Minimalist Theory: The Least Theory That Practice Requires
Michael Scriven
American Journal of Evaluation 1998; 19; 57
DOI: 10.1177/109821409801900105

The online version of this article can be found at:
http://aje.sagepub.com/cgi/content/abstract/19/1/57
Minimalist Theory: The Least Theory That Practice Requires

MICHAEL SCRIVEN

ABSTRACT

This paper has two parts. The first provides a general discussion of the nature of theories—and the reasons why some understanding of this abstract subject is of value in the practice of evaluation. The second part looks more closely at the dependence of good practice on a minimalist theory of evaluation, and at what such a theory comprises.

PART I

Theories

Many kinds of theories fall well within the domain the dictionaries allow: that is, they provide general and economical accounts of a domain, aimed at conceptual simplification and/or explanation of the phenomena, and involve a greater degree of speculation than does direct observations. Note that prediction is just a bonus, not a necessary condition for something to be a theory, even a good theory, and there are many respectable disciplines, from geology to history, where the best theories generate no predictions.

Theories run the gamut from those that are very close to being a straight descriptive hypothesis (e.g., theories about the death of the dinosaurs), all the way to elaborate formal structures proposed as representations of the nature of abstract or invisible entities like gravitational fields, which are far removed from the basic (often but not always observational) data that the theory needs to integrate and/or explain. At the low end of this staircase to the theoretical stars, where we still lack what might be called ‘deep explanations,’ we often talk of providing a ‘conceptualization’ or an ‘account’ rather than a ‘theory’ of the field. With huge fields, like physics or medicine, the entries at the high end of comprehensive theories are so general that we are likely to call them ‘philosophies’ of the field rather than theories, in the strict sense. However, the big fields are usually sub-divided into areas such as particle physics
or relativity, where what we call a ‘general theory’ can still be developed that is recognizably part of the field instead of being actually or nearly part of philosophy of science (or philosophy of medicine, etc.).

Ortho-theories

All of the examples I have used so far are within what we might call the plane of the field to which they refer, i.e., they are about its subject matter, at one or another levels of abstraction. In various orthogonal planes, we have the theories of those from other fields who bring their perspective to bear on the field, but not in a way that supplements or enriches the explanations in the field itself. In the case of science, for example, these ortho-fields include the history of science, the sociology and psychology of science, and the economics of science. Once in a while people in these fields get too big for their britches and decide that the ‘real’ explanation of quantum theory is as a means for the middle class to exploit or cow the proletariat, but we rightly ignore most of this type of category mistake.

Although the philosophy of science is often classified with this set of ortho-theories, in fact most of it is on the same plane as science. This becomes apparent when truly fundamental questions come up, e.g., when quantum theory began to get weird, Niels Bohr and the Copenhagen School went back to the Greek atomists in the search for explanations; or in discussions of the one-way velocity of light; or as today in the artificial intelligence/expert systems field we find computer scientists turning to philosophers’ accounts of consciousness. A good example of the way in which the philosophy of science merges into the science concerns discussion about the nature of a field: is psychology the science of consciousness, as William James thought and Rosenthal and Rosnow (1984) reiterate in one of the most respected methods texts? Or is it about the behavior of an empty organism, as B. F. Skinner thought, following in the tradition of Descartes? Both psychologists and philosophers engage in that kind of dispute and both have important contributions to make to it.

There is another type of entry that is harder to place—methodology. Literally, one might suppose that methodology is at an intermediate level between theory and practice. It is still one level above practice, since it is the study of ways of doing research in the field rather than the use of those ways to study the phenomena in the field. But discussion at the level of methodology often generates its own theory or philosophy of a whole field: for example, behaviorism and operationalism are essentially methodological commitments, not ontological ones (although an occasional enthusiast like Skinner with the “philosophy of the empty organism” goes further). So the philosophy of a field merges not only with its content but also with its methodology. (The two-volume [New Shorter] Oxford explicitly includes general accounts of the methodology of a field as a theory.)

Metatheory

As you can see, it only takes a few minutes of discussion to generate a taxonomy of theories in our metascientific language. First, there are conceptualizations and accounts (the least precise kind of theory) of a whole field; second, simple descriptive hypotheses; third, the more common kind of general theories, usually of subfields; fourth, ortho-theories; fifth, not entirely distinct from the first, theories from the philosophy of the specific sciences; and sixth, not entirely distinct from the philosophy of the subject, methodological theories. Of course,
the lines between these taxons are fuzzy, but they provide a useful framework for our discussions here.

