Logic of evaluation has significant implications for practical methodology.

The Logic of Evaluation and Evaluation Practice

Michael Scriven

This chapter sets out some definitions of common terms used in talking about evaluation and then identifies a set of crucial problems about the nature of evaluation and its components. Two of these problems are addressed here. The first is the fundamental problem of how evaluation is logically possible at all. The second is the recommendations problem: how an evaluation can provide support for recommendations, by contrast with mere evaluative conclusions about whatever is being evaluated. Some implications of each of these problems for evaluation practice are then spelled out.

Basic Concepts

The phrase *logic of evaluation* is used here to refer to the specific principles of reasoning that underlie the inference processes in all and only the fields of evaluation. The general logics of inductive, deductive, and statistical inference, although widely used in evaluation, are not part of the logic of evaluation as the term is used here, as there is nothing evaluation-specific about them. However, particular applications of those general principles may be specific to the practice of evaluation and hence fall under the logic of evaluation.

For convenience, I now make some distinctions that are not clearly implied by the common use of the terms. Fields of evaluation, where, of course, evaluative claims are the main findings, are distinguished by their evaluative subject

Criticisms are most welcome and should be e-mailed to scriven@aol.com or sent to P.O. Box 69, Point Reyes, CA 94956, or faxed to (415)663-1913. Thanks to Deborah Fournier for some valuable comments on an earlier draft.

matter. Some of the more common fields involve the evaluation of programs, proposals, personnel, products, performances, and policies, although there are a dozen other recognized fields of evaluation, ranging from the appraisal of real estate to applied ethics, literary criticism, and the evaluation of evaluations (meta-evaluation). Fields are distinguished from areas: areas are distinguished by their nonevaluative subject matter, the preexisting application discipline to which evaluation is being applied. Typical areas are education, health services, and roadtesting. Some areas can require the application of many fields of evaluation—education involves program and personnel evaluation (and several other fields), whereas road-testing simply involves product evaluation. Evaluative claims are those that attach evaluative predicates to a subject such as those from the list of fields, including a claim that a property is valuable or a program is no longer highly ranked. An evaluative predicate is one drawn from the vocabulary of grading, ranking, scoring, or apportioning—or one that is not definable without recourse to that vocabulary. Finally, evaluation types are approaches used by various evaluators or one evaluator: true experiment, advocate/adversary, goal-achievement, goal-free, naturalistic, and the like. Evaluation type is closely related to evaluation method but includes some approaches that are more ideological and less clearly defined than what are normally thought of as methodological approaches, such as empowerment evaluation.

It is useful to make a rough distinction between two levels of problem in the logic of evaluation, as in deductive logic. The first level involves defining, clarifying, and relating the key concepts via axioms and definitions—the foundations level. Clarifying the foundations includes relating these concepts to other concepts from neighboring disciplines. The second level involves developing what might be called the working principles, roughly the equivalent of useful constructions and theorems. In the work of the surveyor, for example, the analog would be the geometric construction for identifying the midpoint of a line, and Pythagoras' theorem. These provide a key procedure and a key working principle for the surveyor. This is the applications level, but it is still a level within the logic of evaluation or logical methodology. Beyond that, of course, there is the practice of evaluation and all the practical wisdom that goes with it; that is not the domain of logic but of practical methodology. In that domain, corresponding to the practical wisdom the field surveyor acquires with experience about the use and handling of equipment, we can locate generalizations about the right mix of quantitative and qualitative approaches for certain types of problem, the advantages and limitations of goal-free evaluation, how to do useful cost analysis, the relation between the probability of conclusions and their credibility, and so on.[1] The difference between the applied logic of evaluation and evaluation methodology is essentially the difference between conclusions that can be established on logical grounds and those that are based on experience and empirical evidence. The latter may therefore vary somewhat from field to field within evaluation, whereas the former must apply across all fields. Evaluation theory, by contrast with the logic of evaluation, can deal with topics or issues from either the logic or the practice of evaluation. It will generate normative theories in the first case and either normative or descriptive theories in the second case, depending on the topic.

In evaluation the foundational concepts requiring definition include evaluation itself, defined here as the systematic investigation of merit, worth, and significance; the definition of those terms; and the identification and definition of its four distinctive basic operations: ranking, grading, scoring, and apportioning.[2] These all need to be distinguished from closely related and contrasting concepts. Evaluation has to be distinguished from near-synonyms such as assessment (and two dozen others that are less similar), on the one hand, and from substantially different concepts such as monitoring, needs assessment, measurement, explanation, and description on the other; bias needs to be distinguished from commitment, formative evaluation from summative evaluation, and so on. Extensive discussion of these matters has been provided elsewhere, so they are not elaborated here (see Scriven, 1991, 1995). Instead, I concentrate on a selection from the most important logical problems that evaluation faces, problems about the logical constructions that it appears to require and the logical principles that should govern it. These problems are so serious for evaluation, and the attention so far paid to them so slight, that the discipline is running a considerable risk of building on sand, or at least failing to move in the optimal directions for development and influence.[3] It is hoped that this chapter will be a stimulant for criticisms and improvements by others as much as a contribution by the author.

The four key issues in the logic of evaluation on which some comment is provided here, perhaps the four most important problems it faces at the moment, appear to be the following.

Fundamental Problem

The fundamental problem is a construction problem: the problem of whether and how one can get from scientifically supported premises to evaluative conclusions. In the history of thought and science, this kind of inference has been widely believed to be impossible because of two alleged impossibility proofs. The first of these is best known from Hume's work, and it was later reinforced by G. E. Moore's "open question" argument. It is now possible to see that these arguments are without merit—but to see this required some relatively recent developments in logic. This problem is often seen by practitioners as rather remote from the real world of evaluation practice they inhabit, where it seems odd to suggest that every conclusion they reach may be illicit; indeed the problem cannot, realistically speaking, threaten all of them. However, the problem does raise questions that threaten the validity of many, although by no means all, of their conclusions. The specter of the impossibility proofs also explains much of the opposition to the notion of scientific or objective evaluation, and laying it to rest would therefore do much to improve the standing of evaluation in the community of disciplines and the support of leading scientists for doing and following up on evaluations.

