From *Utilization-Focused Evaluation*, 4th ed
by Michael Quinn Patton (2008, Sage

13

The Meanings and Reporting of Evaluation Findings

Analysis, Interpretation, Judgment, and Recommendations

What is the sound of one hand clapping?

—Hakuin

This question was first posed by the Japanese Zen master Hakuin (1686–1769) as a means of facilitating enlightenment. “The disciple, given a Koan [riddle] to see through, was encouraged to put his whole strength into the singleminded search for its solution, to be ‘like a thirsty rat seeking for water . . . ,’ to carry the problem with him everywhere, until suddenly, if he were successful, the solution came” (Hoffman 1975:22). The koan is a technique originated by the Zen masters to shake their students out of routine ways of thinking and acting, open up new possibilities, and help individual students realize their full potential. An effective evaluator can facilitate these same processes. Utilization-focused evaluation helps decision makers and intended users stand outside the program and look at what is happening; evaluations can help shake staff out of routine ways of doing things, open up new possibilities, and help programs realize their full potential.

The Zen search through koans consists of three basic parts: a question, an answer, and interpretation/assimilation of the answer in terms of the student's own life; evaluation involves a question, an empirical answer, and interpretation/utilization of the answer in the context of the program's own dynamics. A fundamental tenet of the koanic method is that the question is as important as the answer; the same principle applies to utilization-focused evaluation. The Zen master carefully matches the koan to the student; the responsive evaluator focuses on questions that are relevant to specific intended users. Finally, the Zen student must struggle to make sense out of the answer to the koanic riddle; in evaluation, the meaning of empirical data emerges from interpretation, dialogue and situational application. Consider the following koanic exchange, titled "A Flower in Bloom."

> *A monk asked Master Ummon, "What is the pure body of truth?"*
> *Master Ummon said, "A flower in bloom."*
> *Monk: "'A flower in bloom'—what's it mean?"*
> *Master: "Maggot in the shit hole, pus of leprosy, scab over a boil"*
>
> —(Hoffman 1975:119).

"What's it mean?" may be a philosophical, religious, or epistemological question. It can also be the very concrete, practical question of program staff laboring over statistical tables. For any given set of data, meaning depends on who is interpreting the data. Some people see flowers; others see maggots.

Evaluators and decision makers can deceive themselves into believing that once data have been collected it will be clear whether or not the program works. But data have to be interpreted. In utilization-focused evaluation, interpretation involves the active participation of primary users because, in the end, they are the ones who must translate data into decisions and action, and evidence into conclusions.

Evidence

What is the meaning of the word evidence?

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. (Safire 2006:18)

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.

Setting the Stage for Use

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—I like to 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 percent for black participants and 75 percent 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.

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 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. The 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 (Taut and Alkin 2003; Taut and Brauns 2003).

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. Athletes, performing artists, astronauts, fire fighters, and entertainers spend hundreds of hours preparing for events that take only a few hours. 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?

Standards of Desirability

A simulated use session also offers a prime opportunity to think about and formalize criteria for making judgments—*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. level at which the program is considered highly effective,

The Power of Interpretative Frameworks: Positive, Negative, Balanced? Expected, Unexpected?

How should an evaluator facilitate stakeholders' thinking about findings? One way to begin the process and set the stage for interpretation is to have primary intended users spend a few minutes thinking about their interpretive tendencies. Do they tend to see the glass as half-full, half empty, or simply descriptively: The 8 ounce glass has 4 ounces of water.

For more than a half-century, "The Power of Positive Thinking" (Peale 1952) has been a dominant mind set in Western culture. Reflecting on this perspective, Harvard medical School Surgeon Atul Gawande (2007) counters that, to increase effectiveness, the key is "negative thinking: looking for, and sometimes expecting, failure" (p. A23).

Gawande tells of visiting the Walter Reed military hospital early in the Iraq war. He participated in a session interpreting eye-injury statistics. The doctors were having considerable success saving some soldiers from blindness, a positive outcome. But digging deeper, the doctors asked why so many severe eye injuries were occurring. Interviewing their patients, they learned that the young soldiers weren't wearing their protective goggles because they were considered too ugly and uncool. They recommended that the military switch to "cooler-looking Wiley X ballistic eyewear. The soldiers wore their eyegear more consistently, and the eye-injury rate dropped immediately" (p. A23). By "negative thinking," Gawande means not just looking for what's going well but asking hard questions about and digging deeply into problems.

Evaluators typically seek to facilitate balance between the positive and negative. The point of introducing the discussion with stakeholders is to get them thinking about what lens they typically bring to interpreting findings.

Another framework involves distinguishing confirming from disconfirming findings. People tend to like to have their opinions confirmed and may view findings from that perspective. In contrast, organizational development researchers Weick and Sutcliffe (2001) found that high performance organizations are always on the lookout for the unexpected. Helping intended users identify their interpretive tendencies can increase their willingness to engage evaluation findings openly.

2. level at which the program is considered adequate, and
3. 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 (data on how participants have changed). Suppose one is 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 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, done before data collection, may also reveal that satisfaction data alone are an inadequate indicator of effectiveness while there's still time to measure additional outcomes.

The process of specifying objectives sometimes involves setting performance targets: for example, "75% of workshop participants will be satisfied." However, 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 percent satisfaction is inadequate or merely "lower than we hoped for but acceptable." Moreover, objectives are often set a long time before the program is under way or well before an actual evaluation has been designed. Reviewing objectives and establishing precise standards of desirability just before data collection increases the likelihood that judgment criteria will be up to date, realistic, and meaningful.

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 percent, 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?

Consider the evaluation of a teacher resource center. One of the implementation issues concerned the extent to which teachers used the center intensively (three or more times) versus superficially (once or twice). Actual baseline data from such a study are shown in Exhibit 13.1 with three categories for primary intended users to set future standards of desirability for the next period of implementation.

Now, suppose the staff assembles in six months to discuss the actual results *without* having set standards of desirability or performance targets.

First staff member: Those results are about what I anticipated.

Second staff member: Plus, remember, the data don't include teachers in our workshops and special classes.

Third staff member: I think the time was really too short to conclude anything. We're still getting established.

First staff member: I agree. And winter is bad. You know, everyone is depressed with winter, and . . .

Soon it becomes apparent that either the findings don't tell staff much about teacher engagement, at least not without other data, or staff members are not prepared to deal with what the data do show. Such resistance and defensiveness are not unusual when staff first interpret evaluation data.

Now, let's try a different scenario. At the outset of evaluation, the program staff discuss their notions of what their task is and how teacher change occurs. They decide that the kind of impact they want cannot occur in one or two visits to the teacher center. "If teachers don't return after one or two visits, we must be doing something wrong." The period of time in question is a full 12-month period. Before the data are collected, the staff fill in the table establishing standards of desirability as shown in Exhibit 13.1.

A record-keeping system of teacher visits must be established that staff believes has credibility. The data will provide clear feedback about the effectiveness of the program's outreach and implementation in attempting to engage teachers on a multiple-contact basis. The key point is that if staff members are unwilling or unable to set expectancy levels *before data collection,* there is no reason to believe they can do so afterward. In

EXHIBIT 13.1

Intensity of Teachers' Use of a Teacher Center: Baseline Data and Standards of Desirability

Category of Visits by a Teacher in a Month	*Baseline Number of Visits per Month*	*Percentage of Total Teacher Visits*
1 or 2	185	80.4
3 or more	45	19.6

Given the baseline data (above), what are the Teacher Center *Standards of Desirability* for the next year?

