interpretations and judgments, either separately or as part of our interactive process. At that point I am playing the role of *evaluator*. In the active-reactive-interactive-adaptive role of a utilization-focused evaluation, I can move back and forth between the roles of evaluation facilitator and independent evaluator. In so doing I am alternating between the tasks of facilitating others' judgments and rendering my own. Some are skeptical that these dual roles of evaluation facilitator and independent judge can be played without confusion about roles or contamination of independence. Poorly executed, those are real dangers. But I find that primary intended users easily understand and value both roles.

I liken this process to that of skilled teachers who engage in both asking students questions (facilitating their critical thinking) and, alternatively, direct instruction (giving them answers and telling them what they need to know).

In facilitating judgments, I typically begin by offering three caveats:

- The quality of your judgment depends on the quality of the findings and thinking that informs it, thus the hand-in-glove link between findings and judgment.
- Don't condemn the judgment of another because it differs from your own. You may both be wrong.
- Forget "judge not that ye be not judged." The evaluator's mantra: Judge often and well so that you get better at it.

Practice Judging

Forget "judge not that ye be not judged."
The evaluator's mantra: Judge often and well so that you get better at it.

Halcolm

Recommendations

Student: What is the major source of problems in the world?

Sage: Solutions

Student: How can one recognize a problem in advance?

Sage: Look for a recommendation about to be implemented.

Student: What does this mean?

Sage: Evaluators who make recommendations are assuring future work for evaluators.

Halcolm

Recommendations are often the most visible part of an evaluation report. Well-written, carefully derived recommendations and conclusions can be the magnet that pulls all the other elements of an evaluation together into a meaningful whole. Done poorly, recommendations can become a lightning rod for attack, discrediting what was otherwise a professional job because of hurried and sloppy work on last-minute recommendations. I suspect that one of the most common reasons evaluators get into trouble when writing recommendations is that they haven't allowed enough time to really think through the possibilities and discuss them with people who have a stake in the evaluation. I've known cases in which, after working months on a project, the evaluators generated recommendations just hours before a final report was due, under enormous time pressure.

Useful and Practical Recommendations: 10 Guidelines

Recommendations, when they are included in a report, draw readers' attention like bees to a flower's nectar. Many report readers will turn to recommendations before anything else. Some never read beyond the recommendations. Given their importance, then, let me offer 10 guidelines for evaluation recommendations.

1. **Whether to include recommendations should be negotiated and clarified with stake-holders and evaluation funders as part of the design.** Not all evaluation reports include recommendations. What kinds of recommendations to include in a report, if any, are a matter for negotiation. For example, are recommendations expected about program improvements? About future funding? About program expansion? About sustainability? Asking questions about what recommendations are expected can clarify the focus and purpose of an evaluation *before data collection*.

2. **Recommendations should clearly follow from and be supported by the evaluation findings.** The processes of analysis, interpretation, and judgment should lead logically to recommendations.

3. **Distinguish different kinds of recommendations.** Recommendations that deal directly with central questions or issues should be highlighted and separated from recommendations about secondary or minor issues. Distinctions should be made between summative and formative recommendations. It may be helpful to distinguish recommendations that can be implemented immediately from those that might be implemented within 6 months to a year and those aimed at the long-term development of the program. In still other cases, it may be appropriate to orient recommendations toward certain groups of people: one set of recommendations for funders and policymakers; others for program administrators; still others for program staff or participants.

Another way of differentiating recommendations is to distinguish those that are strongly supported from those that are less so. Strong support may mean the findings directly lead to the recommendations or that the evaluation task force had strong agreement about the recommendation; other recommendations may be less directly supported by the data or there may be dissension among members of the task force. In similar fashion, it is important to distinguish between recommendations that involve a firm belief that some action should be taken and recommendations that are meant merely to stimulate discussion or suggestions that might become part of an agenda for future consideration and action.

The basic point here is that long, indiscriminate lists of recommendations at the end of an evaluation report diffuse the focus and diminish the power of central recommendations. By making explicit the different amounts of emphasis that the evaluator intends to place on different recommendations, and by organizing recommendations so as to differentiate among different kinds of recommendations, the evaluator increases the usefulness of the recommendations as well as the likelihood of the implementation of at least some of them.

4. **Some decision makers prefer to receive multiple options rather than recommendations that advocate only one course of action.** This approach may begin with a full slate of possible recommendations: Terminate the program, reduce funding for the program, maintain program funding at its current level, increase program funding slightly, and

increase program funding substantially. The evaluator then lists pros and cons for each of these recommendations, showing which findings, assumptions, interpretations, and judgments support each option.

5. Discuss the costs, benefits, and challenges of implementing recommendations. When making major recommendations that involve substantial changes in program operations or policies, evaluators should study, specify, and include in their reports some consideration of the benefits and costs of making the suggested changes, including the costs and risks of not making them.

6. Focus on actions within the control of intended users and those they can influence. Decision makers become frustrated when recommendations suggest actions over which they have no control. For example, a school desegregation study that focuses virtually all its recommendations on needed changes in housing patterns is not very useful to school officials, even though they may agree that housing changes are needed. Is the implication of such a recommendation that the schools can do nothing? Is the implication that anything the school does will be limited in impact to the extent that housing patterns remain unchanged? Or, again, are there major changes a school could make to further the aims of desegregation, but the evaluator got sidetracked on the issue of housing patterns and never got back to concrete recommendations for the school? Of course, the best way to end up with recommendations that focus on actionable variables is to make sure that, in conceptualizing the evaluation, the focus was on manipulability of the problem.