Synthesis Problem

The synthesis problem is the problem of when and how one can integrate several subevaluations (or scores on different dimensions of performance), each referring to a different dimension of the performance or qualities of a particular evaluand[4]—or each referring to different components of the evaluand—into an overall evaluative conclusion. Systematic approaches to this problem—that is, those that do not simply consist of "experienced judgment"—usually involve the use of intuitive weight-and-sum algorithms, of which all the usual ones are invalid. Some complex efforts at a systematic approach have recently been made by social scientists under the heading of multi-attribute utility technology (MAUT), but these efforts are almost totally irrelevant to the practice of evaluation (they require data that are usually unavailable) and have serious problems of validity. (I have recently discussed this problem elsewhere, so it is bypassed here.[5]) Note that the synthesis problem we are talking about here concerns performance synthesis, problem 2A, and is not the problem of combining multiple value profiles on the part of a set of consumers into an overall or optimal value profile of "the market" or "the affluent consumer," and so on. We could call the latter the problem of values synthesis, problem 2B, and it has received more attention from social scientists (such as Kenneth Arrow[6]) without producing very useful conclusions. It does deserve more attention, although it is not as fundamental, for two reasons. First, it does not arise with the single consumer or highly congruent groups of consumers (or in many other cases), whereas the performance synthesis problem will still arise in virtually all of those cases. Second, a resolution of this problem is not needed to come to specific evaluative conclusions that are useful for individuals and groups; some of the reasons for this are given below in an example about product evaluation.

Recommendation Problem

This is the problem of when it is and is not possible to infer from an evaluative conclusion about, for example, a program (and perhaps its components), to a recommendation as to what should be done with the program (or a component). This problem has scarcely been recognized, let alone solved. It is usually supposed that such inferences are obviously valid in principle; in fact, they are invalid in principle, although sometimes defensible in particular circumstances. Because of this dependence on circumstances, it is essential to spell out exactly what assumptions are being made if recommendations are to be put forward on the basis of an evaluation. An indication of the seriousness of this problem is that unless it can be solved, essentially all research on implementation is invalid. This is so because implementation can only be of recommendations, recommendations require validation that is additional to the validation of the evaluation, and no such validation was provided in any of the implementation studies with which I am familiar.

Indicator Problem

For the most part, indicators in the social science literature are accessible (that is, measurable or observable) correlates of—but not conceptually part of—the criterion variables. (Sometimes there is a suggestion that they have to be correlates of causal variables that partially control the criterion, but this is not the general usage.) Evaluators have made considerable use of indicators of one kind or another as criteria of merit. For example, the widely used checklists for evaluation of teachers via classroom observation are usually based on research that shows that certain observable variables are correlated with improved student performances (in the so-called process/product research). However, this particular use of correlates as criteria of merit is totally illicit, as is their use in most other areas of evaluation, such as in the evaluation of programs and organizations. If such correlates were licit criteria of merit, then the use of black skin color, because it is positively correlated with crime rate, could be used as a criterion of (negative) merit in hiring. The reasons against this are not just ethical and legal, but also scientific and statistical. Perhaps the main reason this "fallacy of statistical surrogation"—the use of correlates instead of primary criteria of merit (such as the pupil's acquisition of valuable learning due to the teacher's activities)—has not been recognized, or certainly not stressed, is that the scientific sin here is one of omission rather than commission. (That is, the inference from the generalization to the particular case is licit when and only when we have no detailed job-relevant information about the evaluee, that is, where we have failed to gather any track record data for the evaluee, such as in cases that do not meet minimum standards for serious personnel evaluation.[7])

There are some significant consequences from this point in the logic of evaluation. For example, the use in personnel selection of tests that are not work samples or simulations is illicit. Once again, there are special circumstances such as emergencies or the practical impossibility of obtaining other data, where the use of indicators as criteria of merit can be justified; and there are times when very low intersubject variation in properties makes it less dangerous. The use of a brand name (and model number) in some product evaluation is perhaps the most common example of the latter case; but note the difference between using the indicator as a search heuristic (acceptable) and using it as a criterion of merit (unacceptable in general). Of course, the intersubject variation in personnel evaluation is very high. We sometimes, although not often enough, recognize the invalidity of even highly correlated indicators, for example, in blind reviewing for journals, in the derogatory phrase *guilt by association,* and in preventing juries from having access to prior criminal records of defendants. I have extensively discussed this topic elsewhere (Scriven 1987, 1990), so I will not repeat the arguments here.

Deborah Fournier has raised the interesting question whether this position on indicators threatens "theory-based evaluation," which appears to use theories about the operation of the evaluand to identify indicators for evaluative purposes. It seems possible that this would lead to indicators, which could be used only as criteria of merit in the absence of primary criteria, but the matter needs further study.

Other Problems

One might add to this list of leading problems an equal number that also deserve serious consideration, whose listing may stimulate contributions to their solutions, as well as additional suggestions from readers: the explanation problem of clarifying the relation of evaluation to explanation, the issue on which "theory-based evaluation" commits to one answer; the needs problem, which requires defining needs and needs assessment; and the parameter problem of identifying the parameters that define a particular evaluation problem, to the point where an appropriate design can be suggested (there are at least a dozen of these[8]). To these we should add problem 2B, the values synthesis problem, which is important in certain cases of policy evaluation.

Recommendation Problem

Of the two problems to be addressed here, it seems preferable to begin with this one because it has immediate practical consequences for all evaluators and because it can be handled without any significant excursion into technical logical vocabulary. The effort here is to establish its credentials as a serious problem and introduce a sketch of a treatment.

To understand the recommendation problem and its importance, one needs to see that it constitutes the focus of a fundamental, albeit often unrecognized, disagreement about the very nature of evaluation, especially program evaluation. Program evaluators fall into three groups with respect to the question of when their task is completed: we can call them the minimalists, the purists, and the activists.

The minimalists think that evaluation consists of determining the facts relevant to a decision maker's decision, such as determining what a certain group of children learned as a result of the intervention being evaluated (outcomes) or what the experience was like (process). They believe that placing a value on these facts or perceptions is something best left to the client. A sociological problem with this view is that it marks out no new territory for evaluation, because determining the facts is just traditional applied social science, and "giving the view from the inside" is a more recent (minority) view of legitimate social science. A pragmatic problem is that the client is not in a good position to do objective assembly of a very complex dataset—referring to process and outcome results—into an overall evaluation.

The minimalist view, of which Marvin Alkin is a well-known advocate, is very powerful in the field of program evaluation, as can be seen from the fact that at least two journals with the word *evaluation* in their titles publish little except reports of empirical outcome studies. As mentioned, the notion of evaluators who never draw evaluative conclusions is not only paradoxical, but it makes the notion of an autonomous profession of evaluators redundant. They are no more than relabeled social scientists, as indeed many social scientists think is the case.

