21

The Unruly Logic of Evaluation

Michael Scriven

1. Overview

The interest of this topic from the critical thinking/informal logic point of view is that it refers to the huge and important range of common and professional discourse aimed at supporting or contesting evaluative conclusions, and it raises the possibility of greatly improving the quality of such discussions by means of a largely informal logic approach, an especially interesting challenge given the failure of deontic logic to achieve that end.[1] As a practical goal for work in this field, though this particular presentation is not long enough to justify its attainment, one might propose as an aim for work on this topic that it should try to ensure that every respectable text in critical thinking/informal logic (and why not every text in the philosophy of the social sciences and ethics) will include a section on the logic of evaluation.[2]

Evaluation is the process of determining merit, worth, or significance (m/w/s); an evaluation is a product of that process.[3] Professional evaluation is evaluation done in a systematic and objective way with a degree of expertise that requires extensive specific training or learning. The logic of evaluation is concerned with (i) how, if at all, professional evaluation is possible; (ii) its

[1] See the valuable survey in the *Stanford Encyclopedia of Philosophy* (online) under this heading.

[2] With that goal in mind, the author would greatly appreciate receiving comments, critical or amplificatory; they can be sent to mjscriv1@gmail.com.

[3] This definition is a synthesis of the definitions in most dictionaries and the professional literature, with the unfortunate exception of the mother of them all, the *Oxford English Dictionary*, which seems unable to grasp the move beyond estate appraisal and mathematical formula calculation, despite many decades of prompting. (Recent editions of the single-volume Oxford dictionaries are somewhat less perverse.)

© Michael Scriven

nature and its location in the organization of knowledge, and (iii) the logical structure of its inferences.

Four groups of issues in this branch of logic will be addressed here, even more briefly than they deserve, the first two of them philosophical, the third conceptual, and the last procedural. The philosophical issues concern the barriers that have been supposed to render the entire enterprise impossible—variants on (a) the so-called “naturalistic fallacy”, and on (b) the alleged impossibility of objectivity. The conceptual issues concern various attempts to capture the essence of professional evaluation—either its nature or its operation—usually via a key metaphor. These issues, and the procedural issues, which concern how the practical logic of evaluation does and does not operate, are naturally of interest only if it is not founded on a fallacy. We’ll stay with that sequence, except for beginning with one of the easier conceptual issues—the geography of evaluation.

2. The territory of evaluation

Evaluation has an extremely extensive territory, since it includes the substantial portion of everyday discourse devoted to proposing, attacking, and defending evaluative claims about food products, football teams, human behavior, global warming, and almost everything else. The domain of professional evaluation is still very extensive: we here distinguish seven standard sub-divisions of it, and four other specialized domains which are less commonly categorized or recognized as part of evaluation’s domain, although substantially devoted to it.

The standard sub-divisions, which we’ll call the satellite subjects, are: performance evaluation, product evaluation, personnel evaluation, proposal evaluation, program evaluation (which includes practice, procedural, and process evaluation), policy analysis, and portfolio evaluation (there is a small amount of overlap between some of these). The specialized or core domains include two for which I coined names now quite widely used—meta-evaluation (the evaluation of evaluations), and intradisciplinary evaluation (the evaluation of data, experimental designs, hypotheses, etc., which is part of normal professional practice within every discipline). The other two specialized domains are standard disciplines not commonly classified as divisions of evaluation but substantially concerned with that activity as a main subject matter, not merely for procedural (i.e., intradisciplinary) use: ethics (normative ethics in particular) and logic (because it includes the evaluation of arguments).

The logic of evaluation has general features and processes that are the same in all of these areas, and in the everyday business of evaluation, but it also includes some features that are specific to sub-sets or to individual members of the group. Many of the standard sub-divisions have had specialists working in them for a lifetime and have of course evolved sophisticated technical vocabularies and models of their own. Of course, some of this specialized knowledge will usually have to be mastered in order to make a contribution to the sub-area, although the amount varies considerably by area.

Perhaps the limit case of specialization is in the category of intradisciplinary evaluation, where it is generally thought that to be good at evaluating theories in particle physics, for example, or explanations in medieval history, or programs in international aid, you need to be a fully qualified specialist in the relevant area. However, this turns out not to be entirely correct; the study of the logic of evaluation does yield some results that include refutations of, or substantial improvements upon, the practice of intradisciplinary explanation; and studies of the logic of causal explanation in educational research have uncovered flaws and finer points missed by the mainstream professionals. Two good examples of this can be found in areas of current high activity—the general evaluation of research,[4] and the evaluation of causation in international aid.[5] So the logic of evaluation is sometimes not only of interest in itself, but can be of some value to the general scholarly enterprise and to all the social efforts that depend on results from that enterprise, e.g., the social enterprise of poverty reduction.