7. Exercise political sensitivity in writing recommendations. Ask yourself these questions: If I were in their place with their responsibilities, their political liabilities, and their personal perspectives, how would I react to this recommendation stated in this way? What arguments would I raise to counter the recommendations? Work with stakeholders to analyze the political implications of recommendations. This doesn't mean recommendations should be weak but, rather, that evaluators should be astute. Controversy may or may not serve the cause of getting findings used. But, at the very least, controversies should be anticipated and acknowledged.

8. Be thoughtful and deliberate in wording evaluations. Important recommendations can be lost in vague and obtuse language. Powerful recommendations can be diluted by an overly meek style, while particularly sensitive recommendations may be dismissed by an overly assertive style. Avoid words that confuse or distract from the central message. Here are examples.

Obtuse and meek recommendation: Consider whether current staffing competencies meet program needs and professional standards in light of changing knowledge and skill expectations.

Straightforward recommendation: Increase the amount and quality of staff development to meet accreditation standards.

9. Allow time to do a good job on recommendations. Plan time to develop recommendations collaboratively with stakeholders and time to pilot-test recommendations for clarity, understandability, practicality, utility, and meaningfulness.

10. Develop strategies for getting recommendations taken seriously. Simply listing recommendations at the end of a report may mean they get token attention. Think about how to facilitate serious consideration of recommendations. Help decision makers make decisions on recommendations, including facilitating a working session that includes clear assignment of responsibility for follow-up action and time lines for implementation.

Involving Intended Users in Generating Recommendations

As with everything else, utilization-focused evaluation actively involves primary intended users in the process of generating recommendations based on their knowledge of the situation and their shared expertise. Utilization-focused recommendations are not the evaluator's alone; they result from a collaborative process that seeks and incorporates the expertise of primary intended users.

Putting It All Together: Findings, Interpretation, Judgment, and Recommendations

This chapter has reviewed and discussed the four elements in a comprehensive framework for engaging evaluation results: basic findings, interpretation, judgment, and recommendations. A useful report brings these elements together in a coherent manner and relates them together so that analysis informs interpretations; analysis and interpretations, together, are the basis for judgments; and analysis, interpretations, and judgments lead to and are the explicit basis for recommendations. Exhibit 14.5 shows the outline for an evaluation summary that brings together and reports in sequence the data analysis findings, interpretations, judgments, and recommendation options for an employment training program targeted at high school dropouts.

While the distinction between description (what?) and prescription (so what and now what?) is fundamental in research and evaluation, it is important to note that description and prescription are ultimately intricately interconnected. As the Thomas theorem in social science asserts: What is perceived as real is real in its consequences. Distinguished *New York Times* journalist David Brooks (2010) puts the case more directly:

> Description is prescription. If you can get people to see the world [in a particular way] . . . , you have unwittingly framed every subsequent choice. (p. A37)

EXHIBIT 14.5

Putting It All Together: Basic Findings, Interpretation, Judgment, and Reporting

Evaluation of employment training program for high school dropouts

This shows the outline for an evaluation summary that brings together and reports in sequence the data analysis findings, interpretations, judgments, and recommendation options.

Findings from data analysis:

—All participants admitted to the program met the selection criteria of being high school dropouts who were chronically unemployed

—47% dropped out during the first 6 months this year (45 of 95) compared to a 57% drop-out rate in the same period the previous year.

—The drop-out rate for comparable programs that target a similar population is above 50%

—Of those who completed the program in the past year (35), 86% got a job and kept it for a year, making at least $12 an hour with benefits. The goal was 70%.

Interpretation: The program is serving its target population and exceeding its goal with those who complete the program. The drop-out rate is in line with other programs. The program attained these results at a time when the economy is sluggish and unemployment is somewhat higher than the historical average for this season. No one has solved the drop-out problem. This is a tough target population and difficult problem. The problem remains significant. The program has learned important lessons about how to retain and graduate participants (lessons reported separately).

Judgment: These are positive results. This is a fairly good program addressing an important societal issue. There is room for improvement, and the program shows promise for improvement based on results to date and lessons learned.

Recommendation options:

1. Renew funding at the current level for 2 more years to give the program more time to prove itself.

2. Increase funding to expand the program by 50% to test the program's capacity to increase its impact and go to scale.

Preparing for Use

As discussed in Step 12 on practice simulations, one dramatic way of setting the stage for analysis and use is having stakeholders speculate about results prior to seeing the real data. This can be done prior to data collection or after data collection but prior to actual

presentation of findings. Stakeholders are given an analysis table with all the appropriate categories but no actual data (a dummy table). They then fill in the missing data with their guesses of what the results will be.

This kind of speculation prepares users for how the results will be formatted and increases interest by building a sense of anticipation.

A second and more important function of having stakeholders write down their guesses is to provide a concrete basis for determining the extent to which actual results come close to expectations. Program staff members, for example, sometimes argue that they don't need formal evaluations because they know their clients, students, or program participants so well that evaluation findings would just confirm what they already know. I've found that when staff members commit their guesses to paper in advance of seeing actual results, the subsequent comparison often calls into question just how well some staff knows what is happening in the program. At least with written guesses on paper, program staff and other stakeholders can't just say, "That's what I expected." A database (in the form of their guesses) exists to determine and document how much new has been learned.

You can combine establishing standards of desirability (see Step 12) and speculating on results. Give stakeholders a page with two columns. The first column asks them to specify what outcomes they consider desirable and the second column asks them to guess what results they believe actually will be attained. Having specified a standard of desirability and guessed at possible results, users have a greater stake in and a framework for looking at the actual findings. When real results are presented, the evaluator facilitates discussion on the implications of the data falling below, at, or above the desired response, and why the actual findings were different from or the same as what they guessed. In facilitating this exercise, the outcomes data presented must be highly focused and limited to major issues. In my experience, animated interactions among users follow as they fully engage and interpret the results. Figuring out what findings mean and how to apply them engages us in that most human of processes: making sense of the world. Utilization-focused evaluators invite users along on the whole journey, alternatively exciting and treacherous, from determining what's worth knowing to interpreting the results and following through with action.

