1

Introduction

Why a Focus on Validity?

Evaluation is often described as a practical discipline within the social sciences. It exists to answer questions. Other branches of the social sciences sometimes focus on discovering arcane, or abiding, truths about the world. Evaluation, by contrast, goes about trying to answer timely questions about social issues, social problems, programs and policies, and any number of things that people want to know.

Because the evaluation discipline is built around devising methods to answer pragmatic questions, its practitioners must place high priority on making evaluations both useful and accurate. Indeed, *utility* and *accuracy* are two of the four Program Evaluation Standards that have been developed by the Joint Committee on Standards for Educational Evaluation (Yarbrough et al., 2011). Of these two qualities, utility is the more encompassing one. For example, to be useful, evaluations need to understandable, perceived as relevant, and delivered on time, whether conducted inside or outside of a program or organization.

But clearly, to be useful, evaluations must also be accurate.

What Is Accuracy in Evaluation?

In an intuitive sense we might surmise that an *accurate* evaluation is one in which the technical aspects have been conducted with competence, resulting in a truthful and trustworthy representation of the program, policy, or other phenomenon under study. More specifically, we might conclude that the measures have been thoughtfully chosen, the data collection has proceeded competently, the analyses—quantitative, qualitative, or mixed—are thoroughly and expertly conducted, and the conclusions flow from those measurements and analyses. We might also suggest it means that the

measurements and analyses are free from intentional bias (that is, attempts to arrive at a foregone conclusion), and that both measurement bias and random error are minimized to the extent possible.

In a nutshell, we may consider the information in an evaluation *accurate* to the degree that we believe it reflects the world in a truthful way. Each stakeholder, if they are paying attention, will have their own views on how successfully this requirement has been fulfilled. The opinions of stakeholders may well vary, but in making their judgments they will be working from a shared body of evidence.

The question of accuracy in evaluation has been directly addressed and examined in the 3rd edition of the *Program Evaluation Standards*, which identifies eight specific standards that relate to this category. In general, these standards "are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated" (Yarbrough et al., 2011). Table 1.1 lists the Joint Committee's accuracy standards and their descriptions.

TABLE 1.1 ● The Accuracy Standards of the *Program Evaluation Standards* (3rd Edition, 2011)

Standard	Description
A1. Justified Conclusions and Decisions	Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences.
A2. Valid Information	Evaluation information should serve the intended purposes and support valid interpretations.
A3. Reliable Information	Evaluation procedures should yield sufficiently dependable and consistent information for the intended uses.
A4. Explicit Program and Context Descriptions	Evaluations should document programs and their contexts with appropriate detail and scope for the evaluation purposes.
A5. Information Management	Evaluations should employ systematic information collection, review, verification, and storage methods.

TABLE 1.1 ● *(Continued)*

Standard	Description
A6. Sound Designs and Analyses	Evaluations should employ technically adequate designs and analyses that are appropriate for the evaluation purposes.
A7. Explicit Evaluation Reasoning	Evaluation reasoning leading from information and analyses to findings, interpretations, conclusions, and judgments should be clearly and completely documented.
A8. Communication and Reporting	Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors.

Source: Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users.* Sage. (p. 157)

Inspection of the accuracy standards conveys an appreciation of all the decisions required to achieve a suitable level of accuracy and verisimilitude. One thesis of this book is that the search for accuracy in evaluation—to whatever degree that may be achieved—depends, in part, on the sum total of the quality of the decisions in the planning, conduct, and interpretation of the evaluation study. Other determinants of accuracy include the skillfulness with which those decisions are implemented and the fair and just interpretation of the information that is thereby produced.

Evaluation Questions and the Evaluation Plan

In most cases, an early goal as an evaluation gets established is to determine a formal set of *evaluation questions* to be addressed. Most of my attention in this book will be focused on how the evaluation plan reflects and addresses the established evaluation questions. To reflect the complexity of interventions, most evaluations are multifaceted, involving areas of complexity and attention to a variety of the program's components and often examining those components from a number of angles. As a natural consequence, the set of evaluation questions will usually reflect this complexity, resulting in multiple questions that touch on numerous distinct dimensions of the program.