Judgment	*Percentage and Number of Teachers Who Use the Center Three or More Times*
We're doing an outstanding job of engaging teachers at this level	
We're doing an adequate job of engaging teachers at this level	
We're doing a poor job of engaging teachers at this level	

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; clearly, measuring the frequency of visits is only a starting place.

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

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 with 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.

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 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.

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.

A Framework for Engaging Findings

Four distinct processes are involved in making sense out of evaluation findings: (1) *analysis,* which involves organizing raw data into an understandable form that reveals basic patterns and constitutes the evaluation's empirical findings; (2) *interpretation,* which involves determining the significance of and explanations for the findings; (3) *judgment,* which brings values to bear to determine merit or worth and decide whether the results are positive or negative; and (4) *recommendations,* which involve determining the action implications of the findings. Primary intended users should be actively involved in all four of these processes so that they fully understand the findings and their implications. Facilitating these processes, especially helping stakeholders understand these four fundamental distinctions, requires skills that go well beyond what is taught in statistics courses. Working with stakeholders to analyze and interpret findings is quite different from doing it on one's own as a researcher. Exhibit 13.2 summarizes this framework. We'll now consider each of these processes in greater depth.

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)

EXHIBIT 13.2

A Utilization-Focused Framework for Engaging Findings

Four distinct processes are involved in helping primary intended users make sense out of evaluation findings.

1. *Basic Findings, Description and Analysis*: Organize raw data, both quantitative and qualitative, into a form that reveals basic patterns so that primary intended users can understand the results.
2. *Interpretation*: Engage 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. *Judgment*: Values are added to analysis and interpretations to make judgments. Determining merit or worth means determining the extent to which results are positive or negative, what is good or bad, desirable or undesirable, in the outcomes, and to what extent standards of desirability have been met. Help primary intended users make judgments.
4. *Recommendations*: 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.

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.

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 intimidate 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 to rest content with a presentation of complicated statistics as they originally emerged from analysis.

The first step is realizing that 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 13.3. 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 midpoint and reducing the four categories to two. Sometimes, such an analysis would be very 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 percent) in contrast to social programs (11 percent), rated lowest in priority. Users can go down the list and decide where

EXHIBIT 13.3

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 (Percent)*	*Much Need (Percent)*	*Some Need (Percent)*	*Little Need (Percent)*
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 (Percent)	*Some or Little Need (Percent)*
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 (Percent)*
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

to draw the line on priorities, perhaps after "direct financial assistance" (40 percent). Failure to arrange the data as displayed in the third presentation places decision makers at an analytical disadvantage. This presentation 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.

Balance

The counterpoint to valuing simplicity is that evaluation findings are seldom really simple. In striving for simplicity, one must be careful to avoid simplemindedness. This happens most often in evaluation when results are boiled down, in the name of simplicity, to some single number—a single percentage, a single cost/benefit ratio, or a single proportion of the variance explained. Striving for simplicity means making the data understandable, but balance and fairness need not be sacrificed in the name of simplicity. Achieving balance may mean that multiple 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, e.g., gas mileage instead of price. Politicians often do likewise, picking the statistic that favors their predetermined analysis. For example, Exhibit 13.4 shows how seemingly contradictory statements can both be true: "In the last four years, median incomes for African Americans have risen faster than white incomes" and "after the last years, African Americans are worse off than whites in terms of income." The data in Exhibit 13.4 show that both statements can be true.

Each statement represents only part of the picture. To understand what is happening in the relationship between black and white incomes, one needs to know, at a minimum, *both* absolute income levels and percentage changes. When a report gives only one figure or the other (i.e., only absolute changes or only percentage changes), the reader has cause to suspect that the full picture has not been presented.

EXHIBIT 13.4

Different Data Tell Different Stories: Illustrative Data (Constructed)

	Beginning Level	*Absolute Level 4 Years Later*	*Amount of Change*	*Percentage Increase*
Median white income	$20,100	$21,205	$1,105	5.5
Median African American income	$10,400	$11,336	$936	9.0

These data support seemingly contradictory conclusions, each of which is true: "In the last four years, median income for African Americans has risen faster than white incomes" (9% versus 5.5%) *and* "in the last years, African Americans are worse off than whites in terms of income" (the absolute gap in median incomes has increased in this illustration).

Another example comes from a study of Internal Revenue Service (IRS) audits conducted by the U.S. Government Accountability Office (GAO). 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 to the GAO report pointed out that the same audit cases with $1 million in errors had uncovered over $26 million in errors that led to adjustments in tax. Thus, the $1 million represented only about 4 percent 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. In the data presented by the GAO it is impossible to tell what proportion of the $1 million involved errors of under $100, which are routinely ignored by the IRS as not worth the costs of 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 to the taxpayer from the IRS. In point of fact, the $1 million error would result in virtually no additional revenue to the government had all the errors been detected and followed up.

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 make complex matters understandable without distortion. The omitted information from the GAO report could not be justified on the basis of simplification. The omissions constituted distortions rather than simplification.

Striving for balance means thinking about how to present the full picture without getting bogged down in trivia or extraneous details. It can mean providing both absolute changes and percentage changes; reporting the mean, median, and mode to fully represent the distribution of data; providing multiple measures of an attitude or behavior; categorizing data more than one way to see what differences those categorical distributions make; providing information about mean, range, and standard deviations (represented as straightforward and understandable confidence limits); presenting both positive and negative quotes from interviewees; and finding ways to show the same thing in more than one way to increase understanding.

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, on families, and on 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 such as *farmer* turns out to be enormously variable in its use and definition. When does the weekend gardener become a farmer, and when does the large commercial farmer become an "agribusinessperson?" A whole division of the Census Bureau wrestles with these problems.

Defining *family* is no less complex. There was a time, not so long ago, when Americans may have shared a common definition of *family*. Now there is a real question about who has to be together under what arrangement before we call them a family. Single-parent families, foster families, same-sex "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 different 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.

During the 2000 presidential campaign of George W. Bush, Houston School Superintendent Rod Paige was given credit for "The Texas Miracle," reducing the school system's once-high dropout rate to just 2 percent. Once elected, President Bush named Paige to be Secretary of Education and the Houston's reforms became the basis for the President's "No Child Left Behind" education reform act. It turned out, however, that the celebrated reduction in school dropouts was achieved by redefining what a dropout was and coding dropouts as leaving for acceptable reasons, for example, going on to other things or moving to another school. Independent and external calculations put Houston's true dropout rate somewhere between 25 and

Defining and Measuring "Abnormal Sex"

Definitions matter. They determine results. A study published by the National Federation of Decency measured the decadent content of a daytime television "talk show." One of the categories of analysis included programs that encouraged "abnormal sex." The author of the report later acknowledged that it was probably a bit excessive of the federation to have included breast feeding in this category (Boulder *Daily Camera*, September 30, 1981:2). But, then, definitions of abnormal sex do seem to vary somewhat. Any reader of a research report on the subject would be well advised to look with care at the definition used by the researcher. Of course, any savvy evaluator involved in such a study would be careful to make sure that his or her own sexual practices were categorized as normal!

50 percent (still quite an error range), never anywhere near 2 percent (CBS 2004).

Such examples are not meant to make people cynical about statistics. Many distortions of this kind are inadvertent, due to sloppiness of thinking, unexamined assumptions, or hurrying to complete a final report. Sometimes, of course, they're the result of incompetence or unscrupulousness as asserted in the old adage that "figures lie, and liars figure." Widespread skepticism about statistics 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. A Sufi story reinforces the importance of being clear about definitions before drawing conclusions.