I find that, given the time and encouragement, stakeholders with virtually no methods or statistics training can readily identify the strengths, weaknesses, and implications of the findings. The trick is to move people from passive reception—from audience status—to active involvement and participation. This active engagement of primary intended users is distinct from the job of presenting the findings in a formal report. We turn in Step 15 to increasing the utility of formal evaluation reporting and dissemination of findings. Before doing so, let me close this chapter by emphasizing the skills involved in facilitating users' discussion, interpretation, and sense-making of findings. The challenge, as always, is staying focused on what is useful. That's where the evaluator adds value. An old and oft-told consulting story illustrates this point.

A company's furnace broke down in the midst of winter. Given the urgency of the situation, the chief operating officer authorized calling in a furnace expert who was known to be the best in the business. He agreed to come immediately. Upon arrival, he carefully examined the furnace, testing pipe connections, examining electrical wires, and cleaning away soot-laden areas. He asked questions about the furnace's history and recent functioning. When his questions had been answered and he had completed his inspection, he drew a red X on one side of the furnace. Using a large hammer, he hit the mark hard, the sound reverberating throughout the building. The furnace started working.

The furnace expert presented his bill. The maintenance director looked at the invoice and said, "I can't ask our financial officer to approve a bill of $1,000 for a 15-minute inspection and then hitting a furnace with a hammer." Sputtering he said, "I need to see a detailed invoice."

The furnace expert took the invoice, added a couple of lines, and handed it back to the maintenance director. Detailed invoice: "Hitting the boiler: $50. Knowing where to hit the boiler: $950."

The value added by a utilization-focused evaluation facilitator isn't just getting users to talk about the findings. It's getting them to talk about, reflect on, and reach conclusions about the right things.

Checklist details about what is involved in each step are provided in the summary *U-FE Checklist* in the concluding chapter. See pages 420–421 for the checklist items for Step 14 discussed in this chapter.

PRACTICE EXERCISES

1. Locate an evaluation report on some program of interest to you. Examine how the report handles the distinctions between analysis, interpretation, judgment, and recommendations (see Exhibits 14.1 and 14.5). Give examples of these distinctions from the report. Comment on and critique the extent to which these distinctions are adhered to in the evaluation.

2. Locate an evaluation report that includes recommendations. Examine the connection between findings and recommendations. To what extent can you connect the recommendations to the findings? Looking at the findings on your own, what additional recommendations, if any, occur to you?

3. Locate an evaluation report that includes comparisons. Analyze and discuss what comparisons were made. Generate additional potential comparisons using all the comparison alternatives in Menu 14.1 (Menu of Program Comparisons). You may have to make up data for some comparisons.

15

Prepare an Evaluation Report to Facilitate Use and Disseminate Significant Findings to Expand Influence

The single biggest problem with communication is the illusion that it has taken place.

George Bernard Shaw (1856–1950),
1925 Nobel Prize for Literature

The data from our study of federal health evaluations revealed that much important sharing of and discussion about findings and their significance is interpersonal and informal. In hallway conversations, over coffee, before and after meetings, over the telephone, and through informal networks, the word gets passed along when something useful and important has been found. Knowing this, evaluators can strategize about how to inject findings into important informal networks. This is not to diminish the importance of formal reports and oral briefings which, presented with thoughtful preparation and skill, can sometimes have an immediate and dramatic impact. But the increasing importance of networked communications in the Information Age carries a caution that *evaluators should not confuse producing a report with having communicated findings.*

In all cases, reporting is driven by intended evaluation purpose and the information needs of primary intended users. Formative reporting is different from a summative report. A lessons-learned report is distinct from an accountability report. When a single report serves multiple purposes (and audiences), clear distinctions should be made between sections

of the report. Bottom line: *Communicating and reporting should be strategic* (Torres, Preskill, & Piontek, 1996), which means honed and adapted to achieve use by targeted users. Dissemination to larger audiences then follows.

Formally Communicating Evaluation Findings: Report Menu

In logic model terms, an evaluation report is an output, not an outcome. It can *feel* like an outcome because so much work goes into producing a major evaluation report. But, alas, it is a means to an end, not the end itself, the end being use. Indeed, reports can hinder use when they are poorly written, too long, overly obtuse, and in countless ways anything but user friendly.

As with other stages in utilization-focused evaluation, the reporting stage offers a smorgasbord of options. Menu 15.1 displays alternatives for reporting format and style, content, contributors, and perspectives. Selecting from the menu is affected by the purpose of the evaluation. A summative report will highlight an overall judgment of merit or worth with supporting data. A knowledge-generating report aimed at policy enlightenment may follow

---

MENU 15.1

Evaluation Reporting Menu

Style and Format Options: Written Report

Traditional academic research monograph

Executive summary followed by a full report

Executive summary followed by a few key tables, graphs, and data summaries

Executive summary only (data available to those interested)

Different reports (or formats) for different targeted users

Newsletter article for dissemination

Press release

Brochure (well crafted, professionally done)

No written report; only oral presentations

Style and Format Options: Oral and Creative

Oral briefing with charts

Short summary followed by questions (e.g., at a board meeting or legislative hearing)

Discussion groups based on prepared handouts that focus issues for interpretation and judgment based on data

