Donald T. Campbell
and the Art of Practical
"In-the-Trenches"
Program Evaluation

Burt Perrin

Knowing and describing Donald T. Campbell is akin to the proverbial blind
men describing an elephant, with each touching a different part of the ele-
phant and arriving at a very different portrayal. Campbell was not a single-issue
person and defies easy categorization. As the New York Times indicated in its
obituary (Thomas, 1996), when Campbell “took up his last academic post at
Lehigh University, university officials threw up their hands and simply desig-
nated him ‘university professor.’”

This chapter focuses on Campbell’s contributions to the art of evaluation, in
particular to the art of practical, in-the-trenches evaluation that may be less “sci-
entific” than other forms of evaluation for which Campbell may be better
known. I present a very personal view of how I have been influenced by
Campbell. In these respects, the perspective in this chapter may vary somewhat from
that of other contributors to this volume. My hope is that it will be viewed as
complementary and will help the reader see another side of Campbell.

Plausible Rival Hypotheses

Let me now move from philosophy to perhaps Campbell’s most important
and practical concept. As noted above, the heart of Campbell’s philosophy is
that we can never prove; all we can do is rule out potential hypotheses or ex-
planations. This is expressed in Campbell’s (e.g., 1979; see also Webb, Campbell,
Schwartz, & Sechrest, 1966) concept of plausible rival hypotheses.

This concept is deceptively simple. It suggests that in evaluation and social
research, we should strive to identify potential alternative explanations or
threats to validity, then identify which of these are most plausible. There is no
need to devote research effort or funds to attempt to prove the obvious. Campbell
(e.g., 1979, also see Shadish et al., 1991) suggested that the best way of determining the plausibility of competing explanations is to use common sense.

Different stakeholders in an evaluation are likely to have varying ideas about what they consider "plausible" or not. I find that one of the major challenges in evaluation is to determine these various views, as well as what types and amount of evidence would be credible, in the design stage. It matters little what the evaluator considers as relevant or not: What counts is what the ultimate users of the evaluation think.

In my experience, the major way in which evaluations go wrong is by addressing the wrong questions. Program evaluation typically fails when researchers ask questions of interest to themselves but not of primary interest to the evaluation users, or when they attempt to force onto a program a method that just does not apply. It is necessary to adapt the method to fit the needs of a particular evaluation, and vice versa. In planning an evaluation, it is critical to seek the views of key stakeholders and users of the evaluation about what they view as the key questions about the program and what evidence they would consider convincing or not. I have argued that one should involve the widest and most interesting peer reviews, ensuring their full participation in the evaluation. Participatory approaches to evaluation can help create support for the evaluation process and methods, enhance the relevance of the evaluation, and provide for buy-in to the results and to ultimate action.

Campbell has also advocated using the simplest methods possible to eliminate plausible rival hypotheses. In particular, he has argued that common sense is sufficient to address the key questions (e.g., see Campbell, 1979, 1994; Shadish et al., 1991). In other situations, relatively simple methods, including use of descriptive data, may be all that is required. To employ methods that are more robust than needed or appropriate is to engage in methodological overkill. This increases the time and cost of the evaluation and can reduce its ultimate utility. This also can result in increased cynicism about and resistance to evaluation.

Campbell proposed the concept of quasi-experiments (Campbell & Stanley, 1963; Cook & Campbell, 1979) as a way in which plausible rival hypotheses sometimes can be addressed without resorting to true experiments. The concept recognizes that there is no such thing as the "perfect" method. As Reichardt (1996), for example, indicates, Campbell's work on quasi-experimentation quickly became widely accepted and has been recognized as having more influence on social science methodology than any other work since World War II. The concept, challenging randomized experimental designs as the only acceptable method for demonstrating causality, nevertheless was heresy when it was initially proposed.

"In-the-Trenches" Program Evaluation

The concept of plausible rival hypotheses has major implications for practical program evaluation. When evaluation is carried out for the purposes of contributing to decision making, program improvement, or policy development, there is never time to wait for the "definitive" answer. Decisions have to be—and are—made regularly on the basis of imperfect information. As Campbell has indicated, we should strive not for "truth" but for plausibility and for the reduction of uncertainty.

How much confidence is necessary? That depends on the situation, what is riding on the decision at hand, and what evidence the key users of evaluation would find credible. I have suggested that evaluations should strive to provide the minimum amount of information needed to provide the minimum amount of confidence required for decision making. The art of practical evaluation requires working with the client to assess the trade-offs of alternative approaches and to agree on what information would be most useful in a given situation.

Even though I rarely use quasi-experimental designs as such in my own evaluation work, I nevertheless find the concept critical to my thinking. Campbell (e.g., 1994) has indicated that more important than the application of any particular design is the rationale and thought process in the choice of methods. The analysis of sources of validity, the discussion of which makes up the bulk of Campbell and Stanley (1963) and Cook and Campbell (1979), is absolutely critical to any work in evaluation. As Campbell (1994) observed in his later years, "Do recognize that most of your evaluation tasks will be in settings precluding good quasi-experimental or experimental designs, and that even so, there is much of value that you can do to help improve program effectiveness" (p. 292).