The wise fool Mulla Nasrudin and a friend went to the circus together. They were dazzled by the tightrope walker. Afterwards, Nasrudin's friend kept raving about the performance of the tightrope walker. Nasrudin tired of the conversation, but his companion resisted all attempts to change the subject. Finally, in frustration, Nasrudin asserted, "It wasn't really such a great feat as all that. I myself can walk a tightrope."

Angry at Nasrudin's boasting, the friend challenged him with a substantial wager. They set a time for the attempt in the town center so that all the villagers could be witness. At the appointed hour Mulla Nasrudin appeared with the rope, stretched it out on the ground, walked along it, and demanded his money.

"But the tightrope must be in the air for you to win the wager!" exclaimed the companion.

"I wagered that I could walk a tightrope," replied Nasrudin. "As everyone can see, I have walked the tightrope."

The village judicial officer ruled in Nasrudin's favor. "Definitions," he explained to the assembled villagers, "are what make laws."

They also make evaluations.

Make Comparisons Carefully and Appropriately

> *Noncomparative evaluations are comparatively useless.*
>
> —Michael Scriven (1993:58)

Virtually all evaluative analysis ends up in some way being comparative. Numbers in isolation, standing alone without a frame of reference or basis of comparison, seldom make much sense. A recidivism rate of 40 percent is a 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 U.S. GAO believed that 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 contrived. Working with users to select appropriate comparisons involves considering a number of options. Menu 13.1 presents 10 possibilities plus combinations. Evaluators should work with stakeholders to decide which comparisons are appropriate and relevant to give a full and balanced view of what is happening in the program.

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 miles, 4 miles, 5 miles, or

MENU 13.1

Menu of Program Comparisons

The outcomes of a program can be compared with

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, e.g., those known to be exemplary models (a purposeful sample comparison, Patton 2002a:230–34)
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, e.g., 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.

10 miles a week. Compared with 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 with all the people who don't run at all. On days when they need some incentive to push harder, they may compare themselves with people who run twice as far as they do. Some adopt medical standards for basic conditioning, something of 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 and hours. Some compare themselves with 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.

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

Interpretations

> *Sometimes something historical gives you a better perspective. You can see the latest dumbness as just the end of a long line of dumbnesses that have been taking place for thousands of years.*
>
> —Cartoonist J. B. Handelsman
> (quoted by Franklin 2007:27)

We have been discussing how to analyze and organize data so that primary intended users can engage the evaluation findings. The discussion has included focusing the analysis, clarity of presentation, striving for balance, being clear about definitions, and making comparisons carefully and appropriately. These are all elements of providing intended users with understandable and credible findings that can be interpreted. *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.*

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. Science fiction author and futurist H. G. Wells anticipated the importance of making statistical thinking accessible to nonstatisticians when he observed, "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." For evaluation users, that day is now. Incorporating a training perspective into evaluation (process use) will mean being prepared to help users with statistical reasoning. The logic of inductive qualitative analysis also needs to be made accessible to stakeholders.

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 interocular significance.* Fred Mosteller, the great applied statistician, was fond of saying that he did not care much for statistically significant differences. He was more interested in interocular differences, *the differences that hit us between the eyes* (Scriven 1993:71).

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." As Daniel Patrick Moynihan, a former United States Senator and distinguished social scientist, was fond of saying, "Everyone is entitled to his own opinion, but not his own facts." Once there is understanding of the findings, the evaluator facilitates interpretation by having participants in the process elaborate possibilities and options. 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 in this chapter) follow analysis and interpretations.

> Everyone is entitled to his own opinion,
> but not his own facts.
>
> Daniel Patrick Moynihan (1927–2003)
> U.S. Senator and
> distinguished social scientist

While this kind of facilitation usually occurs with a small number of primary users, the process can be facilitated for very large groups. The following example involved more than 200 people in a half-day process of analysis, interpretation, judgment, and generating recommendations—moving back and forth between small groups and full session reporting and adopting conclusions.

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 with community corrections staff, welfare department staff, court services staff, and members of the county board. That meeting included some 40 of the most powerful elected and appointed officials in the county as well as another 160 field professionals.

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 had data on how variations in recidivism, runaway rates, and juvenile attitudes varied with different kinds of group home environments. We had measured variations in homes with a 56-item instrument. Factor analysis of 56 items uncovered a single major factor that explained 54 percent of the variance in recidivism, with 19 items loading above .45 on that factor. *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 identified information users.* We focused the half-day work session on this issue.

The session began with a brief description of the evaluation's methods, and then the results were distributed. In randomly assigned groups of four, these diverse stakeholders were asked to look at the items in Exhibit 13.5 and *label the factor or theme represented by those items in their own words*. After the groups reported their labels, discussions followed. A consensus emerged around the terms *participation and support* as representing one end of the continuum and *authoritarian and nonsupportive* for the other end. We also asked the groups to describe the salient elements in the factor. These descriptions were combined with the labels chosen by the group. The resulting conceptualization—as it appeared in the final evaluation report—is shown in Exhibit 13.6.

EXHIBIT 13.5

Composition of the Group Home Treatment Environment Scale

The items that follow are juvenile interview items that are highly interrelated statistically in such a way that they can be assumed to measure the same environmental factor. The items are listed in rank order by factor loading (from .76 to .56 for a six-factor alpha solution). This means that when the scales were combined to create a single numerical scale the items higher on the list received more weight in the scale (based on factor score coefficients).

From your perspective, what underlying factor or theme is represented by the combination of these questions? What do these different items have in common?

1. The [group home parent's names] went out of their way to help us.

almost always	30.9%	
a lot of times	10.9%	
just sometimes	34.5%	
almost never	23.6%	*Factor loading = .76*

2. At the foster group home, personal problems were openly talked about.

almost always	20.0%	
a lot of times	9.1%	
just sometimes	32.7%	
almost never	38.2%	*Factor loading = .76*

3. Did you feel like the group home parents tried to help you understand yourself?

almost always	23.6%	
a lot of times	29.1%	
just sometimes	23.6%	
almost never	23.6%	*Factor loading = .74*

4. How often did your foster parents take time to encourage you in what you did?

almost always	27.3%	
a lot of times	20.0%	
just sometimes	30.9%	
almost never	21.8%	*Factor loading = .73*

5. At the foster home, how much were you each encouraged to make your own decisions about things? Would you say that you were . . .

almost always	18.9%	
a lot of times	30.2%	
just sometimes	30.2%	
almost never	20.8%	*Factor loading = .68*

6. How often did the foster parents let you take responsibility for making your own decisions?

almost always	23.6%	
a lot of times	20.0%	
just sometimes	25.5%	
almost never	30.9%	*Factor loading = .67*

7. We really got along well with each other at the foster home.

almost always	23.6%	
a lot of times	29.1%	
just sometimes	32.7%	
almost never	14.5%	*Factor loading = .66*

8. Would the group home parents tell you when you were doing well?

almost always	30.9%	
a lot of times	10.9%	
just sometimes	29.1%	
almost never	9.1%	*Factor loading = .64*

9. How often were you allowed to openly criticize the group home parents?

almost always	14.8%	
a lot of times	7.4%	
just sometimes	24.1%	
almost never	53.7%	*Factor loading = .59*

(Continued)

(Continued)

10. How much of the time would you say there was a feeling of “togetherness” at the foster home?

almost always	27.3%	
a lot of times	23.6%	
just sometimes	32.7%	
almost never	16.4%	*Factor loading = .59*