Half-day or full-day retreat-like work session with primary intended users

Videotape or audiotape presentation

Dramatic, creative presentation (e.g., role-playing perspectives)

Involvement of select primary users in reporting and facilitating any of the above

Advocacy–adversary debate or court for and against certain conclusions and judgments

Written and oral combinations

Content Options

Major findings only; focus on data, patterns, themes, and results

Findings and interpretations with judgments of merit or worth (no recommendations)

(a) Summative judgment about overall program

(b) Judgments about program components

Recommendations backed up by judgments, findings, and interpretations

(a) Single, best-option recommendations

(b) Multiple options with analysis of strengths, weaknesses, costs, and benefits of each

(c) Options based on future scenarios with monitoring and contingency suggestions

(d) Different recommendations for different intended users

Authors of and Contributors to the Report

Evaluator's report; evaluator as sole and independent author
Collaborative report coauthored by evaluator with others involved in the process
Report from primary users, written on their behalf by the evaluator as facilitator and adviser, but report ownership resides with others.

Combinations:

(a) Evaluator generates findings; collaborators generate judgments and recommendations

(b) Evaluator generates findings and makes judgments; primary users generate recommendations

(c) Separate conclusions, judgments, and recommendations by the evaluator and others in the same report

Perspectives Included

Evaluator's perspective as independent and neutral judge

Primary intended users only

Effort to represent all major stakeholder perspectives (may or may not be the same as primary intended users)

Program staff or administrators respond formally to the evaluation findings (written independently by the evaluator); GAO approach

Review of the evaluation by an external panel—*metaevaluation*: Formatively and summatively evaluate the evaluation using evaluation standards to elucidate strengths and weaknesses

a traditional academic format. A formative report may take the form of an internal memorandum with circulation limited to staff. I am often asked by students to show them the standard or best format for an evaluation report. The point of Menu 15.1 is that there can be no standard report format, *and the best format is the one that fulfills the purposes of the evaluation and meets the needs of specific intended users in a specific situation*. In many cases, multiple reporting strategies can be pursued to reach different intended users and dissemination audiences.

E. Jane Davidson, an independent evaluation consultant working out of New Zealand and author of the very useful *Evaluation Methodology Basics* featuring the nuts and bolts of sound evaluation (Davidson, 2005), has emphasized that evaluation reports should be structured around the demand for "actionable questions" (Davidson, 2010c, p. 13). My experience mirrors Jane's. Her reflections are so insightful and her voice so powerful that I want you to experience her thoughts on writing reports in her own words. In reviewing countless evaluation reports, she often found that they were "plagued with the structure of a Master's thesis," which made it quite difficult to figure out what results were important. This academic report format typically begins with a lengthy Executive Summary that presents "lots of introductory information, methodology, sampling, random snippets of findings that fail to give a clear sense of the program's quality or value, plus something incomprehensible about moderator variables." This is followed by

> an Introduction, Literature Review, a theoretical model and detailed explanation of the relevant social science theory explaining the links among some variables (unfortunately not a program logic model, and not even remotely linked to an evaluation question—this part contributed by a university faculty member with no evaluation expertise), Methodology, Findings (about 20 pages of raw data, all presented separately by source and data type with virtually no explanatory narrative, none of it linked back to the questions), Conclusions (some glimmers of hope in here, but by now we are 37 pages into the report and have lost most of our audience), Appendices. (Davidson, 2007, pp. v–vi)

She goes on to note that "for the client, reading a report like this feels like wading through mud. Page after page of graphs and interview quotes, but not a hint of whether or how they were used to answer any question of value. When, oh when, are they going to get to the point?" (p. vi).

In an effort to make reports more sensible and user friendly, Davidson recommends an alternative to the traditional research monograph format.

> One strategy I use is to structure the *Findings* part of the evaluation report into 7 +/– 2 sections, one for each of the 'big picture' evaluation questions used to frame the evaluation. In each section, all data pertaining to that question (qualitative, quantitative, interviews, surveys, observations, document analyses, from different people and perspectives) are presented, interpreted as they are presented, and woven together to form a direct answer to the question.

Next, I write a 2-page executive summary using the same structure: 7 +/– 2 questions with straight-to-the- point and explicitly evaluative answers of 1–2 paragraphs each.

If the client has seven or so major questions about the program that need to be answered, then the first two pages he or she reads (perhaps the only two pages!) should contain direct answers to those questions. And if the client wants to know on what basis those conclusions were drawn, it should be a simple matter to turn to the relevant section of the report and see clearly how 'quality' and 'value' were defined for that particular question, what data were used to answer it, and how they were interpreted together, relative to those definitions of quality/value. (Davidson, 2007, p. vi)

The 1:3:25 Format for Report Writing

The Canadian Health Services Research Foundation has pioneered a user-friendly approach to report writing that is becoming widely used as a way of communicating with focus. The 1:3:25 format specifies:

- One page for main messages and conclusions relevant to the reader
- A three-page executive summary of the main findings, and
- A 25-page comprehensive, plain-language report

Canadian Health Services Research Foundation (2008)

Another resource on structuring an evaluation report and making sure it contains all essential elements is the Evaluation Report Checklist (Miron, 2004). The thing to remember is that, while a report is only one part of the overall process, it is a concrete documentation of what has occurred and a visible representation of major findings. The quality of the report reflects on both the evaluator and the primary intended users. A great report won't ensure use, though it can help, while a lousy report can undermine not only use but future interest in evaluation among those who receive and read it. Take the time to do it well. Nothing undermines producing a quality report more assuredly than treating it like cramming for a final exam and spending an all-nighter just to get it done. Such reports are easy to spot—and undermine both credibility and utility, not to mention stress they induce and the toll they can take on the evaluator's mental health.

