1
Evaluation Theory and Metatheory[1]

MICHAEL SCRIVEN
Claremont Graduate University, CA, USA

DEFINITIONS

What is evaluation? Synthesizing what the dictionaries and common usage tell us, it is the process of determining the merit, worth, or significance of things (near-synonyms are quality/value/importance). Reports on the results of this process are called *evaluations* if complex, *evaluative claims* if simple sentences, and we here use the term *evaluand* for whatever it is that is evaluated (optionally, we use *evaluee* to indicate that an evaluand is a person).

An evaluation theory (or theory of evaluation) can be of one or the other of two types. Normative theories are about what evaluation should do or be, or how it should be conceived or defined. Descriptive theories are about what evaluations there are, or what evaluations types there are (classificatory theories), and what they in fact do, or have done, or why or how they did or do that (explanatory theories).

A metatheory is a theory about theories, in this case about theories of evaluation. It may be classificatory and/or explanatory. That is, it may suggest ways of grouping evaluation *theories* and/or provide explanations of why *they* are the way that they are. In this essay we provide a classification of evaluation theories, and an explanatory account of their genesis.

DESCRIPTIONS

Of course, the professional practice of evaluation in one of its many *fields*, such as program or personnel evaluation, and in one of its *subject-matter areas*, such as education or public health or social work, involves a great many skills that are not covered directly in the literal or dictionary definition. Determining the merit of beginning reading programs, for example, requires extensive knowledge of the type of evaluand – reading programs – and the methods of the social sciences and often those of the humanities as well. To include these and other related matters in the definition is attractive in the interest of giving a richer notion of serious evaluation, so it's tempting to define evaluation as "whatever evaluators do." But

International Handbook of Educational Evaluation, 15–30
T. Kellaghan, D.L. Stufflebeam (eds.)

this clearly won't do as it stands, since evaluators might all bet on horse races but such betting does not thereby become part of evaluation. In fact, professional evaluators quite properly do various things as part of their professional activities that are not evaluation but which they are individually competent to do, e.g., market research surveys; teaching the theory or practice of evaluation; advising clients on how to write evaluation components into their funding proposals; how to meet legal requirements on information privacy; and when to consider alternative program approaches. Those activities are not part of evaluation as such, merely part of what evaluators often do, just as teaching mathematics is part of what many distinguished mathematicians do, although it's not part of mathematics or of being a mathematician. We often include training in some of these activities as part of practical training in how to be successful in an evaluation career; others are just opportunities to help clients that frequently arise in the course of doing evaluations. Of course, some evaluators are better at, and prefer, some of these activities to others, and tend to emphasize their importance more.

In any case, defining evaluation in terms of what evaluators do presupposes that we have some independent way of identifying *evaluators*. But that is just what the definition of evaluation provides, so we cannot assume we already have it, or, if we do not, that we can avoid circularity through this approach.

It is also true that what evaluators do is to a significant extent driven by the current swings of fashion in the public, professional, or bureaucratic conceptions of what evaluation *should* do, since that determines how much private or public money is spent on evaluation. For example, fashion swings regularly occur about outcome-oriented evaluation – we're in the middle of a favorable one now – and so we currently find many evaluators dedicated to doing mere impact studies. These have many flaws by the standards of good evaluation, e.g., they rarely seek for side effects, which may be more important than intended or desired outcomes; and they rarely look closely at process, of which the same may be said. These are in fact merely examples of incomplete evaluations, or, if you prefer, of evaluation-related activities.