The paradox—amounting to an absurdity—is unavoidable, so minimalists should not refer to themselves as evaluators; they do no evaluation, although they do some very valuable footwork for evaluations. The redundancy is suspect, because evaluators can exist and perform a useful and distinctive function in many other areas, such as personnel and product evaluation. And process evaluators, looking at programs in terms of, for example, equity or legal use of funding, work in the social sciences without much dependence on traditional social science. To think that the evaluator's role, as something distinct from data gathering, is impossible or inappropriate in program evaluation is surely just a case of being unclear on the concept.

The purists think that evaluation goes much further than minimalism, but not as far as the activists want to stretch it. They believe it does not end until the merit, worth, or significance of the evaluand has been determined. To determine merit of course requires more than merely determining outcomes or examining processes; it requires weighting the importance of each dimension of merit and then combining those results with all the relevant values, notably needs assessment data and whatever ethical, legal, or scientific standards apply. The purist thinks that the conclusion of an evaluation is a claim about comparative or absolute merit, or worth, or significance, the choice among these depending on the needs of the client. This stakes out a territory beyond that of empirical social science, but one where credentials can be established, as clearly the legitimate and skilled practice of evaluation in this sense goes back several thousand years. Beyond that point, that is, with respect to excursions into recommendations or instructions, the evaluator does not see a place for the evaluator unless he or she is also the decision maker or a remediation specialist.

The activists think that the purists are giving up much too soon, making just the same error of improper limitation that the purists condemn the minimalists for making. Activists believe that the main function, or at least a very important function, of evaluations is to generate recommendations.

The argument here is that the best first approximation is the purist position, but with a couple of significant concessions in the direction of the activists without making that an acceptable primary perspective.

Logical Problems with the Activist Position

The fundamental problem with the activist position is that it is logically unsound. From conclusions about the merit or worth of a program it is generally impossible to draw further conclusions about what should be done by the decision maker. Suppose that the decision maker is a funding agency that has asked the evaluator to look at program X, now up for refunding. The agency makes clear that it feels it still has a commitment to the population targeted by X. Both the purist and the activist agree that we can at least conclude the following:

> C1. Program X represents the most cost-effective and ethical way to meet those needs of its target population that it addresses within the financial resources of the funding agency.

What recommendations follow from this? The most likely candidate is the following:

> R1. Continue to fund X.

But this clearly does not follow. The inference depends on many missing premises. The target population has other needs besides those addressed by X; the agency might now feel it is important to pay some attention to those. The knowledgeable program officer who is an essential part of any further support of X again may have just departed. Funds may have declined since the evaluation was commissioned. And so on and on. All of these circumstances are matters about which the evaluator knows little or nothing, and about which he or she has made no systematic determination of importance. It is careless and presumptuous to propose R1 as if it is a finding from the evaluation. The evaluation finding is C1, and that represents useful input to the decision makers. They are then in a position to make a finding about appropriate action based on many considerations including the results of the evaluation.

Here is a more extreme example (not that the first example is something often encountered). Suppose that the conclusion about merit and worth points in the opposite direction:

> C2. Program X seems clearly to be greatly inferior to several alternatives and in fact seems to be almost completely ineffective.

Surely we can at least infer this recommendation:

> R2. Do not provide further funding for X.

No. The inference in R2 is in general invalid. There are many contextual factors that could make it appropriate, perhaps imperative, to continue with X. The facts about the program do not override these, and hence it does not follow from the evaluation of the program that it should not be refunded. Such factors include cases where the alternatives to X will require expensive installation and training time, for which funding is not available; cases in which there is powerful external political support for X, so that the agency may have to compromise by refunding X in order to do the best overall job with its total program portfolio, or perhaps even to survive or itself get refunded; and cases where it is imperative to avoid violent contention among board members, who are sharply divided about X, because much more important issues confront the agency. Although decisions such as the decision not to refund X are sometimes improper, they are not always, let alone necessarily, improper.

But surely none of these possibilities alter the fact that the evaluator is entitled to, indeed must, say that he or she recommends actions R1 or R2, while understanding that recommendations can be and sometimes should be overruled. To do less would surely be to ignore the plain implications of the results of the evaluation. On the contrary, not only should the evaluator not recommend R1 or R2, but it is simply illicit to suggest that they are implications of the evaluation at all. To produce them is just to produce conflict and confusion within the agency and its clientele, who, seeing the contrast between what the evaluator recommends and what the agency does, think that something improper has occurred. But usually what has occurred is that the evaluator has transcended the legitimate role of an evaluator, and that is always confusing. The confusion is compounded by the fact that many clients expect recommendations but do not provide all the additional information that would be required to make them possible. Of course, even the client cannot provide the expertise that is also required.

Suppose one were to argue that a proper reconstruction of the evaluator's recommendation would show it to have an implicit premise of the form "other things being equal." This reconstruction is often described as saying that R1 or R2 are conclusions with prima facie force, not categorical force (or that they are prima facie implications of C1 and C2). There is indeed a place in the logic of evaluation for the notion of prima facie inference, as shall be seen in the following discussion of the fundamental problem. But that place is not here. One might as well say that a jury's conclusion of guilt should have a recommended sentence attached to it. Judges are in possession of facts, experience, and knowledge of context, prior records, and law that juries do not have. Conclusions about guilt do not justify even prima facie assertions about sentences (with rare exceptions, such as where the sentence is legally tied to the finding of guilt). The bottom line is that the strongest legitimate conclusion from the evaluator is the conclusion about the merit or worth of the evaluand. The consumer and the client know that this conclusion has a certain relevance and value to them in the course of decision making, but the rest is up to them.

The confusion created by the opposite assumption is not confined to the client and the clientele. Evaluators and researchers have also been seriously confused by the fallacious suggestion that evaluations should always or normally end in recommendations. Most of the fuss in the evaluation literature about the frequent failure of clients, especially government clients, to "implement evaluations" was based on a confusion between evaluations and recommendations. Evaluations cannot be implemented. They can be taken into account, but they are not the only source of information relevant to decisions—including ethical and rational decisions—about the future of programs. There is little evidence that evaluations are not generally taken into account, to an appropriate degree, except where they have become irrelevant since the evaluation was commissioned, other factors having overdetermined the decision. It would be extraordinarily difficult to discover such cases through scientific investigation—it certainly can't be done with the usual blunt instrument designs—but once in a while they turn up, occasionally through confessions. Such cases are often cases where the client's intent was nefarious—to get a justification for a hatchet job, which the evaluation did not supply; or the converse cases, where a justification for favorable treatment was sought, which the evaluations did not provide. Those few cases are important because they reveal the true values of the client. They are sometimes highlighted by firing the evaluator but it is much more common for them to be simply concealed by burying the evaluation or by going through a charade of treating it as input and overweighting "other considerations" for not going in the direction it points.