Utilization-Focused Reporting Principles

I've found the following principles helpful in thinking about how to make reporting useful:

1. Be intentional about reporting; that is, know the purpose of a report and stay true to that purpose.

2. Stay user focused: Focus the report on the priorities of primary intended users and answer their questions.

3. Use graphics and other visuals to communicate findings succinctly and powerfully.

4. Prepare users to engage with and learn from "negative" findings.

5. Distinguish dissemination from use.

Let me elaborate each of these principles.

1. Be intentional and purposeful about reporting.

Being intentional means negotiating a shared understanding of what it's going to mean to close-out the evaluation, that is, to achieve use. You need to communicate at every step in the evaluation your commitment to utility. One way to emphasize this point during early negotiations is to ask if a final report is expected. This question commands attention.

"Will you want a final report?" I ask.

They look at me and they say, "Come again?"

I repeat. "Will you want a final report?"

They respond, "Of course. That's why we're doing this, to get a report." And I respond. "I see it a little differently. I think we've agreed that we're doing this evaluation to get useful information to improve your programming and decision making. A final written report is one way of communicating findings, but there's substantial evidence now that it's not always the most effective way. Full evaluation reports don't seem to get read much and it's very costly to write final reports. A third or more of the budget of an evaluation can be consumed by report writing. Let's talk about how to get the evaluation used, then we can see if a full written report is the most cost-effective way to do that." Then I share Menu 15.1 and we start talking about reporting options.

Often I find that, with this kind of interaction, my primary intended users really start to understand what utilization-focused evaluation means. They start to comprehend that evaluation doesn't have to mean producing a thick report that they can file under "has been evaluated." They start to think about use. Caveat: Whatever is agreed on, especially if there's agreement not to produce a traditional academic monograph, get the agreement in writing and remind them of it often. A commitment to alternative reporting approaches may need reinforcement, especially among stakeholders used to traditional formats.

2. Focus reports on primary intended users and their priority questions.

A dominant theme running throughout this book is that use is integrally intertwined with users. That's the thrust of the personal factor. The style, format, content, and process of reporting should all be geared toward *intended use by intended users*. For example, we know that busy, big-picture policy makers and funders are more likely to read concise executive summaries than full reports, but detail-oriented users want—what else?—details.

Some users prefer recommendations right up front at the beginning of the report; others want them at the end; and I had one group of users who wanted the recommendations in a separate document so that readers of the report had to reach their own conclusions without interpreting everything in terms of recommendations. Methods sections may be put in the body of the report, in an appendix, or omitted and shared only with the methodologically interested. Sometimes users can't articulate what they want until they see a draft. Then they know what they don't want and the responsive evaluator will have to do some rewriting.

Beyond Generating a Report to Providing an Information Experience™

Information Experience™ is what *Juice Analytics* calls "the intersection between user experience and information-intensive applications, where success is how effectively a user can consume, understand, and apply that information."

Like sitting behind the wheel of a BMW or my two-year-old flipping through photos on an iPhone, great Information Experiences have less to do with features and more to do with an intimate connection between human and device. Great information experiences tell stories where data is the primary medium for communication. The information appears when it is needed and the device or application seems to anticipate the next question or action. These are the objectives that we apply to the solutions we design and build.

1. **Support the achievement of organizational objectives.** How can the information experience fit into users' existing decision-making and work processes? How can we influence decision making with the right information at the right time?

2. **Direct the user to likely actions in order to "get it done."** What are the important questions a user is trying to answer or tasks the user wants to accomplish? How can the application make it as easy and intuitive as possible to get to results? Does the navigation and user flow feel like an extension of users' thought process?

3. **Present only the information that needs to be seen.** For any given view of data and situational context, what is the most critical information to share with the user? How can information be progressively revealed to give the user what he or she needs to know at any given time?

4. **Present the information in a way that produces understanding and action.** For any given data and situational context, what is the most effective information visualization? What are the best ways to present information given users' experience and sophistication with interpreting information? What is the appropriate level of detail to be displayed given the context and user needs? (Juice Analytics, 2010, p. 1)

Consider this story from an evaluator in our federal use study.

Let me tell you the essence of the thing. I had almost no direction from the government [about the final report] except that the project officer kept saying, "Point 8 is really important. You've got to do point 8 on the contract."

So, when I turned in the draft of the report, I put points 1 through 9, without 8, in the first part of the report. Then I essentially wrote another report after that just on point 8 and made that the last half of the report. It was a detailed description of the activities of the program that came to very specific conclusions. It wasn't what had been asked for in the proposal I responded to, but it was what they needed to answer their questions. The project officer read it and the comment back was, "It's a good report except for all that crap in the front."

OK, so I turned it around in the final version, and moved all that "crap" in the front into an appendix. If you look at the report, it has several big appendices. All of that, if you compare it carefully to the contract, all that "crap" in the appendix is what I was asked to do in the original request and contract. All the stuff that constitutes the body of the report was above and beyond the call, but that's what he wanted and that's what got used.

"Boring! Not a single story in the whole thing."

3. **Use graphics and other visuals to communicate findings succinctly and powerfully.**

Mike Hendricks (1994) has studied effective techniques for executive summaries and oral briefings. The key, he has found, is good charts and graphics to capture attention and communicate quickly. A trend line, for example, can be portrayed more powerfully in graphic form than in a table, as Exhibit 15.1 shows. Hendricks trains evaluators on reporting and he asserts emphatically: "Evaluators have got to learn graphics. I'm amazed at how bad the charts and graphics are that I see in reports. You can't emphasize it too much. Reporting means GRAPHICS! GRAPHICS! GRAPHICS!" This involves "visible thinking," which includes causal mapping and other data displays (Bryson, Ackermann, Eden, & Finn, 2004).