The first proposition I would like to put forward is that almost any professional working in a field needs to understand something about what theories are and what they try to do: for example, they should know how they are different from and related to paradigms, hypotheses, and classifications; how they are confirmed, disconfirmed, and refuted, and how different types of theory from our taxonomy have different functions and scopes. In other words, I think that a smattering of theory about theories (this topic is called metatheory) is a prerequisite for professional practice in any field, just as I think a smattering of statistics is a prerequisite for practice in our field. Now, why should the practitioner bother with metatheory? Without some understanding of the nature of theories of the different types I have mentioned, confirmed antitheoretical practitioners are all too easily misled into making assumptions of one kind or another, and dismissing them as 'all part of theory' which they have decided to eschew, thinking that theories have little impact on practice. But often enough, the assumptions they make are about methodology or the purpose of evaluation, and mistakes in those assumptions can directly and adversely affect their success as practitioners. For example, they can reject or accept a particular assignment because it is “too theoretical” or “not theory-driven enough”, when a more careful look at what is meant by theory in those uses of the term (or its cognates) would make them realize they have made quite a serious mistake.

**Theory-driven Evaluation**

Take the dispute about ‘theory-driven’ evaluation, for example. Although there is a good deal more to it than the points I will discuss here, these may suffice to suggest that some attention to the nature of theories is significant for practice in the field of evaluation. First, people sometimes think that all evaluation should be theory-driven, that this is the natural way to do evaluation. That shows a lack of understanding of the necessity to do black-box evaluation when no theory is available or needed, or when getting the relevant expertise would take resources of time or money we do not have. It is a very practical point to avoid doing what is expensive and not needed; and a little understanding of the general nature of evaluation will avoid this error of practice. Again, one often finds students thinking that identifying the components of an evaluand (i.e., an entity that is being evaluated) is a sign one is doing theory-driven evaluation. On the contrary, it is merely a sign that one is doing one species of analytic evaluation by contrast with holistic or global evaluation, and analytic evaluation—something which had been identified and done long before theory-driven emerged—involves no theory in anything like a proper use of that term. That is, it involves no economical set of principles that accounts for the prescience and relation between components and explains the results of the program by appeal to these principles. It would be weird to suggest that some child who pulls his grandmother’s best alarm clock to pieces has a theory about how clocks work; but he or she has certainly found out what the components are.

Theories are sets of propositions which jointly provide explanations and integration, so a list of components, even if supplemented by an account of how they fit together, is not a theory of operation, because it provides no explanation of the fit, or of the unifying logic of operation. It is simply a set of instructions for assembly. Nor is a set of simple observations about the connections between components a theory of the evaluand; at most, it is a partial set of intermediate level propositions that have to be derived from anything purporting to be a theory. But it may be enough to identify the cause of failure. In an electronic clock, for example,
it is easy to see that there is a battery and a chip and some wires and a gear train driving the hands with input from both the chip and the winding stem. But understanding that, although it is more than enough to explain most failures of such clocks (since the electronic module almost never fails), does not even begin to qualify as a theory of how the clock works in the scientific sense of theory. It is a set of observations about the connections of some of the components, and it requires zero knowledge about electronics on which all the time-keeping functions depend completely.

The practical significance of this is that one often does not need theories or theoreticians in order to fix things and explain their failures. If one is not at least slightly aware of the nature of theories and the difference between theories and a common sense understanding of connections between components, one is likely to be reaching for the phone to call in experts about electronics before looking to see if any wires are broken or gears slipping. That is simply bad practice.

There is a related error concerning the need for theory. In Chen’s (1980) original text, what he says about his meaning of theory is different from what the dictionary tells us. He defines what I will call a “Chen theory” as an account that will explain how the evaluand produces its outcomes (not just outputs), and will predict likely side effects. The first part is more than many theories do, in the usual sense (they just explain outputs), and there are often less fancy ways to do it; it is in fact a theory about the evaluand’s operation as an intervention. Sometimes that is called the program logic. The second part is not often found in theories and can also be done without them.

So, one should not take this as saying that everything that does these things is a theory in the usual sense or that one needs a theory in order to do these things. One may have no theory of X but be able to see how it produces its effects; for example, one may have no idea how the grow tubes in a home greenhouse work, but one may understand that they produce good results by simulating daylight, with the sort of consequences one expects from daylight. One might call this a hypothesis about how the greenhouse works, but it is hardly a theory of how grow tubes work. Conversely, as in the classic example of aspirin, one may have no theory of how it works to produce its effects, but nevertheless be able to predict its effects and even its side-effects—because we found out what they were from direct experimentation. That does not require a theory.