The ways in which evaluation questions get established, and the relation of those questions to the conduct of the evaluation, is a subject that deserves close attention. The process can be political to a significant degree

because it depends on the sometimes diverging viewpoints of diverse stakeholders and other factors having to do with why the evaluation is being launched in the first place.

But once the evaluation questions are established, they provide a structure that guides the activities that follow. To be sure, an experienced evaluator will be alert to new knowledge and unexpected findings that arise from the evaluation process, even if they do not trace back to the themes of the questions. This process of serendipity and surprise can be invaluable—on occasion even the most lasting and significant contribution of the evaluation. But in the normal course of events, most of the evaluation plan and activities are geared toward addressing and answering the predetermined evaluation questions. A major concern of this book is that this process is conducted with accuracy, utility, and fairness.

The "evaluation plan" may be an actual document, which is usually most desirable. But there may be cases, especially when the program is small and the evaluation is undertaken for purposes of purely local utility, in which a formal, documented plan is lacking. Although it is far from an ideal scenario, it may happen that a program practitioner, called upon to evaluate their program as well as plan, conduct, fund, and otherwise ensure the program's existence, collects outcome data without any prior thought as to how those data will be analyzed. Nevertheless, if the evaluation is completed, at some point those analyses will have been conducted, presented, considered, and—with any luck—acted upon. In such cases the "plan" may be deduced from the activities and operations that have been carried out.

In this book, for purposes of simplicity, I will consider the existence of a distinct planning process, in which decisions are identified and represented. So whether the plan is a tangible document or a de facto set of decisions that must be reconstructed from what actually occurred—or some combination of the two—the evaluation plan is the focus of our attention and analysis for most of this book. In addition, the final two chapters discuss phases of the evaluation that occur in the course of implementing the plan: analyzing data, drawing conclusions, and making recommendations.

The evaluation planning decisions might involve the following areas, many of which will be examined in detail in later chapters:

- The determination of the primary evaluation questions to be answered
- The choice of primary outcomes
- The choice of measurement strategies to address those outcomes
- The choice of study participants

- The analytical design, whether based on quantitative or qualitative data, or a mixed-methods strategy
- The choice or development of instruments and other measures
- Timelines for tracking the outcomes
- The number of data points

Evaluations Sometimes Lose Their Way

Tolstoy's *Anna Karenina* famously begins: "Happy families are all alike; every unhappy family is unhappy in its own way" (Tolstoy, 1878/1961). In looking over the landscape of evaluations large and small, I am often reminded of this epigram and sometimes tempted to believe that every unsuccessful evaluation is unsuccessful for its own unique reasons. I am not sure that that is a defensible proposition, but it may serve to illustrate that there are a multitude of ways and reasons, frequently idiosyncratic, that evaluations fail to provide the understandings that are called for in their initial aims or in the original evaluation questions. An example of this phenomenon is provided in Case Study 1.1.

CASE STUDY 1.1 AN EVALUATION THAT WENT OFF-TRACK

In the 1990s I worked as an Extension 4-H Youth Development Specialist in University of California Cooperative Extension. In my capacity as a Program Evaluation Specialist, I advised a county 4-H program in southern California that had grant funds to hire an external evaluator to evaluate their 4-H afterschool program. I assisted my county faculty colleagues in drafting a Request for Proposals and provided some feedback in their team's selection of the successful proposer. Eventually they chose a fairly large research firm and awarded them the contract, which at $25,000 was one of the firm's smallest contracts. (It might be worth noting that $25,000 bought more in the 1990s than it does today.) The county 4-H team was interested in an outcome evaluation, and being responsive, the evaluation contractor proposed a prepost design. However, in the implementation of their design, they needed to deal with multiple challenges that are common to small afterschool programs, one of which is the lack of consistency in the attendance of participating children. If the time period between premeasurement and postmeasurement is long enough, this lack of consistency can result in highly attenuated, potentially biased, samples of children who have contributed scores at both time points.

The contractors began the evaluation. A few months into the process, they contacted the county 4-H team and me, to report that consistency in attendance was indeed a major impediment to their conduct of the evaluation as

(Continued)

(Continued)

planned. That news was not unexpected. The question was what could be done about it to arrive at some answers that could produce some degree of useful information however imperfect. As their preferred solution, the contractor suggested that the evaluation be changed to a series of focus group interviews, in which they could ask the afterschool staff, as well as the participating children and their parents, what they liked about the program. They provided no other options, based on the small budget. After some extended discussion the county team agreed, and that changed plan was followed.