The *Extreme Presentation™ Method* (2010) provides a set of tools for enhancing visual presentations including guidance on effective choice of visual formats and chart options for conveying different kinds of information. The *Periodic Table of Visualization Methods* (2010; Lengler & Eppler, n.d.) offers 100 examples of data visualization options, arranged thematically as a periodic table that you can interact with online. Nancy Duarte (2010) has assembled a set of videos on effective presentations that includes enhancing PowerPoint presentations (Duarte, 2008) as well as a range of techniques to more effectively engage audiences. Susan Kistler (2010a, 2010b), the executive director of the American Evaluation Association, is a leader in monitoring new developments in data visualization, like those cited here, and bringing them to the attention of evaluators.

Skilled visual facilitators can work with evaluators to facilitate, record, and represent the ideas of a group and map relationships between ideas and concepts shared by individuals in a group. California-based evaluator Terry Uyeki (2010) reports that "using graphic facilitation or recording often opens up thinking about patterns, themes, and a sense of the 'big picture' emerging from participant input processes. It is particularly effective when used with culturally diverse groups."

Journalist David McCandless (2010a), reflecting on "the beauty of data visualization," has asserted that *data is the new oil* in terms of its potential power in the Knowledge Age, or the new soil for growing knowledge—but only if it can be effectively accessed and displayed for appropriate understanding and use. Visualization, he explains and demonstrates, is a form of knowledge compression in which "the dataset can change your mindset." His "information is beautiful" website aims to do just that (McCandless, 2010b). No one has illustrated this phenomenon better than Swedish scholar Hans Rosling (2010), who animates statistical trends and graphically illustrates global development in 200 countries over the last 200 years in 4 minutes. His visualizations point the way to the future for evaluation presentations.

EXHIBIT 15.1

The Power of Graphics
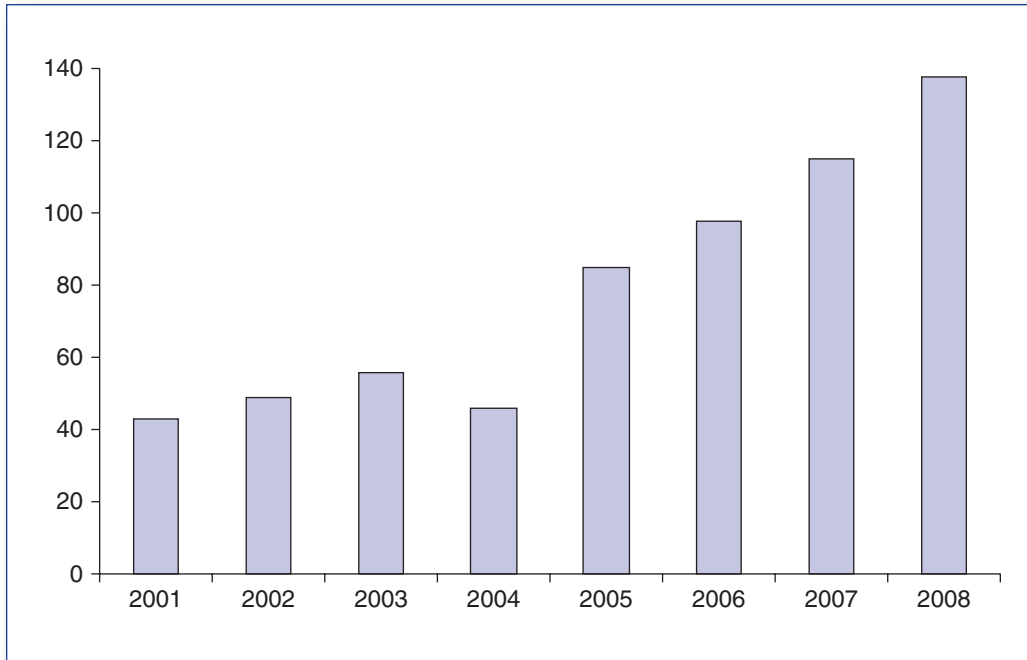
4. **Prepare users to engage with and learn from "negative" findings.**

There is nothing either good or bad, but thinking makes it so.

Hamlet, Act 2, Scene 2, 239–251

William Shakespeare

The program staff's fear of negative results can undermine an evaluation. On the other hand, the absence of negative findings can call into question the evaluator's independence, integrity, and credibility. Here, then, is where evaluation use can take a back seat to other agendas. Staff will resist being made to look bad and will often treat the mildest suggestions for improvements as deep criticisms. Evaluators, worried about accusations that they've lost their independence, emphasize negative findings. As we grapple with these tensions, two points are worth remembering: (1) one person's negative is another person's positive; and (2) evaluators can do much to increase staff receptivity by shifting the focus of reporting to learning and use rather than simply being judged as good or bad.

Few evaluations are all negative or all positive. It's helpful to move beyond a dichotomous win/lose, pass/fail, success/failure, and positive/negative construct on evaluation results. This tendency to label evaluation findings as either positive or negative seems born of a tendency I find common among evaluators and decision makers: to think of evaluation findings in monolithic, absolute, and purely summative terms. This becomes especially true when evaluation findings get into the media—which tends to exaggerate the negative because negative findings make more compelling and attention-grabbing headlines. I reiterate that, in my experience, evaluation findings are seldom either completely positive or completely negative. Furthermore, whether findings are interpreted as positive or negative depends on who is using and interpreting the findings. As the old adage observes: *Whether the glass is half empty or half full depends on whether you're drinking or pouring.* Evaluators can shape the environment and context in which findings are reviewed so that the focus is on learning and improvement rather than absolute judgment. This is part of our overall responsibility to strive for balance.