Micro- Versus Macro-Recommendations

The error of jumping to recommendations is not confined to recommendations about the disposition of programs—what might be called macro-recommendations. Evaluators often make recommendations of a less global kind about what should or should not be done to improve various components or aspects of a program (micro-recommendations). These, too, are usually not implications of the evaluation but jump beyond the evaluative findings, although the subevaluations on which they are based are well founded. The missing premises required to bridge the gap typically include assumptions about the absence of interactive effects from this action on other aspects of the program and assumptions about the remedial effect of the suggested changes, something about which the evaluator is rarely an expert.

It is perhaps easiest to see this by using an example from product evaluation. A good automobile road tester has considerable expertise as an evaluator, and this expertise entitles him or her to come to conclusions about the merits of a car along certain dimensions (handling, reliability, and so on) and, by putting these together, to overall conclusions about the merit of the car (comparatively or absolutely). Additionally, it is legitimate for the road tester to draw evaluative conclusions about certain components, such as the brakes, the headlights, and the wipers, and the legibility of the instruments. These are conclusions about the extent to which these components perform the appropriate functions, and the evaluator has plenty of experience on the basis of which it is possible to say how the level of quality of these components rates against state-of-the-art standards, against minimum acceptable standards, and against the norm, in the price or functional class of vehicle being tested. Many evaluators who have defended the importance of recommendations as part of the evaluator's task in formative evaluation have failed to realize the enormous value of these subevaluations as feedback to the manager or developer looking for guidance toward improvement. Recommendations are not the only path to improvement, and micro-evaluations are usually more legitimate than any recommendations, whether macro or micro.

The moment the road tester begins to make specific recommendations, whether to the manufacturer or to the consumer, the path of justification becomes much rockier. To recapitulate the previous point, the macro-recommendation that everyone about to buy a new small econo-sedan should buy an X—or that no one interested in a new small sedan should buy an X—simply presumes too much about contextual variables. Should they buy it even if they really dislike its looks? If they have to wait a month for delivery? If it lacks a ski pass-through in the back seat, although they are avid skiers? The macro-recommendation is an assertion aimed at a large group of people who vary considerably in respects that are relevant to purchasing. By contrast, the conclusion that this is the best, or worst, car in its class, averaging across the needs of many consumers or idealizing various types of consumers from a mixed population, swings the focus back to the car, identifying it as the best of its type. Naturally, consumers look at their idiosyncratic needs and preferences in the process of applying this evaluative conclusion to their own case. Aesthetics and small-group needs are rightly set aside, or given a passing reference, in professional evaluation of functional artifacts.

For such reasons, road tests often conclude with cautious remarks such as, "This car should be given serious consideration by anyone looking for an economy sedan," which is a recommendation but not much more than a rephrasing of the evaluation. It is a truism that anyone considering the purchase of an expensive product should (rationally speaking) look at relevant facts and evaluations. People making decisions about the future of programs should consider the results of evaluations of those programs. That is trivial: it is always true, but it does not tell decision makers that they should do one or the other of the things they can elect to do. It is a long way from the specific recommendation to refund or withdraw support, and it is these specific recommendations that are often thought by evaluators to follow from the evaluation. I will come back to this in a moment.

First, however, I will follow the product example through to the micro-level. The argument is the same as at the macro-level. It is proper to draw any of the following conclusions about the headlights in the Lexus LS400. They are weak; too weak for safety in senior citizens, whose sensitivity to light goes down by 50 percent; well below the state-of-the-art standard for luxury cars, set by the BMW 750il with its xenon lights; or below the norm for luxury cars, defined, say, as those over $40,000. But the suggestion (micro-recommendation) that one should put xenon lights in the LS400 is unjustified (under heavy loading conditions, they require load-leveling rear suspension to avoid blinding oncoming drivers, which would price the 400 out of its market). One could of course recommend that "something be done about the lights," but that scarcely qualifies as the kind of recommendation that clients seek.

Diffuse Versus Specific Recommendations

This leads to a general point. One of the reasons why it is tempting to think that evaluations entail recommendations is that they can reasonably be said to imply this kind of phantom recommendation, essentially a mere restating of the negative recommendation. So we need to distinguish between these recommendations and recommendations of defined feasible actions. I will call these two types of recommendation diffuse and specific. In these terms, "Interior storage space needs to be increased" is diffuse—just a translation of a poor evaluation into a ghost of a recommendation—whereas "Door pockets should be added" is specific. Another example of a diffuse micro-recommendation would be, "Some way should be sought to improve fuel consumption." This is, so to speak, a way of pointing out the evaluative conclusion (that fuel consumption is not very good by the relevant standards) in the direction of the manufacturer. Because the manufacturer is always looking for ways to improve fuel consumption to avoid the federal penalties on high corporate fuel consumption and already knows the fuel consumption of the model line courtesy of the federally required tests, this is essentially an empty recommendation. An example of a specific kind of micro-recommendation of a highly relevant and potentially useful kind is, "In order to improve fuel consumption, the present intake fuel injection system should be replaced with throttle-body fuel injection." Although it is a much more useful kind of recommendation, it does not follow from the evaluation in a road test. It might of course follow from an evaluation of the options available to the manufacturer, if this were done by an engineer rather than a road-tester, and if the evaluator were given a mass of data not available to road-testers. Analogously, program evaluators are not, per se, social engineers. If they acquire the kind of detailed knowledge that the theory-driven approach suggests, then they have become engineers; but this is more than most evaluators can achieve except by severely limiting the range of programs they can evaluate. There are big differences between health programs and education programs in terms of local expertise, and between types of health programs and types of education programs. Furthermore, there are serious grounds for skepticism as to whether there are reliable program theories in many of these areas.

Particularly with respect to micro-recommendations (for example, increase the size of the glove box, or lighten the clutch pedal), there are many problems about such changes that the evaluator is rarely in a good position to estimate: problems about its cost, about increased complexity of the production engineering of the change (and hence a risk of reduced reliability), about legal liability, and about consequential costs (larger glove box means less leg room, and so on). It seems clear that the evaluator per se should report on the merit of the glove box and the clutch pedal effort but should keep to that. Such remarks have their own significance for actions by the manufacturer and consumer but in each case have to be fitted into a complex composition that will always require some compromises. In general, making evaluative remarks is all that is legitimate. If, however, it seems more useful, the symbolic recommendation that adds nothing more specific may be chosen as the way to express the same point.