Triangulation and the Importance of Using a Range of Diverse Methods

Campbell and Fiske's (1959) "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix" has been identified as the most widely cited paper in the social sciences. In essence, it made the case for triangulation, for the use of multiple methods to accommodate the limitations of any single approach.

As previously discussed, Campbell has indicated that all methods are flawed. All methods are theory laden and are based on questionable assumptions. All researchers bring their values and their biases to their work, and these influence their choice of methods and interpretation of data. There is no such thing as completely objective research or research methods. Quantitative methods are no more objective than qualitative methods; their biases and underlying assumptions are just less visible.
These biases and flaws are inherent in every method and can never be completely eliminated. They can be obviated to some extent by making biases and limitations as explicit as possible, but because values and assumptions are culture laden, this is not always as easy or possible as it may seem. Furthermore, users of evaluation studies are unlikely to take much notice of qualifications and disclaimers buried in dense evaluation reports.

As Campbell has indicated, the best way to control for the inherent biases and limitations of any single method is to use a range of complementary methods. Campbell was a methodological pluralist. We can have greater confidence in results when different methods come up with essentially the same findings. In particular, we can have the greatest confidence in the findings when different types of methods, such as a mix of quantitative and qualitative methods, are used.

Campbell was a strong supporter of innovative and varied methods. For example, in Webb et al. (1966), he advocated the use of unobtrusive or nonreactive methods. One commonly cited example of the "oddball" methods discussed in the book (Campbell's introduction to the book indicates that one of its working titles was "Oddball Methods") is assessing the relative popularity of museum exhibits through identifying the rate at which floor tiles in front of various exhibits wear out. There are alternatives to questionnaires and focus groups!

Campbell supported the use of both quantitative and qualitative methods drawn from competing perspectives and views of the world. Campbell pointed to the futility of searching for the "perfect" research method or for the "correct" view of the world. He supported the use of meta-evaluation to draw on a wide range of different individual studies so as to be able to draw conclusions and generate knowledge not possible through any single study. He also recognized the importance of a systems approach in attempting to make sense of the complexities of social phenomena in the real world.

The concept of triangulation has important practical implications. It means that we usually are better off applying limited resources for evaluation to multiple imperfect methods rather than trying to develop the "perfect" design. Even in low-budget evaluations, such as those I typically conduct, I strive for the use of a range of methods, even if this means limiting to some extent what can be obtained from any particular method. For example, rather than trying to implement the most comprehensive survey possible, I might use a short, focused questionnaire with a small (but representative!) sample, complementing this with qualitative data such as obtained through a small number of in-depth interviews. Use of multiple methods can be more cost-effective than use of any single method, however perfected its execution, and can provide more accurate and useful information as well.

"In-the-Trenches" Program Evaluation

External Validity

Campbell made the critical distinction between internal and external validity (in particular, see Campbell and Stanley [1963] and Cook and Campbell [1979], although he made the distinction much earlier, e.g., Campbell [1957]). Internal validity refers to the ability to draw causal conclusions in a specific situation. External validity refers to the ability to generalize the findings of a particular study to other persons, settings, or times. Campbell indicated the importance of ensuring internal validity; without it, there is nothing to generalize.

Evaluation findings that cannot be generalized, however, if only to the same program with identical characteristics at a future time, are of little or no use. Without being able to identify what factors are responsible for impact, findings about impact have little or no practical value. Not only is it difficult or impossible to hold all variables constant except one in the real world, but evaluation based on this paradigm rarely is meaningful without the ability to consider the role of context and how multiple factors interact with one another. As Kurt Lewin has indicated, the whole is definitely greater than the sum of the parts. Without being able to establish external validity, it is not appropriate to generalize findings from one setting to another, irrespective of the degree of control and internal validity established, unless the exact same conditions are in place. In the real world, this makes little sense.

In the reality of policy and program delivery, especially in today's environment where change is happening at an accelerating rate, little remains stationary. Program activities change continuously, frequently in very substantial ways, in response to changes in the political and social environment, to needs and opportunities, and to events experienced to date. Indeed, as I have indicated elsewhere (e.g., Perrin, 1994, 1998), responsive programs should be changing and adapting. They should be reviewing whether their intended outcomes are still desirable, or if they need modification, supplementation, or replacement. They should be responding to feedback from clients, to new information, and to evaluations. Few programs either can or should stand still long enough to permit experimental evaluation.

What does this mean for practical program evaluation, where the generalizability of findings usually is critical? Rather than attempting to eliminate as many extraneous factors as possible, we should strive instead toward differentness rather than sameness in program elements and contexts. Going back to Campbell and Fiske (1959), we can have greater confidence in the veracity of our findings when they hold up across somewhat different contexts, as well as through different methods. One of the strengths of a cluster evaluation approach is that it can enable the drawing of conclusions across a frequently