11. How much did the foster parents help you make plans for leaving the group home and returning to your real home?

almost always	9.1%	
a lot of times	21.8%	
just sometimes	21.8%	
almost never	47.3%	*Factor loading = .58*

12. How often would the foster parents talk with you about what you’d be doing after you left the group home?

almost always	7.3%	
a lot of times	18.2%	
just sometimes	36.4%	
almost never	38.2%	*Factor loading = .58*

13. How much of the time did the kids have a say about what went on at the foster home?

almost always	13.0%	
a lot of times	29.6%	
just sometimes	27.8%	
almost never	29.6%	*Factor loading = .56*

14. How much were decisions about what you all had to do at the group home made only by the foster parents without involving the rest of you?

almost always	30.9%	
a lot of times	18.2%	
just sometimes	32.7%	
almost never	18.2%	*Factor loading = .56*

15. How much of the time were discussions at the foster home aimed at helping you understand your personal problems?

almost always	23.6%	
a lot of times	23.6%	
just sometimes	18.2%	
almost never	34.5%	*Factor loading = .56*

EXHIBIT 13.6

Group Home Treatment Environment Continuum: Description of Group Home Ideal Types

Supportive-Participatory Foster Homes	*Nonsupportive-Authoritarian Foster Homes*
In group homes nearer this end of the continuum juveniles perceive group home parents as helpful, caring, and interested in them. Juveniles are encouraged and receive positive reinforcement. Juveniles are involved in decisions about what goes on in the home. Kids are encouraged to make their own decisions about the things they do personally. There is a feeling of togetherness, of being interested in each other, of caring about what happens now and in the future. Group home parents discuss the future with the kids and help them plan. There is a feeling of mutual support, and kids feel that they can openly express their feelings, thoughts, problems, and concerns.	In group homes nearer this end of the continuum, juveniles report that group home parents are less helpful, less open with them, and less interested in them personally. Juveniles are seldom encouraged to make their own decisions, and the foster parents tend to make decisions without asking their opinions about things. There isn't much planning things together or talking about the future. Kids are careful about what they say, are guarded about expressing their thoughts and feelings. Kids get little positive reinforcement. There is not much feeling of togetherness, support, and mutual caring; group home parents keep things well under control.

NOTE: *The descriptions presented here are based on stakeholders' interpretations of the factor analysis in Exhibit 13.5.*

EXHIBIT 13.7

Relationship between Different Home Environments and Recidivism

	No Recidivism	*Recidivism*	*Total*
Supportive-participatory homes	76% (*N* = 19)	24% (*N* = 6)	100% (*N* = 25)
Nonsupportive-authoritarian homes	44% (*N* = 11)	56% (*N* = 14)	100% (*N* = 25)

Primary intended users were asked to interpret what this table meant. This table was used in conjunction with Exhibits 13.5 and 13.6.

NOTE: Correlation $r = .33$; significant at .009 level.

The groups then studied accompanying tables showing the relationships between this treatment environment factor and program outcome variables (see Exhibit 13.7). The relationships were statistically significant and quite transparent. Juveniles who reported experiencing more supportive-participatory corrections environments had lower recidivism rates, lower runaway rates, and more positive attitudes. Having established the direction of the data, we discussed the limitations of the findings, the 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-participatory 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 four 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 here about a utilization-focused approach to findings. First, nonresearchers can understand and interpret data when presented with clear, readable, and simplified 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 a 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 do evaluators 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 a 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 or have to wait for a written report.

Making Causal Inferences: Attribution and Contribution

> *Water floats a ship; water sinks a ship.*
>
> —Chinese proverb
>
> *Causal attribution lifts an evaluation; causal attribution sinks an evaluation.*
>
> —Halcolm

Using social science findings generally and evaluation findings specifically means interpreting the significance and relevance of findings—and this typically includes some inference about causality. The extent to which an intervention can be said to have caused observed outcomes is one of the crucial interpretation issues in evaluation. "If an 'outcome' is not caused by a program, it is not an outcome at all; it's a coincidence. Coincidences cannot legitimately be documented as though they are outcomes; some evidence of a causal link is essential" (Davidson 2006b:1).

Chapter 12, on alternative paradigms, explored the debate about whether experimental designs with randomized control groups as counterfactuals are the gold standard for establishing causality (see especially Exhibit 12.6). Since the conditions for implementing high-quality experimental designs limit their applicability in evaluation, we need other ways of dealing with attribution. As we do so, a central question is *what level of evidence is needed.* Evaluation is meant to inform action and decision making. What degree of certainty is needed by primary intended users to use the evaluation findings? Research aims to prove. Definitive proof being elusive under the real-world time and resource constraints of decision making, evaluation more often aims to improve and inform.

An example from Chapter 12 is relevant 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 percent of children have worms, making 70 percent 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: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 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? Is such evidence sufficient to inform decision making? *Primary intended users must ultimately answer these questions.* The evaluator facilitates this deliberative process by illuminating the strengths, weaknesses, and implications of data and design options. Attention to degrees of evidence in relation to the stakes involved for decision makers further informs these deliberations and negotiations (Chatterji 2007).

Direct inquiry into the relationship between worm medicine and school attendance 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.

Michael Scriven has called a related form of causal tracing the *modus operandi method.* This language comes from detective

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.

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 2005:75). 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 of 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). 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.

In working with primary intended users, it can be quite useful to distinguish between attribution analysis and contribution analysis. John Mayne (2007b) 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?

A study in the *British Medical Journal* by Smith and Pell (2003) found that

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, commonsense 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.

SOURCE: http://bmj.bmjjournals.com/cgi/content/full/327/7429/1459?ck=nck

- 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. Outcome Mapping (IDRC 2007) provides a conceptual framework for mapping contributions in complex and dynamic environments with collaborating partners where simple notions of linear attribution are neither meaningful nor accurate (Iverson 2003). Exhibit 13.8 elaborates contribution analysis.

Program Theory and Realist Evaluation Explanations

Program theory can and should play a central role when interpreting evaluation findings. The program's theory of change should specify the expected attribution chain from activities to outcomes and impacts. A map of the anticipated contributions of collaborating programs to desired outcomes also constitutes a theory of change for multidimensional interventions in complex environments. A well-conceived program theory (see Chapter 10) provides a solid basis for interpreting evaluation findings and assessing causal claims (Rogers forthcoming, 2008, 2007a, 2007b, 2005c, 2003, 2000a, 2000b; Davidson 2000). Indeed, program theory is indispensable for testing and interpreting causality. Moreover, and this is quite important, a revised program theory is often a primary product of a theory-based evaluation thus fulfilling a knowledge-building purpose (Mason and Barnes 2007). This theory-testing and knowledge-building function can support and enhance both formative and summative evaluation purposes but should be understood as distinct from them as we "unbundle" and focus intended evaluation uses (Alkin and Taut 2003).Different approaches to theory-building, theory-testing, and explanation make quite a difference in how evaluations are designed and how findings are interpreted (Blamey and Mackenzie 2007). Experimental and quasi-experimental designs frame the explanatory issue as determining whether the program produced the observed outcomes controlling

EXHIBIT 13.8

Contribution Analysis

Contribution analysis (Mayne 2007, 2001) examines a postulated theory of change against logic and evidence to test the theory of change including examining other potentially influencing factors that could explain observed results. The overall aim is to reduce uncertainty about the contribution the program is making to the observed results.

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 as emphasized in Outcome Mapping (IDRC 2007), which also uses the language of contribution rather than attribution in looking at what various collaborating partners contribute to outcomes.