Middle Ground: Where Specific Recommendations Can Be Justified

The gap between the purist and the activist can be bridged in two ways. The first is by expanding on the point just made about the value of focused (micro) feedback for the improvement effort. The second is by spelling out four types of situations for which the evaluator can legitimately add recommendations.

Evaluator Has Expertise at the Micro Level with Respect to This Kind of Evaluand

For example, say that the road-tester is also an automobile engineer. In these cases, the recommendation is then dependent entirely on this expertise and is an add-on to the evaluation. Essentially, the evaluator is saying, "As an evaluator I can show that X needs improvement, and as an engineer it seems to me that Y is the way to improve it, so I recommend Y." The problem with this, as can be seen by looking at what goes wrong with accreditation of professional schools, is that the second part of it is often more a matter of opinion than of expertise, but it receives extra and illicit weight because the first part is provable. In road-testing, this problem can be put by saying that the car was designed by a team of competent engineers, so the opinion of one more, who happens to have been the road-tester, cannot be automatically accepted simply because the road test shows the need for further improvement of this component. Nevertheless, there are some cases in which the contribution from the evaluator with a second area of expertise is useful, provided both sides are clear that the recommendation requires a separate justification from the evaluation and a great deal of extra information not usually available to the road-tester. The evaluation shows that something needs to be improved, but as to how this is to be done, the evaluator becomes just another micro-level consultant.

There are one or two interesting cases in which the gap requires little bridging. For example, in evaluating the self-evaluation process in a program, it takes no extra expertise to recommend, for example, that the program make some use of external evaluators, because evaluation is one's own field of expertise. In another kind of case, in which one discovers improper hiring and promotional practices that discriminate against, say, women, by penalizing longer time-to-doctorate without taking any account of maternity leaves, one not only can but must recommend a change in the process, because ethical transgressions are stand-alone faults. However, the way in which that change is to be brought about is beyond the expertise of the typical evaluator.

Evaluator Has Full Decision-Related Data and Expertise at the Macro Level with Respect to This Kind of Evaluand

For example, say the evaluator works for a foundation and has acquired extensive knowledge of its overall values and resources. After evaluating a particular program, which turns out to have little value, the evaluator is able to see that nothing in the foundation's decision environment provides an overriding reason for retention. Again, the point here is that two areas of expertise are involved, and challenging the second one is not the same as challenging the first. The evaluator may simply be wrong about the political ambiance but completely correct about the evaluation of the program. (Of course, so may the foundation's decision makers, and the situation is complicated by the fact that they may not have a single view.)

Needs Assessment Data Indicate a Homogeneous Population

For example, say the client is an individual who asks for a recommendation, such as which graduate program to enter or which car to buy or whether to hire Jones, Smith, or Robinson. In that situation it is possible for a good evaluator to run a needs assessment on the client that will focus the evaluation to the point where it implies a recommendation. This case extends to the case in which the client asks for an evaluation of options for another group, not including the client, but the other group is homogeneous with respect to the relevant variables. For example, they are all terminal cancer patients with essentially the same life expectancy and the question is what analgesic is best for them. This situation is rare in program evaluation.

Client Asks for an Evaluation of the Decision Alternatives

This is a further step beyond the two previous cases. It is much harder than just evaluating one of the entities that is involved in the decisions, roughly because it involves becoming a macro-expert in the decision space where the client is operating, as well as doing an evaluation of the various evaluands floating around in that space, not to mention thinking up new alternatives. As in case 2, the bottom line is that there are two levels of evaluation involved, and the validity of the inferences and data for each should be considered separately.

Evaluator Wants to Convey a More Specific Estimate of Worth

Even when evaluators lack full knowledge of the decision space for the decision makers who are their ultimate customers, they will sometimes feel uneasy about the limited vocabulary available to them to express worth and may use the recommendations as a way to convey a sense of their estimate. If recommendations are used to serve this function, it should be (but often is not) understood by the evaluator that they are very likely not to be implemented due to factors beyond the evaluator's ken. Thus "This program should be discontinued" has to be read as meaning "Nothing that we know about could justify continuance."

Conclusions

The step to specific recommendations requires separate expertise and certification (of data and inferences) beyond the skills required for an evaluator. It should not be taken, even if requested by the client, without demonstrating that expertise and providing that certification. The purist is right to say that evaluation expertise of itself does not justify this step (very special cases apart); the activist is right to say that we should try to bridge it when we can (that is, when we have the expertise); the cynic is right to say that all too often the evaluator is seduced by the client, or precedent, or the wish to help, or an overoptimistic view of his or her own expertise into thinking that it is a small and natural step from the evaluation to the recommendations.

In case it appears that the preceding argument makes a strong case for using evaluators with expertise in the field of the evaluand, this may be an opportune moment to stress that such expertise nearly always carries with it a lack of experience in looking at the field from the outside, such as from the consumer's point of view, or the point of view of another field competing with the first one for resources, or the point of view of someone without connections to the old boys' network in the field. And the time it takes to acquire the local expertise usually means less skill acquired in general-purpose evaluation. A good principle of practical methodology suggests that teaming local expertise with general evaluation expertise is the best solution.

Fundamental Problem

This is the problem of providing a legitimate basis for inferring to evaluative conclusions: that is, a basis other than arbitrary evaluative premises. This means using either premises that are factually established, theorems from deductive logic or mathematics, or commonly accepted definitions. Although this problem threatens the validity of all evaluative conclusions, logically speaking, few evaluators are seriously concerned about it. Like mathematicians informed that there is a crisis in the foundations of mathematics, they feel that this is a philosophical crisis rather than a real one, and their common sense tells them that it will be resolved in their favor. This is true in a sense; much of their work will be left intact, such as impact data collection. But it is also true that the exact way in which the issue is resolved will have some significant consequences for practice, notably in defining the range of legitimate activities for the evaluator. Solutions to the recommendation problem, for example, depend on how the fundamental problem is solved. It is also important that resolving the issue in a convincing way is likely to have a significant effect on the status of evaluation among the company of disciplines, because, in the eyes of many others, the specter of invalidity has long haunted this field.