Another way to put the difference between Chen theories and more usual theories about the evaluand is to say that we are talking about two kinds of theory here: (i) internal theories, which explain how an entity manages to produces its outputs; and (ii) external theory (also known as intervention theory); which explains how its outputs produce the effects they are used to produce, which we often refer to as outcomes by contrast with outputs. The point being made here is that it is quite possible to understand the external theory of a program, i.e., what kind of effects certain kinds of intervention can have, and know how to avoid some common side-effects, without knowing anything about how the program produces the outputs, i.e., about the internal theory of the program.

So we need to be very careful about the notion of a “program theory’ as defined by Chen. In general, it is supposed to do both the job of internal and external theories, and internal theories rarely if ever can do that. The key practical point here is that it is often a waste of time to become learned about internal theory, e.g., management theory, in order to improve one’s evaluation of the effects of changes in management style on the rest of the organization or on the success of the company.
In short, do not sell common sense short in matters of explanation. And if you want to extend it, go to program logic, which is a kind of organization chart of the links between inputs, components, outputs, and outcomes. Like organization charts in general, program logic may be more programmatic than realistic, but it is highly relevant to those whose conception it is, and who are interested in its results. As the evaluator, yours is not to argue on organizational theory grounds whether the staff or administration were unrealistic about all this; yours is to find out whether they were wrong, and if wrong, how wrong, at the level that you can determine with the skills in the evaluator's toolkit. For that, all you need to know is what they thought the machinery was, not what it really was, and to be able to contrast this with what machinery is really there, and how it is really operating. Drivers can find out how well cars run without knowing anything about auto mechanics; and even so, they can tell what is wrong in a hundred cases without any technical knowledge—the wheel fell off, the puncture-proof tire punctured, the lights went out, the gas ran out, the engine seized up, or the transmission lever came off in one's hand. We do not need to have anything properly called a theory of how cars work to explain many cases where they did not. And even if you think this is a kind of theory, you have to understand it is not a Chen theory—it would not come close to predicting any of these events or side effects connected with them.

It is true that if you want to explain why these events happened in any depth, a Chen theory will do that for you, if you can find one. But the primary task of the evaluator is to determine how well the car runs, and for that you do not need the auto mechanics. Moreover, even without that theory you can give good explanations of why it failed to run, a secondary but frequently useful task. Deep explanations, on the other hand, are the business of subject-matter experts, not evaluators; that is why the best expert road-testers do very nicely without being members of the Society of Automobile Engineers. But even deep theories will not predict side-effects; the accidents that killed two other people because the car blew a tire on the freeway. That is too much to ask of a theory of any kind, even an extended theory, such as a Chen theory. That is magic.

In a particular case, we may or may not need intervention theory; we may well have a client whose principal concern is whether or not the intervention works, i.e., produces the outputs it is supposed to produce. If it does not, then we may look at the components to see if any are faulty; and/or at simple program logic to see if there are any faults in that common sense account, on the evidence we have accumulated. We may never need to know in any detail how the inputs to the program produce its outputs, or how those produce outcomes, which is the domain of Chen theories.

Of course, there are linkages between internal and external theories but that does not make them all part of one theory any more than neurophysiology is part of social psychology, although both are about people. Applications and evaluations have a life of their own, requiring little or nothing from the theories that provide what are in some sense deeper explanations.

Evaluation Theories

Turn now to the view of Shadish, Cook, and Leviton (1991), which is that a good theory of evaluation should provide an account of the utilization, or lack of utilization, of evaluation. Following the line sketched above, I think this could better be expressed as the view that a complete multi-dimensional treatment of the field of evaluation should include some coverage of utilization, just as a complete account of the phenomena of mathematics would call on several ortho-disciplines to cover its use, history, psychology, and sociology. None of these
involve any internal mathematical theory; and in the same way, the study of the utilization of evaluation involves nothing about internal evaluation theory—the theory of evaluation as an intellectual discipline, i.e., the substantial or internal theory of the entity that is studied by the external subjects. A theory about evaluation’s utilization is an external theory, simply part of the sociology—and perhaps the history and psychology and economics and politics—of evaluation. It is a perfectly legitimate theory, but not a theory about the abstract entities and methodologies that define evaluation: it is about what gets done with the results of applying those entities to certain problems, how people came to think of those entities, and the like.

On this account, it would be completely inappropriate to expect a single ‘theory of evaluation’ to cover both the properties of merit and worth and such phenomena as evaluation use or abuse, with which we are all familiar. In a loose sense, we do think of all of this as grist to the mill for theorizing about evaluation, and rightly so; but when we get down to the details, we should quickly separate out the different tasks of internal and external theories, as outlined above. Some people must work on the evaluand itself, on developing the theory of what defines evaluation, as some do with mathematics; others must work on the question of explaining and describing how whatever there was of a core discipline at the time was used or abused, applied or not applied, historically or sociologically. And that brings me seamlessly to the question of what evaluation is, to the internal theory of evaluation, and in particular to the minimalist theory.