When I heard this new proposal from the contractors I filed it in my memory bank as a glaring example of how an evaluation can lose its way. The evaluators steered toward a design that was possible and plausible, and that was economical within the constraints of the project budget. One might view all of these considerations as reasonable, under the circumstances. The one fly in the ointment is that their newly proposed design diverged radically from the original questions—which had reflected the needs of the program team—and, in essence, substituted a completely different set of evaluation questions, based on expediency. The evaluation was completed, and it was competently done in its way. The county team reviewed it briefly and didn't use it. For the purposes of their grant and their local information needs to gather evidence that might support program sustainability, their highest priority need was to assess program effectiveness. Thus, this evaluation must be considered a failure in achieving its aim of utility. Its implementation abandoned the original evaluation purposes.

In terms of losing sight of the original evaluation questions and losing the ability to answer those questions, Case Study 1.1 represents an example on a grand scale. But more subtle examples can also be cited. Case Study 1.2 describes a measurement-related decision that could have gone off-track with regard to what was being measured.

CASE STUDY 1.2 SEEKING A MEASURE OF CAREER AWARENESS IN ADOLESCENTS

Some years ago, in preparation for developing a youth development program on the topic of career development, I co-led a needs assessment study with Extension colleagues and graduate students, involving career development of high school seniors in northern California. We described our primary construct, *career awareness*, as follows:

> We use the term *career awareness* to refer to adolescents' reflectiveness about their future work lives. This may or may not include a specific career decision, in which a student can identify with certainty the career field that he or she wants to pursue. However, our definition does include

(Continued)

> students' active consideration of the elements upon which a decision can be based, such as understanding one's own talents and interests or understanding the opportunities and requirements of various career fields. (Braverman et al., 2002, p. 55)

To examine this construct, we used an established instrument called the Career Decision Profile (Jones & Lohmann, 1998), a 16-item scale with six subscales including decidedness, level of comfort regarding vocational decision status, understanding of one's own interests and abilities, and other characteristics. We analyzed students' scores on these subscales in relation to their grades, part-time work experiences, and demographic variables. From these analyses, we drew some conclusions about school opportunities for career preparation programs.

Our study, when completed, provided information that was considered valuable by the schools and our Extension colleagues at the University of California. But our team's path to understanding our variables was not straightforward and in fact the evaluation planning period proved to be a learning experience for us. We discovered a plethora of instruments relating to adolescents' vocational development, many of which varied, sometimes subtly, with respect to the variables on which we were focusing. Some of these identified constructs included career decision-making, career maturity, vocational identity, and career beliefs, among others. Thus, if we had accepted the first, or most convenient, instrument that turned up in our search, without intensive discussion and clarification of the specific variables that we wished to examine, we would have been misled.

Evaluation as Argument

Ernest House was an early proponent of the position that evaluation can be viewed as a process of argument, in which the evaluator presents evidence and builds a case toward a proposition or conclusion, e.g., regarding the value or effectiveness of a program. This position views evaluation as a particular form of inquiry and contrasts it with a view of evaluation as, for example, a fundamentally scientific endeavor. *Science* searches for enduring truths about the world—if not permanent truths, then truths that are sufficiently stable for the conditions of replicability to be determined and applied. An argument-based perspective, by contrast, relies on persuasion and the overall weight of evidence. House wrote:

> Evaluation aims at persuading a particular audience of the worth of something or that something is the case by an appeal to the audience's reason and understanding. For this purpose, uncertain knowledge is

> useful although the ideas themselves are always arguable. The appropriate methods are those of argumentation, which is the realm of the "credible, the plausible and the probable" rather than the necessary....In summary, evaluation persuades rather than convinces, argues rather than demonstrates, is credible rather than certain, is variably accepted rather than compelling. (House, 1980, p. 73)

A more recent exposition of this viewpoint is provided by Thomas Schwandt (2015):