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) that 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,
- a discussion of the roles of other influencing factors, and
- a discussion of the quality of the evidence provided, noting weaknesses.

for context. In contrast, *realist evaluation* makes understanding context a primary basis for explaining variations in outcomes. *Realist evaluation* begins with skepticism about the generalizability of intervention effects across participants and contexts and seeks, instead, to understand and explain *what works for whom in what ways through what mechanisms in what contexts*. This is, fundamentally, a different question than whether aggregate outcomes can be attributed to the program without regard to variations in participants and contexts. *Realist evaluations* view programs as theories that, once actually implemented, are embedded in open social systems and must be understood in interaction with and in the context of the systems within which they operate. *Realist evaluation* doesn’t treat the overall program as the intervention to be tested but rather looks for the actual *mechanisms* that elucidate and explain what it is about programs and interventions that produce observed outcomes. In this way, *realist evaluation* is attentive to both intended and unintended outcomes, and seeks to test a "context-mechanism-outcome pattern configuration" (Pawson and Tilley 2005:365). *Realist evaluation* critiques experimental designs as overly simplistic and overgeneralized, seeking instead more nuanced and context-specific explanations (Blamey and Mackenzie 2007; Greenhalgh, Kristjansson, and Robinson 2007; Pawson and Tilley 2005, 1997; Pawson 2002a, 2002b; Mark, Henry, and Julnes 2000).

To Explain or Not to Explain?

Both attribution analysis and contribution analysis involve explaining the findings. Each attempts to interpret and explain the relationship between the program intervention and observed outcomes. Theory of change approaches make explaining causality a primary evaluation function and responsibility. Scriven (2007) cautions that offering such explanations may exceed the evaluator's responsibility and, more importantly, the evaluator's competence. He argues that the purpose of program evaluation is to determine the effects of the program, if any, and judge the program's merit, worth, or significance. It is *not* the task of the evaluator to *explain* these effects, he insists, other than by showing that there *are* effects of the program. Whatever effects are documented may lead to questions about why those effects occurred (or why desired effects did not occur), and the evaluator may or may not be able to help with this task, but it's a completely different task, a task of explaining a phenomenon, not of evaluating it. Scriven argues that explaining causation requires the specialized substantive knowledge of people who are professionally expert in the program's arena of focus. The evaluator is not likely to be one of these, he suggests, and does not acquire that expertise by evaluating one or several programs in that field, any more than a radiologist who is an expert at telling whether the patient has or has not got cancer becomes an expert about what causes cancer as a result of acquiring radiology skills. Scriven acknowledges the allure of offering explanations but thinks this temptation is due to confusion between the role of the scientist and the role of the evaluator. Scientists explain how the world works. Evaluators judge whether a program works. The reason why this is not just a semantic quibble, he insists, is that since trying to find out why a program works is sometimes more than *anyone* has so far been able to do, even scientists expert in the field, for evaluators to think they can do it will sidetrack the often vital task of finding out *whether* the program works, which, he emphasizes, is the primary obligation—and vital service—of the evaluator.

A utilization-focused evaluator can solve this problem, at least in part, by supporting and facilitating primary intended users, including program staff and substantive experts, to engage in the process of 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 doing so is deemed important.

Making Claims

The level of evidence needed in an evaluation challenges the evaluator to determine just what kinds of claims are appropriate and what level of certainty is required to make findings useful (Davidson 2000; Smith 1987, 1982, 1981). One way of meeting this challenge is to engage with primary stakeholders, especially program 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: (1) participants like the program, (2) participants get jobs as a result of the program, (3) the dropout rate is low, (4) changes in participants last over the long term, (5) the program is cost-effective, (6) the program does not work well with people of color, etc. Having generated a list of possible claims, I then have them sort the claims into the categories (or cells) shown in Exhibit 13.9. 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.

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 understand and engage the findings, (2) facilitating interpretation, (3) facilitating 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" depends on what values inform that judgment. If one supports gender equity or racial integration, it is good. If one opposes gender equity or racial integration, the findings are 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 1994, 1991a, 1967). 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, 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

EXHIBIT 13.9

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 a Claim of MAJOR IMPORTANCE

- Involves 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, accomplishes 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 a STRONG CLAIM

- 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 (pre-, post-, follow-up)
  With other groups
  With general trends or norms

(Continued)

(Continued)

- 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.

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.

Related Distinctions

1. Program *premises* are different from but related to and dependent on program *claims.*

Premises are the basic assumptions on which a program is based, for example, that effective, attentive parenting is desirable and more likely to produce well-functioning children who become well-functioning adults. This premise is based on research. The program cannot “prove” the premise (though supporting research can and should be provided). The program’s claims are about the program’s actual implementation and concrete outcomes, for example, that the program yielded more effective parents who are more attentive to their children. The program does not have to follow the children to adulthood before claims can be made.

2. Evidence is different from claims—but claims depend on evidence.

Claim: This program trains welfare recipients for jobs and places them in jobs, and, as a result, they become self-sufficient and leave the welfare rolls.

Evidence: Numbers and types of job placements over time; pre-, post-, and follow-up data on welfare status; participant interview data about program effects; employer interview data about placements; and so on.

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 (Chapter 6), I can move back and forth between the roles of evaluation facilitator and evaluator. In doing so, 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 both 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).

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

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.

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 reporting session, under enormous time pressure. In our follow-up study of federal health evaluations, we asked 20 decision makers about the usefulness of the recommendations they had received. The following reactions provide a flavor of typical reactions to recommendations:

- I don't remember the specific recommendations.
- The recommendations weren't anything we could do much with.
- It was the overall process that was useful, not the recommendations.
- I remember reading them, that's about all.
- The recommendations looked like they'd been added as an afterthought. Not impressive.

Useful and Practical Recommendations: Ten 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. *The focus of recommendations should be negotiated and clarified with stakeholders and evaluation funders as part of the design.* Not all evaluation reports include recommendations. The kinds of recommendations to be included 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 and important to distinguish between recommendations that can be implemented immediately, those that can be implemented in the short term (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 program 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.* A major source of frustration for many decision makers is that the recommendations in evaluation reports relate mainly to things 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, with the evaluator getting sidetracked on the issue of housing patterns and never getting back to concrete recommendations for the school? Of course, the best way to end up with recommendations that focus on manipulable variables is to make sure that, in conceptualizing the evaluation, the focus was on the 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, 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.

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,* 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.

Controversy about Recommendations

An evaluation without a recommendation is like a fish without a bicycle.

—Michael Scriven (1993:53)

While evaluators such as Mike Hendricks and Elizabeth Handley (1990) have argued that "evaluators should almost always offer recommendations" (p. 110), Michael Scriven has disagreed. Earlier I noted Scriven's insistence on distinguishing rendering judgments from offering explanations (only the former being a core evaluative responsibility). In a similar vein, he has been vociferous in warning evaluators against the "logical fallacy" of thinking that judging the merit or worth of something leads directly to recommendations. He considers it one of the "hard-won lessons in program evaluation" that evaluators seldom have the expertise to make recommendations and that they are generally well advised to stop at what they are qualified to do: render judgment.