Note, second, that evaluation is not just the process of determining facts about things (including their effects), which, roughly speaking, we call research if it's difficult and observation if it's easy. An evaluation must, by definition, lead to *a particular type* of conclusion – one about merit, worth, or significance – usually expressed in the language of good/bad, better/worse, well/ill, elegantly/poorly etc. This constraint requires that evaluations – in everyday life as well as in scientific practice – involve three components: (i) the empirical study (i.e., determining brute facts about things and their effects and perhaps their causes); (ii) collecting the set of perceived as well as defensible values that are substantially relevant to the results of the empirical study, e.g., via a needs assessment, or a legal opinion; and (iii) integrating the two into a report with an evaluative claim as its conclusion. For example, in an evaluation of a program aimed to reduce the use of illegal drugs, the empirical study may show (i) that children increased their knowledge of illegal drugs as a result of the drug education part of the program,

which is (we document by means of a survey) widely thought to be a good outcome; and (ii) that they consequently increased their level of use of those drugs, widely thought to be a bad outcome. A professional evaluator, according to the definition, should do more than just report those facts. While reporting such facts is useful research, it is purely empirical research, partly about effects and partly about opinions. First, a further effort must be made to critique the values, e.g., for consistency with others that are held to be equally or more important, for the validity of any assumptions on which they are built, and for practicality, given our relevant knowledge. Second, we must synthesize all these results, mere facts and refined values, with any other relevant facts and values. Only these further steps can get us to an overall evaluative conclusion about the merit of the program. The values-search and values-critique part of this, and the synthesis of the facts with the values, are what distinguish the evaluator from the empirical researcher. As someone has pithily remarked, while the applied psychologist or sociologist or economist only needs to answer the question, "What's So?", the evaluator must go on to answer the question, "So What?"

In this case, the reason the knowledge about illegal drugs is thought to be good is usually that it is expected to lead to reduced use (a fact extracted from interviews, surveys, or focus groups with parents, lawmakers, police, and others). Hence the second part of the factual results here trumps the first part, since it shows that the reverse effect is the one that actually occurred, and hence the synthesis leads (at first sight) to an overall negative conclusion about the program. However, more thorough studies will look at whether *all* the consequences of the use of *all* illegal drugs are bad, or whether this is just the conventional, politically correct view. The social science literature does indeed contain a few good books written on that subject, which is scientifically-based values-critique; but the significance of these for the evaluation of drug education programs was not recognized. The significance was that they showed that it was perfectly possible to combine a scientific approach with critique of program goals and processes; but this was contrary to the existing paradigm and hence just ignored.

There's a further complication. It's arguable that good and bad should be taken to be implicitly defined by what the society *does* rather than what it *says*. That means, for example, that good should properly be defined so that alcohol and nicotine and morphine are acceptable for at least some adults in some situations, perhaps even good (in moderation) in a subset of these circumstances. With that approach, the overall evaluative conclusion of our program evaluation example may be different, depending on exactly what drugs are being taken by what subjects in what circumstances in the evaluation study. If we are to draw any serious conclusions from such studies, it is essential to decide and apply a defensible definition of social good and to answer the deeper questions as illustrated above. These are the hardest tasks of evaluation. Of course, these challenges doesn't come up most of the time since there is usually little controversy about the values involved, e.g., in learning to read, or in providing shelters for battered women and children. But it's crucial in many of the most important social interventions and policies. Avoidance of this obligation of evaluation vitiated or

rendered trivial or immoral the research of many hundreds, perhaps thousands, of social scientists who did not question the common assumptions on these matters, for example in the notorious case of social science support of dictators in South America.

These further steps into the domain of values, beyond the results of the empirical part of the study, i.e., going beyond the study of what people do value into the domain of what the evidence suggests they should value, were long held (and are still held by many academics) to be illicit – the kind of claims that could not be made with the kind of objectivity that science demands and achieves. This skeptical view, known as the doctrine of value-free science, was dominant throughout the twentieth century – especially in the social sciences. This view, although also strongly endorsed by extraneous parties – for example, most religious and political organizations, who wanted that turf for themselves – was completely ignored by two of the great applied disciplines, medicine and the law. For example, no doctor felt incapable of concluding that a patient was seriously ill from the results of tests and observations, although that is of course an evaluative conclusion that goes beyond the bare facts (it is a fact in its own domain, of course, but an evaluative fact). This legal/medical model (partly adopted in education and social work as well) would have been a better model for the social sciences, whose chosen theory about such matters, the value-free doctrine, rendered them incapable of addressing matters of poverty, corruption, and injustice because, it was said, the mere identification of those things, since the terms are value-impregnated, represented a non-scientific act. Consequently, many areas languished where social science could have made huge contributions to guiding interventions and improving interpretations, and people suffered and died more than was necessary. Ironically, many of those who despised excursions into the logic or philosophy of a domain, thinking of themselves as more practical for that choice, had in fact rested their efforts on one of the worst logical/philosophical blunders of the century, and thereby had huge and highly undesirable practical effects.