> In its completed form, an evaluation is an argument for the value (or lack of value) of some particular program or policy. An argument is an attempt to persuade some particular audience of the conclusion about program value by giving reasons for accepting that conclusion as evident. The use of evidence in an argument demonstrates that evidence is not synonymous with data or information. Rather, evidence is data or information introduced at a specific point in an argument in order to persuade a particular audience of the veracity of the conclusion about program value. (p. 83)

Schwandt (2015) goes on to point out that an evaluation argument is notable in several respects. First, it does not purport to provide a proof of its evaluative proposition, but rather a credible, convincing case that can persuade a specific audience. Second, an evaluation argument is rooted in a particular context: it is concrete and immediate rather than abstract and universal. Third, Schwandt notes that such an argument is dialectical, meaning that it is part of an exchange of ideas. Fourth, the emphasis on persuasion as a goal requires that evaluation is concerned not just with accuracy and evidence but also, in part, with processes of communication, presentation, and rhetoric. Fifth, Schwandt notes that the argument itself can be the object of evaluation, in addition to the program or other evaluand.

The ideas and recommendations that I present in this book are consistent with the view of evaluation as a process of argument. My focus is on how evaluators can shape their evaluation plans with an eye toward building the most compelling, convincing case for the eventual interpretations of data and overall conclusions—well before data have been collected and certainly before those conclusions have been formulated.

The Role and Significance of Validity

The central concept in this book is validity, which relates squarely and directly to issues of quality, accuracy, and fairness in the evaluation process. Validity is a multifaceted construct that takes numerous forms, with a long history in the research tradition. As we will discuss more fully in Chapters 2 and 3, there are a number of perspectives as to what validity is, what it can

do, and how it can contribute to the quality of research and evaluation efforts. But in the majority of its incarnations, validity refers to the adequacy of interpretations, judgments, conclusions, inferences, and knowledge claims. Validity is interwoven with the act of drawing interpretations from measurements, study designs, analyses, and entire studies.

In the social sciences, validity had its origins in measurement theory and practice, as will be discussed in Chapter 2. But if validity is a property of an inference rather than a measure, then it can be applied to other kinds of inferences beyond simply the meaning of test scores. As we will see in Chapter 3, Donald Campbell and his colleagues, particularly Thomas Cook and William Shadish, expanded the applications of validity to the research process itself, through the development of a four-part validity typology that they applied to research designs. Those subtypes were *internal validity, external validity, construct validity, and statistical conclusion validity*. Construct validity was already well established, having been introduced by Cronbach and Meehl (1955). But these theorists examined its use for purposes of assessing the quality of experimental and quasi-experimental designs in addition to the measurement context.

A different approach to validity, and one directly applied to the practice of evaluation, was offered by House, who approached the examination of validity in a way that is more connotative than scientific. He argued that the validity of an evaluation rests on a set of considerations on which the evaluation can be assessed overall, beyond attention solely to research rigor. Thus, he proposed that the quality of an evaluation can be judged with respect to three characteristics: truth, coherence, and justice (more in Chapter 3). One of House's unique contributions was to propose that in the determination of validity for an evaluation, accuracy is just one leg of a three-legged stool, accompanied by effective framing and communication (coherence) and the advancement of social justice. To the extent that these three cornerstones are well represented in an evaluation, the evaluation will be valid or, in House's words, worthy of recognition.

A final theoretical conception of validity, also described in Chapter 3, is multicultural validity. This form of validity is closely tied to the construct of cultural competence in evaluation, which was described as a critical skill set for evaluation professionals by the American Evaluation Association in a 2011 public statement (AEA, 2011). Originally developed and proposed by Karen Kirkhart, multicultural validity refers to the accuracy and fairness of evaluative inferences and conclusions considered in light of the specific, intersecting cultural contexts that characterize any evaluation setting (Kirkhart, 2010). It reflects a recognition that knowledge claims cannot be made in the absence of cultural understandings, and the validity of any evaluation is affected by how successfully those cultural dimensions have been incorporated into all phases of the evaluation. Multicultural validity also encompasses the promotion of social justice in evaluation, in common with House's validity model, and places high priority on the consequences of

evaluations, including interpretations, conclusions, and actions that are taken in response to evaluation results.