In closing this first major section, let me just sum up the thrust of it in two sentences. I have been arguing that a minimalist theory of evaluation, the minimum package that a practitioner must have, includes, surprisingly enough, some understanding of the nature of theories. We often think of that kind of issue as two levels more abstract than practice, as more abstract even than theory; but I have argued that putting it aside in that way leads to bad understanding of what practice can and cannot do, and hence to bad practice.

**PART II**

**Minimalist Evaluation Theory**

A minimalist theory is more than the simplest possible description or conceptualization of a field, because it must, for the reasons just given, include some metatheory. In particular, along with some understanding of the nature of theories in general, it must include some theory—at least some conceptualization and rationale—about the distinctive nature and the limits of (for example) evaluation. These ‘criteria of demarcation’ draw the lines between and around evaluation and other cognitive activities. A minimalist theory also states some fundamental principles as part of or in addition to integrating the phenomena via new concepts. But it is well short of a general theory, because it does not try to provide the complete basis required for answers to all the hard questions in a field, many of which spring from theory itself rather than practice. One might say that minimalist theories are the irreducible stubble left after you shave with Occam’s Razor. William of Occam did not recommend the avoidance of all theoretical entities, he only recommended avoiding them whenever possible—exactly the spirit of minimalism.

Now, in what sense are minimalist theories essential? In what sense does everyone need them? Do not practitioners get along very well without theories? I believe that they cannot, but this point has to be proven. I recently dealt with this issue in some detail (Scriven, 1996) in a
review of a valuable anthology edited by our distinguished and highly experienced colleagues Wholey, Hatry, and Newcomer (1994). Here I will pick up on a few themes from that treatment that still seem to me worth stressing a year later.

The first thing that strikes the thoughtful reader is that nowhere in this anthology's 25 chapters and 600 pages, mostly devoted to elaborations of social science methodology to suit the special needs of evaluation, is there any reference to evaluation-specific methodology. Not theory, but methodology. Now, that of course commits the editors to an implicit conception—if you like, a theory-based definition—of evaluation that is close to the one in Rossi and Freeman's (1993) text, the view that evaluation is an application of social science methods to assess social intervention programs. So, let us take a moment to look at that latter definition. If Rossi and Freeman had called their book "...Program Evaluation," as did Wholey, Hatry, and Newcomer, rather than "Evaluation,...), with no limitation, that definition might make a little more sense as an operational definition—since it would then be less obviously irrelevant to twenty fields like product and personnel evaluation—but even then it fails because it leaves out ethics, the law, and a few other things like cost analysis that are not normally taken to be part of social science methodology. (Wholey and his colleagues do have a very sophisticated chapter on cost analysis, even though it does not fit that definition.) The Rossi and Freeman definition also excludes programs that are aimed at studying phenomena that were not interventions, but just occurred in the natural course of human affairs. Now there are a good many of these, such as presidential campaigns, nuclear leaks, and market crashes, that have been evaluated by historians, economists, sociologists, and many others, and a good many more that will be, and perhaps others that should be. None of them seem to pose problems that are beyond the reach of a competent program evaluator with appropriate field-specific knowledge, so it seems odd to exclude the study of them from program evaluation. Then of course there is the slight problem of circularity: the Rossi and Freeman definition is couched in terms of the notion of assessment, which is very close to being a synonym for evaluation.

Definitions that do not match the field to which they are said to apply and that do not manage to avoid circularity are not tremendously enlightening. Now Wholey, Hatry and Newcomer also rely on a definition in terms of assessment, although with its own peculiarities (e.g., what is assessed is said to be only "the results of the program"). Surely programs with unethical or illegal personnel practices should be faulted for that in the course of serious program evaluation. Surely one must include some attention to process in program evaluation. Perhaps one can best understand these colleagues' concept of program evaluation by exclusion. Some of the terms commonly involved in evaluation but missing from their index include: needs assessment, worth, values, ethics, law, personnel, grading, ranking, scoring, apportioning, standards, non-money costs, meta-evaluation (or the evaluation of evaluations), integration of sub-scores—and of course evaluation theory. Program theory, on the other hand, gets a strong endorsement, backed by the well-known quote usually attributed to Kurt Lewin, "There is nothing so useful as a good theory". It is interesting that it did not occur to these colleagues to apply that to evaluation of programs, and not just to programs. Of course, that omission might be due to the absence of anything worthy of the name of evaluation theory, which brings me to the substantive task of setting out the elements of minimalist evaluation theory.