> It is widely thought that program evaluations should always conclude with a recommendations section, but this view is based on a misunderstanding of the logic of evaluation, and the misunderstanding has seriously unfortunate effects. The conclusion of an evaluation is normally a statement or set of statements about the merit, worth, or value of something, probably with several qualifications (for example, These materials on planetary astronomy are probably the best available, for middle-school students with well-developed vocabularies). There is a considerable step from the conclusion to the recommendations (for example, You should buy these materials for this school), and it is a step that evaluators are often not well-qualified to make. For example, in teacher evaluation, an evaluator, or, for that matter, a student, may be able to identify a bad teacher conclusively. But it does not follow that the teacher should be fired or remediated or even told about the result of the evaluation (which may be informal). In making one of those recommendations, the evaluator must have highly specific local knowledge (for example, about the terms of the teacher's contract, the possibility of early retirement, and temporary traumas in the teacher's home life) and special expertise (for example, about the situation), both of which go a long way beyond the skills necessary for evaluation. If the evaluator is looking at recommendations aimed not at actions but at

> improvement (for example, suggested changes in the way in which the teacher organizes the lesson and changes in the frequency of question-asking), then he or she moves into an area requiring still further dimensions of expertise. (Scriven 1993:53)

While Scriven's counsel to avoid making recommendations if one lacks expertise in remediation or design is wise as far as it goes, he fails to take the added step of making it part of the evaluator's responsibility to seek such expertise and facilitate experts' engagement with the data. Utilization-focused evaluation does offer a way of taking that extra step by actively involving 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 very expertise Scriven says is necessary for informed action.

Moreover, in recent years, in response to engagement with clients and what they want, Scriven (2007) has had a change of heart. He has acknowledged that clients want recommendations and he recommends doing a simulation to show they are realistic. He also recommends following up on what has happened to evaluations in long-term relationships.

> We're finishing our third year of doing an impact evaluation of the overseas efforts of a large international aid charity. Each year we've made recommendations. In the plans for the fourth year, it seems to me we really should include a minor study of the impact of our prior recommendations. . . . Moreover, I think this should be standard operating practice for all continuing evaluation relationships. I also think this will have considerable impact, so you should be careful about doing it. Some organizations will be very nervous about having anyone check on whether they actually use the evaluations/recommendations they commission and some evaluators will not be too keen to have past recommendations dragged back into the light of day, since some of them will look less plausible now than they did at the time, and others will look even more vacuous than they did then (both possibilities strike me as good reasons for doing this, so that you can refine them and admit errors). (Scriven 2007:1)

A Futures Perspective on Recommendations

> *Show the future implications of recommendations.*
>
> —Hendricks and Handley (1990:114)

Recommendations have long struck me as the weakest part of evaluation. We have made enormous progress in ways of studying programs, in methodological diversity, and in a variety of data-collection techniques and designs. The payoff from those advances often culminates in recommendations, but we have made comparatively less progress in how to construct useful recommendations. I have found that teaching students how to go from data, interpretation, and judgment to recommendations is often the most challenging part of teaching evaluation. It's not a simple, linear process. A common complaint of readers of evaluation reports is that they cannot tell how the evaluators arrived at their recommendations. Recommendations can become lengthy laundry lists of undifferentiated proposals. They're alternatively broad and vague or pointed and controversial. But what recommendations always include, usually implicitly, are assumptions about the future.

The field of futures studies includes a broad range of people who use a wide variety of techniques to make inquiries about the nature of the future. Futurists study the future in order to alter perceptions and actions in the present. Evaluators, on the other hand, study the past (what programs have already done) in order to alter perceptions and actions in the present. In this sense, then, both futurists and evaluators are interested in altering perceptions and actions in the present, the impact of which will be a changed future. Evaluators do so by looking at what has already occurred; futurists do so by forecasting what may occur.

In effect, *at the point where evaluators make recommendations, we become futurists*. Recommendations constitute a forecast of what will happen if certain actions are taken. These forecasts are based on our analysis of what has occurred in the past. The accuracy of such forecasts, as with any predictions about the future, is subject to error due to changed conditions and the validity of assumptions that are necessarily made. Futurists have developed approaches for dealing with the uncertainties of their forecasts. Some of these approaches, I think, hold promise for evaluation. For example, futurists have developed techniques for constructing alternative scenarios that permit decision makers to consider the consequences of different assumptions and trends. These are variations on "if → then . . ." constructions. There are often "three to four different scenarios constructed: a pessimistic scenario, an optimistic scenario, and one or two middle-of-the-road or most likely-case scenarios.

The very presentation of scenarios communicates that the future is uncertain and that the way one best prepares for the future is by preparing for a variety of possibilities. General Robert E. Lee is reputed to have said, "I am often surprised, but I am never taken by surprise." That is the essence of a futures perspective—to be prepared for whatever occurs by having reflected on different possibilities, even those that are unlikely.

The advantage of scenarios in evaluation presentations is threefold. First, they permit us to communicate that recommendations are based on assumptions and thus, should those assumptions prove unwarranted, the recommendations may need to be altered accordingly. Second, the presentation of scenarios directs attention to those trends and factors that should be monitored so that as future conditions become known, program actions can be altered in accordance with the way the world actually unfolds (rather than simply on the basis of how we thought the world would unfold). Third, they remind us, inherently, of our limitations, for the "results of a program evaluation are so dependent on the setting that replication is only a figure of speech; the evaluator is essentially an historian" (Cronbach et al. 1980:7).

Communicating Evaluation

Communication is part of all program evaluation activities. Indeed, it is probably not an exaggeration to say that evaluation without communication would not be possible.

The very conduct of an evaluation is, itself, communication.

The evaluator needs to draw on all the resources available to him or her to consider the context of the program, to encourage processes that will animate reporting (that is, that make the results come alive), and to consider alternative forms of communication, appreciating the fact that multiple forms of presentation may increase the likelihood that a larger audience will be reached, realizing that what clicks with one stakeholder may not click with another.

—Marv Alkin, Tina Christie, and Mike Rose
(2006:384, 401–402).

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

This chapter has reviewed and discussed the four elements in a comprehensive framework for engaging findings: analysis, 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 13.10 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. With this integrating framework in mind, the concluding sections of this chapter discuss additional factors that can increase the utility of evaluation reporting.

Utilization-Focused Reporting

In utilization-focused evaluation, use does not center on the final report. Traditionally, evaluators and users have viewed the final written report as the climax—the end of the

EXHIBIT 13.10

Putting It All Together: Analysis, Interpretation, Judgment and Reporting

Evaluation of employment training program for high school drop-outs

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 drop-outs who were chronically unemployed

- 47% dropped out during the first 6 months this year (45 of 95) compared with a 57% dropout rate in the same period the previous year.
- The dropout 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 dropout rate is in line with other programs. The program attained these results at a time when the economy was sluggish and unemployment was somewhat higher than the historical average for this reason. 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).

(Continued)

(Continued)

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 two 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.

evaluation—and the key mechanism for use. From an academic perspective, use is achieved through dissemination of a published report. Moreover, use often doesn't emerge as an issue until there is something concrete (a report) to use. In contrast, *utilization-focused evaluation is concerned with use from the beginning, and a final written report is only one of many mechanisms for facilitating use.* The Minnesota Group Home Evaluation reviewed earlier illustrates this point. Major use was under way well before the report was written, as a result of the half-day work session devoted to analyzing the results with major stakeholders. The final report was almost an anticlimax, and appropriately so.

The data from our study of federal health evaluations revealed that much important reporting is interpersonal and informal. In hallway conversations, over coffee, before and after meetings, over the telephone, and though 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 oral briefings, which, presented with thoughtful preparation and skill, can have an immediate and dramatic impact.

In all cases, reporting is driven by the 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. Where 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, and Piontek 1996), which means honed and adapted to achieving use by targeted users.