Great care is indeed needed in addressing the validity of value judgments, but science is no stranger to great care; nor, as we'll see in a moment, is it any stranger to objectively made value judgments. So neither of these considerations is fatal to scientific evaluation. The real problem appears to have been the desire to "keep science's nose clean", i.e., to avoid becoming embroiled in political, theological, and moral controversies. (It is clear that this is what motivated Max Weber, the originator of the value-free doctrine in the social sciences.) But getting embroiled in those issues is what it takes to apply science to real world problems, and balking at that challenge led to a century of failed service in causes that society desperately needed to press. To the credit of educational researchers, some of them followed the medical/legal model, which led to half a century of pretty serious evaluation of various aspects of education. In fact, to put it bluntly, educational research was several decades ahead of the rest of social science in the search for useful models of evaluation, and still is, to judge by most of the evaluation texts of the new millennium (see references). Sadly

enough, although this seems clear enough from a glance at the literature, it is rarely acknowledged by mainstream social scientists who have gotten into serious evaluation, a shoddy example of misplaced arrogance about the relative intellectual importance of education and the mainline social sciences.

Part of the explanation for the avant garde role of education in improving evaluation approaches may be due to three factors. First, a typical premier school of education, with its historians, philosophers, statisticians, and qualitative researchers, is remarkably interdisciplinary and less controlled by a single paradigm than the typical social science (or other science) department. Second, education is heavily committed to the application of its research, in this respect it is like engineering, medicine, and the law. And thirdly, it is usually an autonomous college, not so easily driven by the fashions espoused by fellow high-prestige fellow departments.

One must concede, however, that it was difficult to conceptualize what was going on, since most educational researchers are, appropriately enough in most respects, strongly influenced by social scientists as role models, and there was no help to be found from them. Not surprisingly, there emerged from this confused situation a remarkably diverse zoo of models, or theories of evaluation, or, we might equally well say, conceptions of evaluation. And since many people who came to do the evaluations that government funded had been brought up on the value-free doctrine, it is not surprising that that conception – it's really a denial of all models rather than a model in its own right – was very popular. This negative view was reconciled with actually doing evaluation, as many of the value-free doctrine's supporters did, by saying that evaluators simply gathered data that was relevant to decisions, but did not draw or try to draw any evaluative conclusions from it. This was "evaluation-free evaluation", perhaps the most bizarre inhabitant in the evaluation-models zoo. Let's now look in slightly more detail at this and some other evaluation models.

MODELS OF EVALUATION: EIGHT SIMPLIFIED ACCOUNTS

Evaluators play many roles in the course of doing what is loosely said to be evaluation, and, like actors, they sometimes fall into the trap of thinking that their most common role represents the whole of reality – or at least its essential core. There seems to be about eight usefully distinguishable cases in the history of evaluation in which this has happened. I list them here, calling them models in order to bypass complaints that they are mostly lacking in the detailed apparatus of a theory, and I append a note or two on each explaining why I see it as providing a highly distorted image of the real nature of evaluation. In most cases, the conception is simply a portrayal of one activity that evaluators often perform, one function that evaluation can serve – no more an account of evaluation's essential nature than playing the prince of Denmark provides the essence of all acting. Then I go on to develop the general theory that is implicit in these criticisms, one that I suggest is a more essential part of the truth than the others,