5. Distinguish dissemination from use.

Dissemination of findings to audiences beyond intended users is distinct from the kind of use that has been the focus of this book. Studies can have an impact on all kinds of audiences in all kinds of ways. As a social scientist, I value and want to encourage the full and free dissemination of evaluation findings. Each of us ought to be permitted to indulge in the fantasy that our evaluation reports will have impact across the land and through the years. But only a handful of studies will ever enjoy (or suffer) such widespread dissemination.

Dissemination takes us beyond intended use by intended users into the broader concept of *evaluation influence* (Kirkhart, 2000), both intended and unintended, and longer-term *evaluation consequences* generally (Mark, 2006). This includes instances where planned dissemination *hopes for* broader influence but can't be sure if or where this will occur.

Dissemination efforts will vary greatly from study to study. The nature of dissemination, like everything else, is a matter for negotiation between evaluators and decision makers. In such negotiations, dissemination costs and benefits should be estimated. The questions addressed in an evaluation will have different meanings for people not directly involved in the painstaking process of focusing the evaluation. Different individuals and audiences will be interested in a given evaluation for reasons not always possible to anticipate. Effective dissemination involves skills in extrapolating the evaluation specifics of a particular study for use by readers in a different setting (raising issues of external validity and generalizability).

The problematic utility of trying to design an evaluation relevant to multiple audiences, each conceptualized in vague and general terms, was what has led to the emphasis in utilization-focused evaluation on identification and organization of primary intended users. Dissemination can broaden and enlarge the impact of a study in important ways, but the nature of those long-term impacts is largely beyond the control of the evaluator.

What the evaluator can control is the degree to which findings address the concerns of specific intended users. That is the use for which I take responsibility: intended use by intended users. Dissemination is not use, though it can be useful.

Exhibit 15.2 depicts the complex, dynamic relationship between use and diverse strategies for dissemination and influence. At the center, as the bull's-eye, is **intended uses by intended users.** The utilization-focused evaluator works with intended users to plan formal dissemination and influence strategies if the findings are of sufficient import to merit sharing more widely; these formal pathways are depicted by the solid box on the left. At the same time, informal networks can be energized for dissemination, as shown in the dotted-line box to the right. Informal networks are likely to generate some unexpected and emergent opportunities for further dissemination and influence. Moreover, in a complex dynamic system, some formal pathways will manifest links to informal networks, as shown in the feedback arrow at the bottom of the diagram, even as some informal dissemination networks may generate and lead to formal dissemination strategies, like publications and conference presentations. The dynamic links that begin informally and opportunistically but then morph into formal and planned strategies are represented by the meandering arrows at the top. The intended uses by intended users can further generate unanticipated but important new opportunities for dissemination and influence. As Exhibit 15.2 shows, dissemination can be a multifaceted, many-splendored phenomenon, like the diffusion of birth control pills, mobile phone applications, and the uses of the Internet for information dissemination, but always at the core, always in the spotlight, always in high-definition focus, is *intended uses by intended users*.

Use Is a Process, Not a Report

Analyzing and interpreting results can be exciting processes. Many nights have turned into morning before evaluators have finished trying new computer runs to tease out the nuances in some data set. The work of months, sometimes years, finally comes to fruition as data are analyzed and interpreted, conclusions drawn, recommendations considered, and the evaluation report finalized. Great relief comes in finishing an evaluation report, so much relief that it can seem like the report was the purpose. But use is the purpose and, as this book has emphasized throughout, use is a process, not a report or single event.

I remember fondly the final days of an evaluation when my co-evaluators and I were on the phone with program staff two or three times a day as we analyzed data on an educational project to inform a major decision about whether it met criteria as a valid model for federal dissemination funding. Program staff shared with us the process of watching the findings take final shape. Preliminary analyses appeared negative; as the sample became more complete, the findings looked more positive to staff; finally, a mixed picture of positive and negative conclusions emerged. Because the primary users had been intimately involved in designing the evaluation, we encountered no last-minute attacks on methods to