Constructs and Construct Validity

Another central concept in this book is the understanding and analysis of *constructs*. A construct, as commonly understood in psychology, is an individual characteristic that exists only in a theoretical sense but can be useful in explaining patterns of activity, behavior, or abilities. General intelligence, reading ability, and mathematical aptitude are all constructs. So are all psychological traits, including the so-called big five traits of personality psychology (agreeableness, conscientiousness, emotional stability, extraversion, and openness; Poropat, 2009). Athletic ability is a construct, although running time in the 440-yard dash is an objective measure of running speed—at least as demonstrated on a particular day. In a psychometric framework, construct validity refers to the degree to which scores can be interpreted as providing information about the underlying theoretical construct. The validity theorist Bruno Zumbo (2009) observed: "In short, construct validity involves generalizing from our behavioral or social observations to the *conceptualization* of our behavioral or social observations in the form of the construct" (p. 68).

Outside of the realms of psychometrics and psychological theory, however, Shadish et al. (2002) argued that "... the creation and defense of basic constructs is a fundamental task of all science" (p. 65), and they defined *construct*, broadly and simply, as "a concept, model or schematic idea" (p. 506). They established the concept of construct validity as one component of the overall quality of research studies, and noted that the task of identifying, specifying, and describing relevant constructs applies not only to outcomes and variables but also to a study's units, settings, and treatments.

The specification of the constructs in an evaluation or research study is sensitive to larger contexts in ways that can be overlooked or that can rapidly change. An example is gender. Shadish et al., in describing the relationship between constructs and their conceptualization, wrote:

> It may help, however, to give examples of construct validity of persons, settings, and treatments. A few of the simplest person constructs that we use require no sophisticated measurement procedures, as when we classify persons as males or females, usually done with no controversy on the basis of either self-report or direct observation. But many other constructs that we use to characterize people are less consensually agreed upon or more controversial. (Shadish et al., 2002, p. 70)

Shadish et al. go on to describe the example of racial and ethnic identity as a type of construct, in contrast to gender, for which there is much less consensus and which presents challenges for determining the appropriate

categories, the labels for those categories, and the best measurement procedures. It is striking that in the two decades after the publication of that volume, we can immediately recognize how much the concept of *gender* has changed, becoming a highly complex topic in society, politics, and social relations, as well as in research practice. Social conceptions of gender are transforming with remarkable rapidity, amid intense debate with consensus nowhere to be found. And in current research practice, a self-report survey item that inquired about gender by presenting the two-standard, mutually exclusive response categories would probably be viewed as unacceptable by most institutional review boards, as well as by many evaluators and researchers. Operationally in survey practice, the old construct of gender is now often divided into multiple variables (e.g., gender assigned at birth, followed by current gender identity). And recommendations for best research practice sometimes involve the presentation of multiple items with multiple response categories (GenIUSS Group, 2013). As this example illustrates, the quality and usefulness of a study, whether conducted for evaluation purposes or in a research context, can depend on considerations of its construct validity quite as readily as the more widely accepted considerations of its internal and external validity.

Construct Validity's Utility in Assessing Program Theory and Evaluation Plans

If we borrow the idea of trying to represent real, observable events through a theoretical or abstract conceptualization, we land very close to the purposes that evaluators ascribe to a *program theory* as the representation of a functioning program. A program theory is a schematic, highly simplified representation of how and why the program works (Chen, 2015; Lipsey, 1993). And for a given intervention, a program theory can be judged according to how well it succeeds in this task. Thus, it is reasonable for us to apply the concept of validity—particularly construct validity—to program theories as we seek to determine how accurately they help us to understand what is going on with the program on the ground.

The evaluation questions that an evaluation study is designed to answer or shed light on are characterized by constructs, including the program or policy being evaluated as well as its desired outcomes, and the relations between them. The decisions made in an evaluation plan represent the operational translations of these constructs in ways that allow investigation and examination through, in House's terminology, the logic of evaluative argument. A rigorous evaluation study, from this perspective, is one that preserves those intended core concepts and relationships in ways that create a strong evidence base for the process of evaluative inference. In this book we focus on how to build an evaluation plan so that its elements can adequately address the representations and assumptions inherent in our evaluation questions.