The first thing that a theory of evaluation should do, it was suggested above, is to lay out some criteria of demarcation. It should tell us how evaluation is different from other types of investigation, such as prediction, data analysis, diagnosis, research, explanation, description, generalization, and classification. The definition Wholey, Hatry and Newcomer use tells us
that if these are used to assess programs—and they all have been used for that purpose, in one context or another—they are evaluation. But what is it that identifies "using the research for assessment"? Since assessment is close to being a synonym for evaluation, this hardly clarifies the situation. Is that all we can say to students beginning to study the subject, or to colleagues we wish to enroll in a project? It is not surprising that most of the book is about standard methods in the social sciences; for them, that is all there is to evaluation. But evaluation-specific methodology is serious stuff, and considering it seriously leads to greatly improved efficiency in practice, as for example matrix sampling led to greatly improved efficiency in the practice of survey work.

Minimalist theory begins by addressing the demarcation issue. But it goes on to set out a core doctrine about the nature and content of evaluation. The suggestion in this paper is that without some general recognition of this set of propositions we do not have an understanding of our own enterprise, and our practice will suffer seriously. So here are the basic elements of a minimalist theory of evaluation.

1. The discipline of evaluation undertakes the systematic, objective, determination of the extent to which any of three properties are attributable to the entity being evaluated: merit, worth, or significance. (Merit is roughly equivalent to quality; worth is roughly equivalent to value or cost-effectiveness; significance is roughly equivalent to importance.) Each of these concepts is context-dependent, especially significance, and understanding the difference between context dependence and arbitrariness is part of understanding the logic of evaluation.

2. All evaluative conclusions are expressed in terms of just four predicates: (i) ranking; (ii) grading; (iii) scoring; and (iv) apportioning. Each can refer to any of the three properties mentioned in the previous paragraph, and each is logically distinct and hence each requires a different investigatory design (for example, grading requires defining standards of, for example, merit, other than by reference to the distribution).

3. Recommendations and explanations can almost never be deduced from evaluative conclusions. They require further knowledge and evidence beyond that required for the core evaluative tasks, and the nature and source of this further knowledge need to be made explicit and verified carefully whenever evaluators elect to move from a purely evaluative conclusion to an explanation or recommendation. One can quite often make the extra step using plausible premises, but to make it without examination of the assumptions required is a flaw in an evaluation.

4. The general outline of an evaluative investigation will normally involve determining some and often all of the following: (i) the nature of the questions, assumptions, and context (e.g., client, audiences, stakeholders, history, reasons for the evaluation) that define the entry point to the evaluation; (ii) the nature of the entity being evaluated (the evaluand); (iii) the sources and validation of values that will be used in order to generate answers to the evaluative questions (e.g., via needs assessment, existing codes, standards, principles, strategies, law, ethics, management or employee preferences, competitor performance, generalizability, costs, objectives, conceptual analysis); (iv) the criteria of merit (or worth or significance) for an entity of this kind in this context (e.g., access, outcomes, reduction of alcohol use), and their justification; (v) the relative importance or weight of each of the criteria, and their justification; (vi) the identification of standards ("cutting scores") on the (qualitative or quantitative) scales on which these criteria run (if grading is required—standards are not required for
ranking, scoring, and apportioning) and the justification for these standards; (vii) the empirical or analytical determination of the achievements of the evaluand on each of these scales (using measurement, observation, experimentation, expert testimony, logical analysis, etc.); (viii) the integration (internal synthesis) of the achievements and weights into an overall conclusion about the merit (etc.) of the evaluand (this step is dispensable in a few special types of evaluation); (ix) the conversion of the results into an appropriate report or set of reports, which may be verbal, written, or graphical. Items (iii), (iv), (v), (vi), and (viii) are evaluation-specific, and are not normally covered in social science training; not understanding how to perform these steps undermines, usually fatally, the ability to perform practical evaluations. There are aspects of items (i), (vii), and (ix) that are also evaluation-specific and in some evaluations these aspects will be absolutely essential.

5. Evaluation is one of a small group of disciplines referred to as transdisciplines: the list includes statistics, measurement, decision theory, experimental design, and real logic (which includes fallacy theory, fuzzy logic, and informal logic, but not the body of formal logic, which is merely a stand-alone branch of pure mathematics, a sidetrack in the history of logic). They serve the major function of providing tools to other disciplines, although they also have stand-alone status. (This is quite different from interdisciplines, which are created or emerge as hybrids of one or more other disciplines, although many subfields in evaluation, e.g., program evaluation, are also interdisciplinary.) Among these transdisciplines, logic is perhaps the most general, but evaluation either ties it or comes in a close second, since it provides essential tools not only in the sciences, but in the crafts and physical disciplines where logic has a secondary role. (For example, the difference between a discipline and a pseudo-discipline is itself a complex evaluative judgment.)