Report Menu

As with other stages in utilization-focused evaluation, the reporting stage offers a smorgasbord of options. Menu 13.2 displays alternatives for reporting format and style, content, contributors, and perspectives. As just noted, selecting from the menu is affected by the purpose of the evaluation (see Chapter 4). 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 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 13.2 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. For a comprehensive discussion of evaluation strategies for communicating and reporting aimed at "enhancing learning in organizations," see Torres, Preskill, and Piontek (2004).

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.
3. Organize and present findings to facilitate understanding and interpretation.
4. Avoid surprising primary stakeholders.
5. Prepare users to engage with and learn from "negative" findings.
6. Distinguish dissemination from use.

Let me elaborate each of these principles.

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. Use of the evaluation findings and processes is the desired outcome, not producing a report. A report is a means to an end—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 13.2 and we start talking 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

MENU 13.2

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

(Continued)

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 ownership of the report residing 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 against . . . pertinent standards, so that its conduct is appropriately guided and, on completion, stakeholders can closely examine its strengths and weaknesses" (Stufflebeam 2007:A11).

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.

Focus Reports on Primary Intended Users

A dominant theme running throughout this book is that use is integrally intertwined with users. That's the thrust of the personal factor (Chapter 3). The style, format, content, and process of reporting should all be geared toward *intended use by intended users*. For example, we've learned, in general, 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, put 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. 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. [EV367:12]

Organize and Present Findings to Facilitate Understanding and Interpretation

I emphasized this point earlier in this chapter (e.g., Exhibit 13.3), but it's worth repeating and re-emphasizing with another example. Michael Hendricks (1994, 1984, 1982) 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 13.11 shows. Mike Hendricks regularly 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 et al. 2004).

Avoid Surprising Stakeholders: Share Findings First in Draft Form

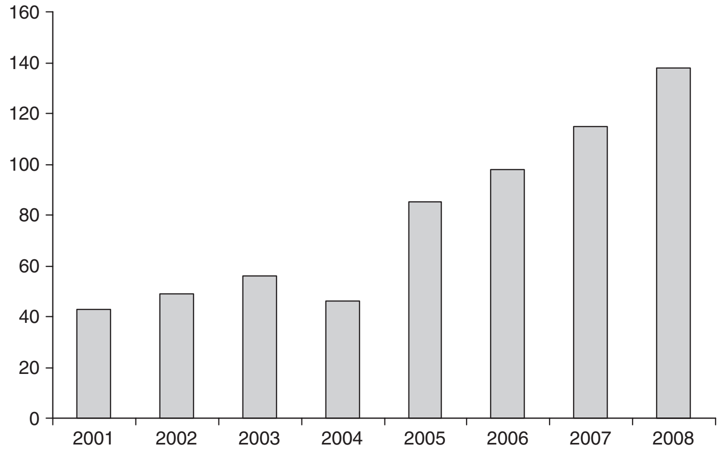
The story just told emphasizes the importance of sharing draft reports with primary users in time to let them shape the format of the final report. This doesn't mean fudging the results to make evaluation clients happy. It means focusing so that priority information needs get priority. Collaborating with primary users means that evaluators cannot wait until they have a highly polished final report prepared to share major 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. Formative feedback, in particular, is most useful as part of a process of thinking about a program rather than as a one-shot information dump. In the more formal environment of a major summative evaluation, surprises born of the public release of a final report are not going to be well received by important stakeholders caught unawares.

EXHIBIT 13.11

The Power of Graphics Data in a Table

2001	43 graduates
2002	49
2003	56
2004	46
2005	85
2006	98
2007	115
2008	138

The same data in graphic form



In our study of the use of federal health evaluations, we asked the following question:

> Some suggest that the degree to which the findings of a study were expected can affect the study's impact. Arguments go both ways. Some say that surprise findings have the greatest impact because they bring to light new information and garner special attention. Others say that surprises will usually be rejected because they don't fit in with general expectations. What's your experience and opinion?

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 that guarantee the relative predictability of the content of the final report.

> 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. [DM346:30–31]

The evaluator for this project expressed the same opinion: "Good managers are rarely surprised by the findings. If there's a surprising finding it should be rare. I mean, everybody's missed this insight except this great evaluator? Nonsense!" [EV364:13]. Surprise attacks may make for good war strategy, but in evaluation, the surprise attack does little to add credence to a study.

Prepare Users to Engage with and Learn from "Negative" Findings—and Think Positive about Negatives

John Sununu: (while Governor of New Hampshire in 1988, discussing the economy and upcoming presidential election): "You're telling us that the reason things are so bad is that they are so good, and they will get better as soon as they get worse?"

James A. Baker: (then President Reagan's Secretary of the Treasury): "You got it."

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. In the next chapter, on politics and ethics, we'll revisit this confrontation of perspectives. In this section, I want to make two points: (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.

The context for these two points is a general belief that most evaluations have negative findings. Howard Freeman (1977), an evaluation pioneer, expressed the opinion that the preponderance of negative findings diminished use. He recommended, somewhat tongue-in-cheek, that "in view of the experience of the failure of most evaluations to come up with positive impact findings, evaluation researchers probably would do well to encourage the 'biasing' of evaluations in the direction of obtaining positive results" (p. 30). He went on to add that evaluators ought to play a more active role in helping design programs that have some hope of demonstrating positive impact, based on treatments that are highly specific and carefully targeted.

Freeman's colleague Peter Rossi, coauthor of one of the most widely used evaluation texts (Rossi and Freeman 1993), shared the view that most evaluations show zero impacts on targeted clients and problems. He asserted, also tongue-in-cheek, that "only those programs likely to fail are evaluated." This led him to formulate Rossi's Plutonium Law of Evaluation: "Program operators will explode when exposed to typical evaluation research findings" (quoted in Shadish et al. 1991:386–87).

On the other hand, Michael Scriven (1991b) has observed presumably the same scene and concluded that evaluations display a "General Positive Bias" such that there is a "strong tendency to turn in more favorable results than are justified" (p. 175).

The problems I have with either stereotype, that most evaluations are negative or most are positive, is, first, they are impressions, not the result of any systematic analysis, and, second, they impose a dichotomous win/lose, pass/fail, success/failure, and positive/negative construct on evaluation results that, in my experience, actually display considerable diversity and balance. 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. 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.*

Consider these data. In our 20 federal health evaluation case studies, respondents described findings as follows:

Evaluator and decision maker independently agreed that the findings were:

Basically positive	5
Basically negative	2
Mixed positive-negative	7
Evaluator and decision maker disagreed in characterizing the findings	6

Our sample was not random, but it was as systematic and representative of federal evaluations as we could make it given the difficulty of identifying a "universe" of evaluations. Only 2 of 20 were basically negative; the most common pattern was a mix of positive and negative; and in 6 of 20 cases, the evaluator and primary decision maker disagreed about the nature of the judgment rendered. Moreover, in only one case did any of our respondents feel that the positive or negative nature of findings explained much about use. Because we encountered few summative decisions, the overall positive or negative nature of the evaluation was less important than how the findings could be used to improve programs. In addition, the positive or negative findings of a particular study constituted only one piece of information that fed into a larger process of deliberation and was interpreted in the larger context of other available information. Absolute judgments of a positive or negative nature were less useful than specific, detailed statements about levels of impact, the nature of relationships, and variations in implementation and effectiveness. This shifts the focus from whether findings are negative or positive to whether the evaluation results contain useful information that can provide direction for programmatic action.

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 (see Torres et al. 2004). Placing emphasis on organizational learning, action research, participatory evaluation, collaborative approaches, developmental evaluation, and empowerment evaluation—approaches discussed in Chapter 5—can defuse fear of and resistance to negative judgment.