The problem is simply explained. We can, by the usual process of scientific observation and inference, reach empirical nonevaluative conclusions. We can also, by deduction, reach mathematical and logical conclusions from common definitions, and from the axioms of mathematics and logic, which are themselves either definitions or self-evident in some strong sense. It appears that none of these premises or conclusions are evaluative propositions, and none of these inference rules bring them in, and hence it appears that one cannot reach any evaluative conclusions from such proofs. Thus it appears that the only way to reach evaluative conclusions is to add evaluative premises. These will, it appears, be arbitrary, because no evaluative premises can themselves be established by these processes of logical, mathematical, or scientific inference. And if the premises are arbitrary, one might equally well adopt contrary premises and derive contrary evaluative conclusions. Hence no evaluative conclusions can be established that deserve to be treated as any more true than their denial; that is, no evaluative propositions can be established that meet the minimum requirement for true propositions. This is a development of the Humean argument, and its conclusion is what I will call the dismissive view (of the status of evaluative claims).

The main weakness in this line of argument—and in Moore's rather different one leading to the same conclusion—is that it is based on grossly oversimplified notions of definition and inference, notions that have been refined since Moore's time. That there is something wrong with the arguments is clear enough if we hold firmly on to our common sense, because of their reductio ad absurdum (overkill) nature. The absurdity is clear enough from, for example, the fact that according to the dismissive view all published product evaluations are guilty of a logical fallacy, although they proceed by good inference from facts to evaluative conclusions. Their missing premises are the common conceptions of the common products they evaluate, such as knives, radios, and houses, plus a general understanding of the concepts of merit, worth, and significance—that is, what it means for something to be better or worse among its kind, more or less valuable or significant. But seeing that the dismissive argument and view are unsound is one thing; seeing why they are unsound takes a little more work. The extra work is worthwhile because it shows exactly what the mistakes were that led to the view, and once seen, these are deprived of any residual charm. I will approach the refutation by reconsidering the classical views of definition and inference. The sequence is important because the reanalysis of definition is the key to the reanalysis of inference. This is not easy reading for those who do not remember their logic courses, but it is important and applies to almost all evaluands and evaluations.[9]

Definitions. The concept of a definition in traditional logic—still to be found in most current texts—was that it consisted of an equivalence between the defined concept, X, and a set of terms, C1, C2, and so on, that could be substituted for that concept without loss of meaning. This meant that (1) each member of this set of Cs was a necessary condition for the concept, often put by saying that they are "individually necessary,"[10] and (2) the set as a whole was jointly sufficient (that is, it is a contradiction to say that all are present but the concept does not apply). This was indeed true of concepts in mathematics, from which the notion came, and it is true of many neologisms at the time of their introduction. But it is a misrepresentation of the situation with most everyday concepts (such as apple or friend) and all the major concepts in science (temperature and so on), as well as most of the more complex concepts in the dictionary. These are cluster concepts, that is, ones whose meaning is learned and explained by reference to a number of criteria that do not constitute a set that is jointly sufficient and individually necessary. The set of criteria for X has three weaker properties:

1. *Although some of the criteria may be necessary conditions for X, in many cases none of them have that status.*
2. *Although some subsets of criteria from the complete list must be sufficient—for otherwise we could never know when to apply the concept X—they are sufficient only for all practical purposes, but not with logical certainty.*

That is, it is conceivable, although unlikely, that such a set, which everyone thinks is sufficient to show that X is present, is really not sufficient. Moreover, the complete list of criteria is often very long, often not known to anyone (in the sense that anyone could list it when asked), often contains functional accounts of the way the term is used (for example, with terms such as *and* or *large* or *good*), and often contains contradictory components. Of course, at most one of the contradictory components will be present in any particular case. Contradictory components, familiar enough in the definition of disease entities, are usually of the form C5 or (not-C5 and R). For example, disease X is characterized by symptoms C1 and/or C2, and so on, and by C5 (a mottled pink skin rash) or not-C5 (no rash but small subcutaneous nodules (C6)).

3. *Although the set of criteria is not sharply defined, it is loosely distinguished from other factors (known as indicators) that may often be found in conjunction with the defined concept by the fact that criteria are, and indicators are not, sometimes put forward and defended as part of the meaning of the concept.*

For example, sunrise is a preoccurring indicator for sunset but not a criterion for it. The line between criteria and indicators is obviously not sharp, and it shifts with time and our growth of knowledge. Criteria also vary in weight at a given time and over time, and there is some agreement on gross differences in weight among experienced users of the concept, but nothing like a complete or precise account of these weights.

This imprecise relation between a concept and its meaning applies just as much to the key concepts in science as it does to well-known slippery notions in everyday talk, such as *democracy* or *equality*. This can be seen in the various definitions of temperature that have reigned in the history of thermodynamics. The definitions did not encapsulate the whole meaning of the concept; they were just leading criteria that it was convenient to enshrine in a definition for a while, until the theory to which they were connected began to look a little shopworn. Then we called into service some of the other criteria that had been occupying a background role for a while, often embedded in a theory, and elected them to definitional status.

Many of us puzzled over the allegedly empirical status of Newton's Laws of Motion, especially the third law ("For every action there is an equal and opposite reaction"), thinking that it looked very much like a definition of *reaction*, but in fact its use in the parallelogram of forces construction in statics gives it some indirect empirical content (that is, buildings constructed on that basis tend not to collapse). Hence *reaction* is a concept that can be defined criterially but not in terms of a classical definition. This is indeed the most common situation with deeply theory-embedded concepts in the sciences: whatever definition was used to introduce them originally, their present meaning is more complex, and usually only specifiable in terms of criteria.

Inference. What is the nature of the inference from a set of apparently adequate criteria to the concept itself? As stated above, although this is sometimes a deductive step (when the term is classically definable and we have listed all the criteria that occur in the classical definition), in practice it is often just a very reliable but not absolutely irrefutable inference. Yet it is more than that, too, because the criteria (and their relations) are, collectively, all there is to the meaning of the concept. So the inference is at least nearly as strong as inference via a classical meaning rule, for example, from *three-sided plane figure* to *triangle.* That is, it is nearly as strong as deduction (sometimes it is deduction), and it is not just an empirical inference from indicators that are connected by mere correlations with the concept. This kind of inference is a leading type of what has been called "probative inference."[11] It is neither deduction nor plain inference from observed generalizations, and it is the key to handling the logic of evaluation.