6. There are more than a score of named and developed fields of evaluation. These are at different levels of development, but each has the same underlying logic outlined above. Program evaluation has received more attention than most, but personnel evaluation, performance evaluation (notably educational testing) and product evaluation/technology assessment are in most evaluation-specific respects more sophisticated, and no less difficult. (Note, for example, the discussions of the modified Angoff procedure for setting standards in educational testing). The methodology of many of these fields involves errors which are avoided in other fields and are rendered obvious by spelling out the logic of evaluation in the terms outlined above. For example, they may use rating scales for the “sub-scores’ (which may be qualitative) that mix ranking and grading anchors (e.g., the anchors may include both Excellent and Average, or Outstanding and Satisfactory; or they may use some version of “grading on the curve’); or they may use scales that have no provision for a minimum necessary level on particular scales (an error in most proposal evaluation).

7. There are two fields of evaluation that are not commonly recognized but are of special importance. One is intradisciplinary evaluation and the other is meta-evaluation. Let me touch on both briefly.

The imprisoned giant of the sciences, intradisciplinary evaluation, provides the skeleton on which the meat of all scientific investigation depends. (This is one reason why it is completely wrong to conceptualize evaluation as an application of the social sciences, and absurd to talk of value-free science.) All scientific work rests on continual, repeatable, intersubjective, multiple processes of evaluation, such as the evalua-
tion of data, hypotheses, classification, taxonomies, measurements, instruments, experimental designs, interpretations, theories, scientific papers, student work, scientists, and so on. Hiding this fact under the dirty clothes of the value-free doctrine was one of the most extraordinary confidence tricks in the history of thought, not least because it deceived those who did it. What has just been said about the sciences is also true of engineering, jurisprudence, philosophy, and, for that matter, tai-chi.

When we get down to making the criteria and the process of evaluation explicit in intradisciplinary evaluation, there emerges the possibility of improving it. We certainly can improve some of its applications in fringe areas of scientific activity, such as proposal evaluation and the evaluation of research output, by contrast with evaluating mainline science. What has until now been acquired through osmosis in the process of apprenticing to a particular science, and then generalized very shakily to other areas of application (such as the evaluation of proposals for funding at NSF), can now be brought out from the closet and—under the harsh light of scrutiny by evaluation methodologists—turns out to be rather sloppy. Scientists recognized this point rather slowly, as they gradually moved to institute blind reviewing for journals. Some of them have done enough in science education to see ways to improve the standard assessment procedures in science education, including the grading of students in graduate courses. In mainline social science, two of many examples where the logic of evaluation makes a contribution to progress are the study of significance testing and the use of style variables in personnel evaluation.

The other especially significant but only recently emergent field in evaluation is meta-evaluation, the evaluation of evaluations. Its significance is that it demonstrates the reflexivity (self-reference) of the discipline of evaluation. As evaluators have developed their critical sense in evaluating evaluations, they have been forced to make the criteria of merit that are being used explicit, and doing this has led evaluation to tidy up its own act. This bootstrap effect is a powerful force for self-improvement in the field, and some of us are hoping and intending that we can do more of it in the public domain of the AEA journals in the next few years, e.g., in the new series on exemplary evaluations, or perhaps in an even more directly evaluative section devoted explicitly to meta-evaluation of prior evaluation studies.

The study of meta-evaluation has produced three techniques for doing it (replication ab initio, procedural analysis, and the use of special checklists) and even if used only by the practitioner, these typically produce substantial improvement in the quality of the product. If they can be done by an independent evaluator—which is possible at a minimal cost level, if the second and third approaches are employed—the results are often better, sometimes startlingly so. In either case, there is a benefit to the practitioner, and in the second case, especially if the meta-evaluator reports directly to both the evaluator and the client (the recommended alternative), there is also a notable increase in credibility.