Operationalizing in Evaluation

It would be appropriate to consider much of the subject matter of this book as an analysis of the process of *operationalization* (Singleton & Straits, 2018). This can be thought of as the ways in which concepts and ideas get translated into real-world research or evaluation activities. Operationalization involves multiple, significant decisions by the evaluator, requiring expert judgment and sometimes entailing a degree of research creativity.

For target outcomes and other discrete variables examined in the evaluation, this process can be readily understood. For example, the evaluator of a nutrition education program that focuses on promoting "healthy eating" can measure that construct through survey self-report, detailed food logs, meal observations (e.g., plate waste), pantry inventories, and other strategies. Even within the confines of a specific measurement procedure (most notably survey self-report), the specifics of wording can produce significant alterations. For example, evaluations of smoking cessation programs can specify their definition of "current smoking" with wide variation on timeframe (e.g., past seven days vs. past 30 days), product (e.g., inclusion vs. non-inclusion of vaping products or hookah), and other particulars. These alternatives for representing what may seem like a clearly understood construct will result in very different choices regarding the operationalization of the variable. These topics are explored further in Chapters 5 and 6.

Beyond the specification of outcome variables, other aspects of an evaluation study also require skillful operationalization. For example, an evaluator who wishes to study whether the effectiveness of a community intervention strategy varies across different kinds of neighborhoods will need to identify the features and components that determine what a *neighborhood* actually is, and then decide which of those various features will be measured, and how that will take place (more on this in Chapter 5). Entire program treatments must be operationalized as well, in ways that may be more or less adequate for the purpose of addressing study questions, e.g., they may be oversimplified or instituted at sub-optimal levels (more in Chapter 4).

Because operationalization refers to the process of turning theoretical concepts into specific measures and other forms of practice, it is, by definition, perhaps the most fundamentally pragmatic of empirical activities relating to evaluation and research. Every evaluation plan must address the process in order to achieve on-the-ground implementation of the study. Therefore it is curious that the subject has received remarkably little attention in the evaluation literature, even though it can be done well or poorly, with significant implications for the eventual quality—the validity—of the evaluation study.

In the next chapter, we examine the development of validity theory in its original domain of psychometrics.

Chapter Summary

- Evaluation is a practical discipline within the social sciences that aims to answer timely questions about social issues, programs, and policies. Accuracy is an important value in evaluation. The Program Evaluation Standards, which are widely accepted within the discipline, include eight standards that relate specifically to dimensions of accuracy in evaluation.
- Validity is a multifaceted construct that addresses issues of accuracy, quality, and fairness in evaluation, research, measurement, and other endeavors. It usually refers to the accuracy and adequacy of judgments, interpretations, and conclusions.
- Once an evaluation study's primary evaluation questions have been determined, the evaluator develops a plan with the aim of answering those questions in the most accurate and valid way possible, within the limits of available resources. This book takes the position that validity in evaluation depends in large part on the quality of the decisions that are made in the planning, conduct, and interpretation of an evaluation study.
- Evaluation can be viewed as a process of argument, in which evidence and critical reasoning are used to build a case for evaluative conclusions. In outcome evaluations, those conclusions generally address the effectiveness of programs and policies. The perspective of evaluation as argument suggests that evaluation seeks to develop a body of evidence that can be persuasive to stakeholders and other audiences in making judgments about an intervention. This book focuses on how to design and carry out an evaluation plan that persuasively answers the evaluation's guiding questions.
- Operationalizaton is the process through which concepts and ideas get translated into real-world research or evaluation activities. Every evaluation study requires multiple decisions regarding how abstract constructs will be operationalized into variables, measures, designs, timelines, analysis strategies, and other activities necessary for the evaluation to be completed.

Questions for Reflection—Chapter 1

1. In your experience, did you ever encounter an evaluation that largely failed because it lost track of its original aims?
   - What were the ways in which it went off-track?
   - What were the reasons that this occurred?
2. The Accuracy standards shown in Table 1.1 are part of the *Program Evaluation Standards* (Yarbrough et al., 2011). In reading through these standards, do they expand the concept of *accuracy* beyond the ways that you would intuitively interpret that term as applying to evaluation? If so, in what ways do they expand it?
3. How might the conception of "evaluation as argument" change the ways in which evaluation is planned, conducted, and interpreted, in comparison to perceiving evaluation as a form of scientific investigation?