Payoff Time. Suppose we are evaluating a cooking pot. If we understand the meaning of the term, then we know what it is for (its purpose), as this is something whose definition essentially brings in function.[12] Knowing what it is for and knowing what it means in general to do something well (or badly), we can quickly generate a comprehensive list of dimensions along which a cooking pot can be ranked for merit: leakproof, able to withstand the heat of fires and flames (and preferably ovens), not easily cracked or split by typical kitchen tools and treatment, preferably able to withstand being dropped, preferably corrosion resistant and inert to chemicals contained in food or generated in the cooking process, preferably easy for most cooks to lift, preferably with lifting handles that stay cool, preferably with a lip that allows pouring without spilling, and so on. Someone who knows the meaning of the term without ever having used a cooking pot will recognize that this list covers most and perhaps all of the dimensions along which pots vary in merit. If we add the desirability of low cost, we have a checklist for evaluating worth or value. If we add aesthetic considerations, we start to consider the pot as an art object or as in part an art object, and we find that that dimension allows only poor objectivity.

Hence, from the common meaning of the term and the common meaning of the concept of merit, we can generate enough criteria of merit to allow a solid inference from high performance on each or most of the dimensions on that list to the conclusion that the pot is a good (or bad or better) one. Determining whether the performance is high involves determining standards for *high* and then applying them. The standards mainly or entirely emerge from the definition (for example, *absolutely leakproof* is as high as one can get on that dimension). That is a definitional matter, involving the definition of a term with a common meaning, and hence not vulnerable to the charge of arbitrariness. Applying these standards is an empirical matter, and hence not vulnerable to those embracing the dismissive thesis (the value-skeptics).

It will take knowledge of the market prices to infer to good value, and knowledge of the range of pots available to infer to most, although not all, judgments of relative merit. But that further knowledge is empirical knowledge, readily obtainable and intersubjectively testable, like the most reliable scientific knowledge.

Now the objections to inferring from facts to values have been eliminated. The inference is probative, not deductive or classical inductive, but good enough to bet the farm on, and it is based on understanding two concepts—pot and merit—in the common language, which cannot be classically defined, only criterially defined.

What comeback from the value-skeptics is possible? The premises that brought in the values are at most of the form, "The more leakproof a cooking pot is, across a range of several days' use, the better it is." Or "The lighter a cooking pot is, within the range that current materials make possible, then, other things being equal, the better it is."[13] These are provably true, and the proof comes from an understanding of the concepts involved and perhaps to some extent an understanding of the context of use. To challenge these as arbitrary statements, which could just as well be denied as accepted, is to be speaking another language, not our native tongue.[14]

Conclusions for Practice

The discussion of inference here shows that the logic of evaluation at most requires identification of the criteria of merit for the evaluand, not a classical definition of either the evaluand or merit.[15] This suggests that the effort to set out the criteria of merit being used should be made more explicit in many evaluations. Doing this in the case of program evaluation, for example, makes it rather obvious that one cannot plausibly restrict the criteria of merit to those based on the goals and objectives of a program. It is necessary to brainstorm other possible effects and functions that may turn up as side-effects, just as in the evaluation of cooking pots one may thereby think of the fact that pots are typically more useful if other pots can nest inside them, thus greatly reducing the space required for storage, or of the fact that they are more useful if they have a handle made to facilitate hanging the pot, say, on a hook. This approach to side-effects should complement the usual one of looking for them blindly or through open-ended questions in interviews.

The bottom line of the inference section here is of course to establish not only the possibility but the relative ease of rock-solid inferences from unchallengeable definitional and factual premises to evaluative conclusions. This type of premise is what consumer product evaluation has always used, and it is the type that the professional program evaluator and personnel evaluator should be using.

The discussion of definitions, apart from serving as the basis for the introduction of probative inference, also emphasizes the possibility of providing entirely satisfactory definitions of complex concepts without submitting to the tyranny of classical definitions. It is only necessary to indicate as many criteria for the term as possible, just as in evaluation we try to list as many criteria of merit as possible. It is not crucial that we find them all, but it is sufficiently important that we find nearly all of them that we should try out our attempts as hypotheses, in this case hypotheses about meanings, so the trial should involve a sample of experienced users of the terms.

The section on recommendations also has implications for practice, which are spelled out there in some detail, but can be summarized in the leading sentence of the concluding section of that part: The step to specific recommendations requires separate expertise and certification (of data and inferences) beyond the skills required for an evaluator. Although some evaluators have this further expertise, many and perhaps most of them are compromised as evaluators by the ties they develop in the course of acquiring the expertise. Hence, evaluators should never feel that they are on balance weaker as evaluators if they lack local expertise about the particular type of evaluand. What they lose on the swings they make up on the roundabouts. However, they should make an effort to form an evaluation team with a locally expert evaluator or use local experts as consultants to the evaluation (two very different approaches).

Overall, it is hoped that the chapter shows how the logic of evaluation is not only of intellectual importance as the backbone of the discipline of evaluation but worth studying for its significant implications for practical methodology.

Notes

[1] A longer list of principles, for one field of evaluation, is proposed in Scriven (1993). A good example of a relatively undiscussed problem in this area is disclosure, that is, the problem of how much and when the external evaluator should show drafts or final evaluation reports to those who are evaluated (see Scriven and Kramer, 1995).

[2] There is a rough relationship between this set of concepts and the fundamental concepts in the logic of measurement: nominal, ordinal, interval, and ratio scales. However, only two are closely equivalent (ranking provides an ordinal scale on merit, worth, and so on, and the usual type of scoring supposedly, although in fact rarely, provides a ratio scale).

[3] Three of the examples here are construction problems, and construction problems sometimes prove impossible. A good example of this in geometry is the problem of squaring the circle, on which a vast amount of time was wasted before an impossibility proof was discovered. We need to decide what constructions are and are not possible in evaluation, a task for the logic of evaluation.

[4] The term *evaluand* is used here to refer to whatever is being evaluated, such as a program, performance, person, proposal, product, or policy. *Evaluee* is used to refer to any person who is, is part of, or is a stakeholder in an evaluand, such as the staff of a program being evaluated. (*Evaluatee* has some currency as a term for the person evaluated in personnel evaluation.)

[5] The background of the problem is discussed in some detail, and its crucial importance is illustrated by reference to a number of familiar examples, in "The Final Synthesis," in Scriven (1994). A crude version of a technical solution, using a qualitative weighted sum approach, is in Scriven (1991). It is much improved in the fifth edition (1995) and is also available by request to those sending a stamped self-addressed envelope. These references contain a justification for the remarks about the invalidity of the usual weighted-sum approaches and the irrelevance of MAUT.

6. Arrow's impossibility theorem has some connection with this issue. It is a trivial result that, contrary to various popular interpretations, has no implications for the nonfeasibility of democracy.