8. The process of bringing evaluation out of the closet makes it possible to identify the huge range of applied fields in which very similar work, including work on evaluation methodologies for dealing with certain problems, is being done. This for the first time raises the hope that we can avoid the continual process of reinventing the wheel in each field, a process that resulted from the denial of the legitimacy of the discipline of evaluation, and hence of the existence of a common logic for the many applied fields that sprang up because practical needs demanded it. For the practitioner, this has a
9. Evaluation skills are of crucial importance in a number of practical activities—other than evaluation as such—that suffuse everyday life, managerial life, professional life, craftwork, and governance. These include planning, designing, needs assessment, goal-clarifying, diagnosing, recommending, auditing, mentoring, explaining, mediating, decision making, selecting, trouble-shooting, leading, and the formulation of regulations and legislation. At one time or another, the practical evaluator is involved, and should be involved, in all of these efforts, and all of them can be improved by attention to the relevant tenets of minimalist evaluation theory and the results that it has generated and is generating.

Theory in Practice

A. In the above outline, I have interspersed quite a number of illustrations of how theory pays off for practice, but it is worth spelling out how examples from this list relate to Wholey, Hatry, and Newcomer’s (1994) handbook (hereafter referred to only as “Handbook”), and to other practical evaluation efforts. The fourth numbered paragraph in the foregoing set of nine features of minimalist theory lays out the steps or components (sometimes implicit) in the logic of evaluation. What does it suggest for practice? Item (iii) in that paragraph concerns the source of the values used in the evaluation and it reminds us of the necessity for (a) identifying these, instead of just assuming their legitimacy, and (b) validating them where there is any question about them. A leading source of values is often the needs assessment, but for most programs there are several other relevant values such as legal, ethical, professional, scientific, political, and situation-specific values. In the Handbook, as in most practical evaluation books, we find virtually no discussion of these other values that should be checked (usually quite quickly) for relevance in every evaluation. The exception is that the ethical issues that refer to methodology are discussed, e.g., the rights of human subjects. But there is no discussion of how to deal with ethical issues concerning the program itself. Surely with family planning clinics, for example, one can hardly fail to address such matters, or at least explain the reasoning behind failing to do so.

B. In the Handbook, as in most evaluation books, there are few references to the tough logical problems of defining and designing needs assessments. For example, the practitioner has to understand how it is different from (and why it is more important than) wants assessment (which surveys and focus groups are more likely to turn up), or ideals assessment (which program planners tend to overweight); or exactly how one goes from a needs premise to an evaluative conclusion about an intervention. The common attention to considerations of internal and external validity, part of the basic menu in social science methodology, should be matched by equal vigilance about the logic of the move to evaluative rather than causal conclusions.

C. The fourth item in the minimum list for evaluators is essentially the task of developing evaluative checklists, a topic so important as to deserve full treatment in a future
paper, rather than trying to squeeze its discussion into this article. The fifth and eighth items refer to two different parts of the process of evaluative synthesis, about which I have already written enough to make repetition here boring (e.g., in Evaluation Thesaurus, 4e, Sage, 1991). None of these are discussed in the Handbook.

D. The sixth item—standards determination—is very tough. It is still being explored by social scientists, especially those in educational testing and measurement, but so far without results that are satisfactory to them, let alone to logicians. However, it is clear that present approaches are substantially better than doing it by eye, and the method used—some method will have to be used in any evaluation that leads to a grading conclusion—needs to be justified. It is not discussed in the Handbook.

E. In the third numbered paragraph of the previous section, mention was made of the need to justify the step from evaluative conclusions to recommendations. In the Handbook, naturally enough, recommendations play a large part; but there is no discussion of this step. Consequently, what amounts to a serious flaw in many evaluations is left open for critics to attack; this is entirely analogous to the omission of justification for the synthesis of sub-evaluations into an overall evaluation, when that step is necessary, as if usually the case. The recent extensive discussions of theory-driven evaluation have also raised our consciousness about the need to examine our assumptions, and this paper may persuade some to reach out in some other directions with the same intent.

F. For practitioners, the issue of getting decision-makers to act on the basis of an evaluation is a common concern. It is, perhaps for that reason, common for practitioners—and some theorists—to suppose that the failure to do so act is a sign of weakness in the evaluation. I believe that this is a complete misunderstanding, for several reasons, most obviously because there are often contextual considerations surrounding the decision that cannot be fully conveyed to the evaluator. Such conveyal is often difficult because these contextual factors, including legitimate political ones, are (1) both complex and remote from the program being evaluated (e.g., overall funding considerations for an agency), and (2) they often arise or crystallize between the time the evaluation is commissioned and the time it is concluded. These can fully justify ignoring its main thrust without casting the slightest aspersion on its validity or prima facie utility. It seems likely that a main reason for the misconception here is a confusion about the logic of evaluation and its connection to recommendations, explanations, and decisions. If one does not fully understand their key logical differences, as well as their more obvious but context-dependent connections, one is likely to imagine that failure to act on the basis of an evaluation is a rejection of the evaluation, whereas the evaluation should always be taken as only relevant to one dimension of the action decision.