As this discussion illustrates, when striving to get evaluations used, we often focus on overcoming resistance to findings, especially negative findings. The opposite problem, less often discussed, is overcoming apathy—or boredom. In Exhibit 13.12, experienced internal evaluator Gene Lyle offers advice about overcoming boredom when presenting routine performance data to the same intended users on a regular basis (e.g., annual reports).

EXHIBIT 13.12

Overcoming the Boredom of Repeat Presentations

As I was writing this chapter, Gene Lyle retired from a long and distinguished career as an internal evaluator in Ramsey County Community Human Services Department in Minnesota. The American Evaluation Association recognized Gene for his work by awarding him the Alva and Gunnar Myrdal Government Award in 2000. I asked Gene if, as an internal evaluator having presented findings many, many times over the years, there was some special evaluation reporting issue that he had never seen addressed in the literature. He quickly responded: "Boredom." I asked him to elaborate. Here, then, is the wisdom of experience speaking on the challenge of making evaluation presentations interesting.

On Boredom

Since internal evaluation often involves the management of systems that generate routine reports over time—performance measurement being an example—the evaluator often must present essentially the same data sets over and over again, often with little variation in outcomes. This can lead to a dulling of interest. Therein lies one of the skeletons in the internal evaluator's closet: boredom!

When data from a reporting system are first presented—we're pretty much talking formative evaluations here—there is usually curiosity, questioning, high interest, and motivation to explore the data. Let's say we're looking at data about the effectiveness of social work and foster care in returning children to their parental homes, and this information is being presented once every three months. The first two or three times people will be eager to see the results, study them, ask for more information than what may already be in the evaluation, and maybe even make decisions based upon the findings. But if the results do not change significantly over time, if they remain static (even if that's a good thing), they may become fainter and fainter blips on the intended users' radar screens.

How to deal with this eventuality? Here are some suggestions.

- Pre-review the results before meeting with reviewers. Anticipate questions they may ask based on the data and results.
- Present results in person if at all possible. Also, if possible, review results with the people who are actually and directly providing the service.
- Encourage dialog about the results.
- Do not overinterpret. Encourage the reviewers to interpret for themselves.

- When the review is finished, take time on your own to reflect on what you heard and saw. You may later recall suggestions or concerns that were overlooked in the review itself.
- Follow up on any recommendations for change to the evaluation itself. Negotiate with the reviewers to incorporate their suggestions.
- At the very least, do an annual review of the evaluation with the stakeholders to determine where it meets, and doesn't meet, their needs. For example, are the measures still appropriate? Do we need different data from different sources? Are the programs we're evaluating operating any differently now compared with when the evaluation was established? (But don't wait until an annual review to ask these questions and make changes if they are necessary.)

In the real world evaluations can only be sustained for a limited time in the format in which they were first developed. They will inevitably need to be changed within a year or two of their initiation. Within five years they will either be totally different or will have disappeared altogether. There are too many forces operating against static evaluations. Boredom is one of the indicators that the wise evaluator will use to monitor the need for change.

SOURCE: Reprinted with permission of Gene Lyle.

Finally, it's worth remembering, philosophically, that the positive or negative nature of evaluation findings can never be established with any absolute certainty. As Sufi wise-fool Mulla Nasrudin once advised, a heavy dose of humility should accompany judgments about what is good or bad. Nasrudin had the opportunity to render this caution at a teahouse. A monk entered and said:

> *"My Master taught me to spread the word that mankind will never be fulfilled until the man who has not been wronged is as indignant about a wrong as the man who actually has been wronged."*
> *The assembly was momentarily impressed. Then Nasrudin spoke: "My Master taught me that nobody at all should become indignant about anything until he is sure that what he thinks is a wrong is in fact a wrong—and not a blessing in disguise."*
>
> —Shah 1964:58–59

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 an 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, both intended and unintended (e.g., Kirkhart 2000). This includes instances where planned dissemina-

Studying Use

We have very good methods for gathering social science knowledge but considerably less good advice about how to put it to use. What we most need to study is not how to do social science *but how to use it.*

—Michael Scriven (2005e:78).

tion *hopes for* broader influence but can't be sure if or where this will occur. Exhibit 13.13 reminds us of the boundaries between intended use of both evaluation findings and processes and unintended influences that can accompany dissemination. In between is a gray area of hope and possibility that is less than fully intentional and more than laissez-faire unintentional use.

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 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

EXHIBIT 13.13

Matrix of Intentionality and Use/Influence

	Findings Use/Influence	*Process Uses/Influences*
Intended	Intended use by intended users	Includes explicit, planned evaluation capacity building as well as other process uses
Intended/unintended—gray area	Intentionality focused on primary intended users, but planned dissemination *hopes for* broader influence (though can't be sure if or where this will occur).	Evaluator facilitates the evaluation process to build capacity, but this is implicit and those stakeholders who are involved are motivated by and focused on findings use.
Unintended	Unplanned influence of findings beyond primary intended users—and even beyond original dissemination	Evaluation capacity building is implicit (an artifact of participation in the evaluation)

use for which I take responsibility: intended use by intended users. Dissemination is not use, though it can be useful.

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, and recommendations considered. 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. This chapter has emphasized that the challenges and excitement of analysis, interpretation, and judgment ought not to be the sole prerogative of evaluators. Stakeholders can become involved in struggling with data, too, increasing both their commitment to and understanding of the findings.

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 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.

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. In that spirit, Marvin Alkin (1990:148) suggested a tee shirt that user-oriented evaluators could give to intended users:

COME ON INTO THE DATA POOL

Follow-Up Exercises

1. Locate an evaluation report. Identify the major findings. Now imagine that you were working with primary intended users for the evaluation *before* the data were collected. (a) Fabricate simulated findings that could be used to prepare intended users for subsequent data engagement, for example, findings that are better and worse than those actually attained. Format the simulated data for presentation to intended users. What facilitation questions would you ask to help the intended users engage with the simulated data? (b) Construct standards of desirability that would guide intended users in interpreting the findings. What

facilitation questions would you have asked to help the intended users of this evaluation generate standards of desirability before data collection?

2. 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 Exhibit 13.2 and Exhibit 13.10). Give examples of these distinctions from the report. Comment on and critique the extent to which these distinctions are adhered to in the evaluation. Based on your reading and review of this report, how important do you believe it is to adhere to these distinctions? What are the challenges of adherence?

3. Identify a commonly used and widely cited statistic like the crime rate, inflation rate, unemployment rate, poverty rate, dropout rate, HIV infection rate, or divorce rate. Locate at least media reports where this statistic is used. Find out how it is defined. Identify alternative definitions, meanings, and uses. What issues would you raise about this statistic if primary intended users wanted to include it in an evaluation as a basis of comparison for a program outcome? Discuss how you would use such statistics in an evaluation and how you would discuss their use with primary intended users.

4. Locate an evaluation report that includes recommendations. Convert those recommendations into (a) at least three options for each recommendation (as if you were presented primary intended users with choices) and (b) three *futures-oriented scenarios,* one based on optimistic assumptions, one based on pessimistic assumptions, and one that assumes that the future will be much like the recent past. (Note: Items a and b are separate tasks.)

5. Using Menu 13.1 (Menu of Program Comparisons), use the actual outcome for a real program from that program's evaluation and discuss what comparisons were made. Construct comparisons using all the alternatives in Menu 13.1. (You may have to make up data for some comparisons.)

6. Using Menu 13.2, 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.