7. When we do have track record data, we "know too much" to be able to use a generalization that applies only to a population defined more generally. This is clear in the limit case, where we have a paradigm of a good teacher, that is, one who scores outstandingly on all primary criteria but who goes against the research-based indicators. (We know there are many of these teachers because the correlations are quite modest.) Should we mark them down because they do not follow the rules? This would obviously be unjust and, on consideration, absurd, because the correlations were established based on an independent criterion of good teaching, that is, doing well on the primary indicators. We can hardly use the indicators to invalidate the very examples from which they draw their own, secondary legitimacy.

Suppose blue-eyed golfers are statistically better than brown-eyed ones. That fact can be used to make a weak inference to merit as long as one does not also have the testimony of good witnesses as to the golfer's average number of strokes per round in the current season. At that point, it is ridiculous to argue that that testimony should be downweighted against information that someone has blue eyes. The situation is not improved by using the elegance of someone's swing instead of eye color, that is, a criterion that has some plausible connection with merit. Once usable evidence about primary criteria of merit is available, the evidence about secondary criteria (correlates of primary criteria) becomes worthless.

The only exception occurs for a generalization that relates to a subclass defined by common evidence about primary criteria. If all we know about Janet is that she is a teacher who uses high time-on-task, we can infer that she is more likely to be successful than Robert, about whom all we know is that he scores low on time-on-task. But if we also know that Janet has mixed references from previous supervisors, whereas Robert's are uniformly good, we cannot offset that by weighting Janet's style in her favor. She lost on the primary indicators and is no longer part of a population of teachers about whom we have only secondary indicator data..We can now use her time-on-task in her favor only if we also have a generalization that says teachers with weaker references are better than those with better references if they exhibit high time-on-task. In short, so-called research-based indicators cannot be used once we have a typical dossier, which includes evidence about primary criteria of merit (knowledge of subject matter, communications skills, assessment skills, ability to relate to peers and parents, and so on); that is, they are of no use in practical cases.

8. There is a first try at this in Scriven (1989), which is condensed in the entry on Evaluation Parameters in *Evaluation Thesaurus.*

9. The following analysis is a development of the one originally set out in Scriven (1959). That account represented an attempt to render more explicit Wittgenstein's analysis of meaning. The entry under Logic of Evaluation in Scriven (1991) provides more detail but is slightly less accurate.

10. Saying that a criterion is necessarily connected with a concept is to say that it is a contradiction to assert that the concept applies to something but that the criterion does not. However, contrary to another classical notion, a contradiction is not sharply distinguishable from a highly implausible or almost meaningless claim. Nevertheless, it remains a useful notion because there are plenty of clear cases of contradiction even if there are some debatable borderline cases.

11. The legal term *probative* is defined in the *Oxford English Dictionary II* as "having the quality or function of proving or demonstrating; affording proof or evidence; demonstrative, evidential." The term was introduced to logical theory in Scriven (1987a). Probative inference also covers cases of inference to a conclusion that is established beyond a reasonable doubt but not to a conclusion that represents only the best interpretation of the data, the balance-of-evidence criterion.

12. The *Oxford English Dictionary II* defines *pot* as "a vessel of cylindrical or other rounded form, and rather deep than broad, commonly made of earthenware or metal (less commonly glass), used to hold various substances, liquid or solid, for . . . cooking or boiling." Of

course, this isn't a classical definition; we would have no trouble in calling a square pot a pot, or a fireproof plastic container a pot. In fact, for many years cooking pots in some cultures were commonly made of bark, but that does not show the definition to be incorrect, owing to the insulating function of the term *commonly*. In the list that follows, the term *preferably* serves an insulating function.

13. Actually, one needs only a weaker claim: "Cooking pots that have many of these properties are better than those with few of them." To establish at least one evaluative conclusion, one needs only "Cooking pots lacking all these properties are not good," or the converse.

14. I have omitted all discussion of Moore's open-question argument, which raises no new issues. It works only if we assume that the only legitimate form of inference to evaluative conclusions has to be deduction and the only legitimate form of definition has to be classical. Once one sees that only probative inference and criterial definitions are required, here as in all nonartificial discourse, the force of the argument evaporates.

15. As a matter of interest, there are a few cases where the dismissive thesis is completely wrong on its own terms because there are a few arguments with evaluative premises that are definitional truths and not arbitrary. Hence there are a few arguments that can legitimately proceed by using these propositions, along with factual claims as the other premises, and reach nonarbitrary conclusions. For example, the proposition "murder is wrong" is true by definition because any case of killing for which there is a justification is not a case of murder in the commonly used sense of that term. Hence, we can establish the conclusion that X is a murderer by using this premise plus factual premises that rule out all of the finite list of justifications for killing. John Searle has used the premise "deliberately breaking a promise is wrong" for the same kind of argument. Some hard fighting has gone on in the philosophical literature to avoid this crack in the wall between facts and values, but most would agree that it has not weakened Searle's case.

References

Scriven, M. "The Logic of Criteria." *Journal of Philosophy*, 1959, *56*, 857-868.

Scriven, M. "Probative Logic: Review and Preview." In F. H. van Eemeren, R. Grootendorst, J. A. Blair, and C. A. Willard (eds.), *Argumentation: Across the Lines of Discipline. Proceedings of the Conference on Argumentation*. Dordrecht, The Netherlands: Foris, 1987a.

Scriven, M. "Validity in Personnel Evaluation." *Journal of Personnel Evaluation in Education*, 1987b, *1* (1), 9–23.

Scriven, M. "The Logical Specifications of an Evaluation." *Australasian Journal of Evaluation*, 1989, 2 (1), 31–38.

Scriven, M. "Can Research-Based Teacher Evaluation Be Saved?" *Journal of Personnel Evaluation in Education*, Fall 1990.

Scriven, M. *Evaluation Thesaurus*. (4th ed.) Newbury Park, Calif.: Sage, 1991.

Scriven, M. (ed.). *Hard-Won Lessons in Program Evaluation*. New Directions in Program Evaluation, no. 58. San Francisco: Jossey-Bass, 1993.

Scriven, M. "The Final Synthesis." *Evaluation Practice*, Oct. 1994.

Scriven, M. *Evaluation Thesaurus*. (5th ed.) Newbury Park, Calif.: Sage, 1995.

Scriven, M., and Kramer, D. "Risks, Rights, and Responsibilities in Evaluation." *Australasian Journal of Evaluation*, Jan. 1995.

MICHAEL SCRIVEN is an attention-span challenged scholar who has spent large amounts of time in departments of mathematics, philosophy, psychology, and education; and in developing the new disciplines of technology studies, computer studies, evaluation, and informal logic. He is now director of the Evaluation & Development Group, a small consulting company in Inverness, California.