EXHIBIT 15.2

Complex, Dynamic Relationship Between Use and Dissemination Approaches, Both Planned and Unplanned
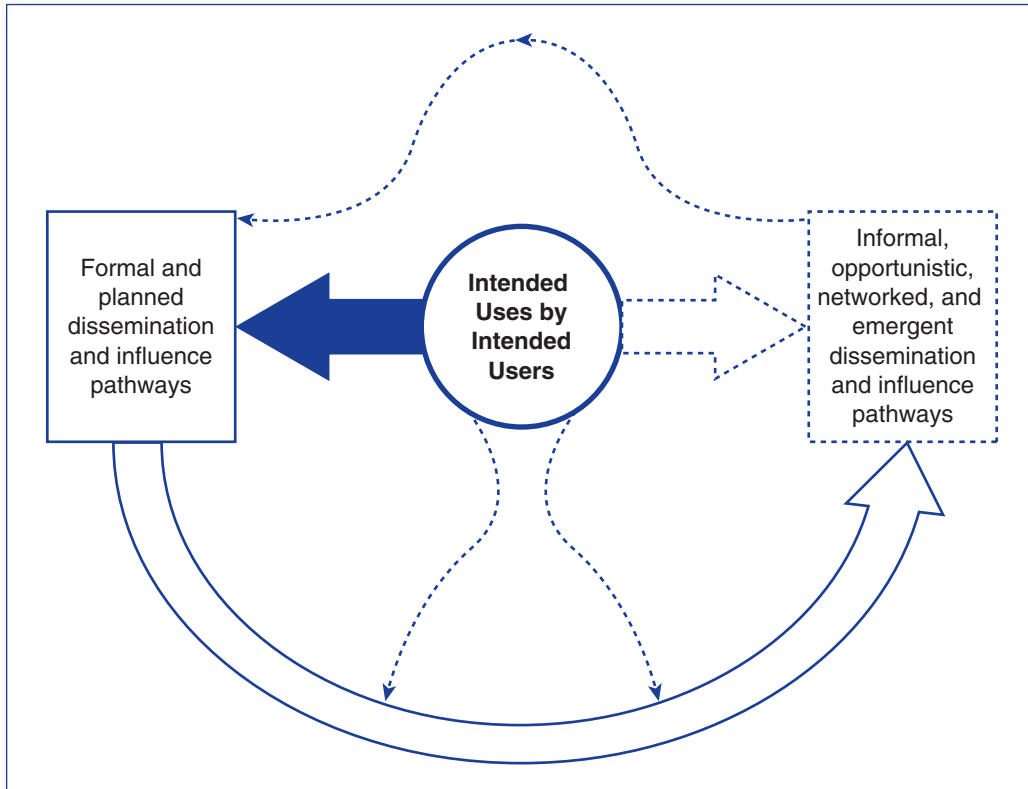
explain away negative findings. The program staff understood the data, from whence it came, what it revealed, and how it could be used for program development. They didn't get the dissemination grant that year, but they got direction about how to implement the program more consistently and increase its impact. Two years later, with new findings, they did win recognition as a "best practices" exemplar, an award that came with a dissemination grant. The highly polished summative evaluation report made that recognition possible and was a central part of the dissemination process. One intended use of the evaluation was to influence thinking and practices generally, but to achieve that dissemination and influence

purpose, the evaluation first had to be useful to the primary intended users who funded, developed, implemented, and adapted the program in the years leading up to the final summative evaluation findings and report.

Ultimately, of course, utility is linked to accuracy. As reports get disseminated, any inaccuracies will take the spotlight and subject the report to potential ridicule. One of the most obvious examples of this phenomenon is premature obituary reports, as when the *New York Journal* published the obituary of American humorist Mark Twain. Upon hearing the news, Twain famously replied: "The reports of my death are greatly exaggerated." A more recent example is that of Dave Swarbrick, a British folk-rock violinist, who was killed off mistakenly by the *Daily Telegraph* in April 1999. The newspaper reported that he had been hospitalized in Coventry, where he subsequently died. His witty response to the news: "It's not the first time I have died in Coventry."

And what have these obituary examples to do with evaluation reports? Just this: In the Internet age, reports have a long life. Inaccurate reports seem especially hard to kill, as fake news, fabricated statistics, and distorted findings circulate round and round, generating affirmation by repetition rather than rigor. As *New Yorker* journalist Jonah Lehrer (2010a) found in a recent review of scientific publishing, "Many scientific theories continue to be considered true even after failing numerous experimental tests. . . . Although many scientific ideas generate conflicting results and suffer from falling effect sizes, they continue to get cited in the textbooks and drive standard medical practice" (p. 57). The long life expectancy of widely disseminated evaluation reports in the Internet age raises the stakes for determining what findings deserve a wide audience.

Taking Report Writing Seriously

Report writing is serious business, often with high stakes, and typically consuming considerable time and resources—which is why evaluation conferences often include a lot of self-deprecating humor about report writing, like this advice:

> Remember that people don't actually want to read an evaluation report. The purpose of reports is to decorate offices, not inform readers. So the most important things are cover color and size. Don't ask intended users about content and substance. Ask their preferences about cover color and size so that the report will nicely match their office décor and therefore be more likely to be prominently displayed. That's a *best practice* dissemination strategy.

Or this, on writing annual accountability reports:

> Those that think you do good work don't need to know your warts. Those that think you are a waste of space are not going to be convinced otherwise. So write for the great mass in between who just don't want to know. Help them not know. Nurture them.

The safest path is to appear to have said a good deal, without having said anything at all. With a little practice, you can incorporate the ideas presented to produce impeccable annual reports that will confuse people so badly they won't even know they are confused. Look for the knowing nods of people when they read your report. That will tell you you have succeeded. (Bacal, 2009, p. 1)

Skepticism about the value of reports notwithstanding, effective report writing is an essential evaluator competence. Work at writing. Take time to do it well. Get feedback on how well you do. Evaluate use.

In doing so, you will be engaging in "utilization focused communication"—which is not an oxymoron according to international communications expert Ricardo Ramírez. He further asserts that "*communication focused evaluation* is not an oxymoron." These observations flow from reflections in which "evaluation and communication approaches and methods keep on interconnecting in my mind" (Ramírez, 2011). In essence, when thinking evaluation, think communications; when thinking communications, think evaluation. And make both utilization-focused.

Checklist details about what is involved in each step are provided in the summary *U-FE Checklist* in the concluding chapter. See pages 421–422 for the checklist items for Step 15 discussed in this chapter. (See pp. 375–376 for discussion of this distinction.)

PRACTICE EXERCISES

1. Using Menu 15.1, discuss the reporting situation that would be appropriate for each of the Style and Format Options (both Written and Oral). Show that you can match the reporting option to a situation for which that option is a good match. Make it clear how the situation you describe lends itself to each reporting option.

2. Find an evaluation report on the Internet. See if you can distinguish use (intended use by intended users) from dissemination. Discuss the implications of the distinction between use and dissemination using the particular example you've found to illustrate the implications you generate.