G. For a practitioner, the question of what counts as disqualifying bias in an evaluator or a subject-matter expert is often an important one. For example, there is often some discussion as to whether a high degree of involvement in the field almost always creates bias. It usually does lead to certain conclusions about the main issues in the field. Is that the same as bias? This is a theory issue and quite a tricky one. Surely it is also an issue where a Handbook offering advice on practical program evaluation might be helpful.

H. There is no reference in the Handbook (or in most program evaluations or evaluation handbooks, for that matter) to personnel evaluation. But programs are run by people,
and there are serious concerns about whether they are treated properly (equitably, legally, sensibly) that come up in the course of program evaluation from time to time. Can one really approve a program without any consideration of the way it treats its staff? Should not these matters be checked as a standard practice? Is this not one of the areas where one can be most useful in formative evaluation? There is no reference to this subject in the Handbook, and little about it in the leading evaluation textbooks that I have scrutinized. Minimalist theory warns us that evaluation is in fact one discipline, although artificially separated into many. For this reason, the sub-fields often fail to take notice of the relevance of other sub-fields to a particular evaluation task which happens to come up under their own heading. It is hard to avoid the conclusion that this means the evaluation is invalid.

I. The same argument applies to the artificial separation of program evaluation from product evaluation, proposal evaluation, and policy evaluation. For example, much educational evaluation these days involves looking at programs that use computer software and hardware; but those doing the evaluation all too often demonstrate that they either nourish the bizarre notion that the programs can be evaluated without looking at the quality of the products they use (or produce), or that they lack any knowledge of the state of the art in computer product evaluation. Similarly, many large programs do some contracting via Requests for Proposals (RFPs). Should we not therefore look at the process of evaluating proposals submitted in response to RFPs, especially since aspects of its usual practice are clearly invalid (e.g., scoring them by the allocation of points out of 100)? And it is common to see program evaluations done with no reference at all to policy implications or dependencies that are obviously relevant and important for the client.

CONCLUSION

To conclude, it is hoped that a case has been made for the necessity for practitioners to acquire some familiarity with theory. The general outline of such a theory has also been provided. The claim that the material here outlines a theory rests on these features: (1) it is integrative across many fields; (2) it provides a map of the territory, based on the dictionary definition of the subject, that demarcates evaluation from many other adjoining fields with which it has been confused; (3) it provides a concise listing of the core abstract concepts of the field; and (4) the theory implies a number of results from these concepts that are of substantial value in the applied fields.

Note that the account here is not an account of, let alone a theory of, the profession or practice of evaluation, only of the discipline. Accounts of the profession, like accounts of the profession of physicists or lawyers in the U.S. today, are part of the sociology of the subject, not its content. For example, some professional evaluators spend much of their time on the activity of staff development (i.e., teaching evaluation to program staff or senior management); but this should not be interpreted as evidence that evaluation itself is or includes the teaching of evaluation, any more than mathematics should be regarded as including the teaching of mathematics just because many mathematicians spend much of their time teaching mathematics. There are other evaluators who are especially interested in explaining the results of evaluations (or lack of results) or explaining the behavior of the evaluatees. Again, these are interesting and valuable activities, but not essential parts of the process of evaluation,
although they often provide a useful and oft-requested supplement to the evaluation effort, just as a fiscal audit will often do the same.

The case made here for limitations on what counts as evaluation probably cuts closest to the bone in its cautions about recommendations. Clients (and evaluators) often assume that these will be part of the evaluation report, and sometimes they do indeed fall out of the investigation. But that is not always true, and to justify them frequently requires a great deal of knowledge to which the evaluator is not privy, e.g., knowledge about overall available funding levels at the moment of decision. (This is required because one cannot recommend continuance, or a 25% cut/increase in funding, without knowing the opportunity cost of the dollar differential.) Other apparently obvious recommendations are inappropriate because of legitimate political or personnel or social considerations to which the evaluator is not—and cannot be made—privy.

So the theory here is a theory of evaluation as such, and keeping the topic focused in this way is important, not just for logical reasons but also for the simple reason that it is hard enough as it is. Doing more often means doing it less well, which reduces the validity and credibility of the effort. A minimalist theory of evaluation, like the core theory of measurement or of relativity, deals with the core concepts of evaluation: it is not and should not include a theory of evaluation-related explanations, recommendations, education, or application. Although it lays the foundations for those and sets up the hooks for connection to them, they are secondary enterprises. The proper task of a theory of evaluation is to provide understanding of the subject matter.

REFERENCES