From the comments already made about the geography of evaluation, it should be clear that evaluation has a rather unusual dual role in the country of the mind. It is of course a discipline, or family of disciplines, in itself—in fact, even within one subdivision such as program or personnel evaluation scores of entirely specific books, journal issues, and conferences mark every year. But it is also a “service discipline” in the same way as logic or—to a more limited extent—statistics, because every discipline is heavily involved in evaluation. This goes somewhat beyond the way in which each discipline uses intradisciplinary evaluation, which mainly involves the evaluation of methodological entities, since every discipline evaluates its own students as well as practitioners (personnel evaluation), research proposals submitted for funding (proposal evaluation), its journals, websites, doctoral programs, and associations (product and program evaluation), etc. For this reason, evaluation is sometimes referred to as a transdiscipline—one that operates across

[4] Documented in the Summer 2008 issue of *New Directions in Evaluation* (Vol. 2008, Issue 118, pp. 1-105) eds. C. Coryn and M. Scriven; the entire issue is devoted to the evaluation of research.)

[5] Will be supported in the discussions with the Ministry of Foreign Affairs at Tokyo (July, 2007), and in an article in #8 of the online *Journal of MultiDisciplinary Evaluation*. Earlier issues of this journal contain earlier discussions of this topic.

many other disciplines as well as autonomously.[6] Logic is the only other transdiscipline that operates across all academic disciplines, but evaluation also operates across all physical disciplines, e.g., tai chi, marathon running.[7] But is it based on a logical error?

3. The possibility of evaluation

The Humean argument that is a precursor to the naturalistic fallacy was appealing enough to rationalize the ban on evaluative research (the "value-free doctrine") that largely prevented the social sciences from directly addressing most issues of social concern for the greater part of the 20th century. Following the conventional interpretation of Hume, it was argued that since one cannot validly infer evaluative conclusions from "factual" (here used, incorrectly, to mean naturalistic, a.k.a. empirical) premises, it followed that science cannot support them. This is an invalid argument with a false premise into the bargain. The premise is false, following Searle, since from the observable fact that Jones promised to repay his debt to Smith, it follows that he should do so—an evaluative conclusion— using the tautological (definitionally true) inference warrant that promises incur obligations. The argument is invalid, since science includes far more than factual (i.e., non-evaluative) claims: it is, by practical necessity and of its very nature as a discipline, suffused with evaluative claims, without which it would not be either functional or distinguishable from charlatanry. These claims range from evaluation of the quality of observational data and instruments, and the merits of experimental designs, hypotheses and theories, to evaluations of the worth of research proposals and scientific papers submitted to journals for publication, and to evaluations of the merit of scientific works and scientists. The difference between astrology and astronomy is essentially a difference in the quality of their data and the inferences from it. In fact, the nature of any discipline, including not just the sciences but also history and jurisprudence—the very reason they deserve the name at all—lies in the extent to which they meet the standards of quality that learning to practice requires every practitioner to master. Since science itself contains, at the heart of its practice and conclusions, many evaluative claims, the assumption that it is only concerned with

[6] Note the difference from 'interdisciplinary' meaning "relating to more than one discipline", and from 'multidisciplinary,' implying relevance to more than two disciplines.

[7] However, if one thinks that all 'logic of….' subjects are part of logic, logic would include the logic of evaluation, and hence, in this extended sense, the rules governing evaluation in the physical disciplines.

claims that can be established as (non-evaluative) facts directly, or by inference from other facts, is wrong.

Against this commonsensical refutation of the value-free doctrine, two defenses are often produced. It is sometimes argued that the doctrine was only meant to exclude the intrusion of *political and ethical* values into science, not these workaday values. However, political and ethical values are excludable simply on the grounds of irrelevance—when they are indeed irrelevant—so there was no need for a new doctrine to keep them out. And a glance at the arguments provided for the doctrine shows that they are arguments that are entirely general about value claims, not restricted to a sub-species.

Again, it is sometimes argued that the kind of evaluative claim that is commonplace in science is acceptable "because it can be translated (loosely speaking) into factual claims": but, of course, it is plausibly arguable that this is also true, loosely speaking, about most evaluative claims—and equally, of course, it was exactly this claim about translatability that G. E. Moore rejected, and labeled the naturalistic fallacy.

Here we come to a number of further variations in the general case against evaluative claims as scientifically (or logically) respectable. To a substantial extent the case was based on three confluent mistakes—one error of misclassification, one of overgeneralization, and one about definability and translatability—and each of these comes into one or another of several defenses of the value-free position or its grounds. The first error was the supposition that the typical value claim was a statement of personal taste or preference, which of course lacks the quality of universality that was thought to be requisite for objectivity—here taken to mean universality in the sense of intersubjectivity—and hence scientific status. The error here was to over-generalize: while personal preferences are indeed one kind of value, they are not the only kind. For example, "Murder is wrong," "The reliability of cars is a prima facie-valid criterion of their merit," and "The average vehicle on a United States highway still contributes more than twice as much in emissions as the average Chinese vehicle" are all clearly evaluative but strongly general, two true by definition and one a matter of evaluative fact (albeit timebound and with particular names as descriptors).

The second error, closely related, was the supposition that objectivity requires anything approaching universal generalization, clearly an error since large parts of science (as well as history and jurisprudence, where objectivity is equally important) are both objective and highly particular, obvious scientific examples being geology, forensic science, and planetary astronomy.

The third error, Moore's error, was the assumption that definability/translatability is a matter of logically necessary connections. In fact, most common concepts—especially the most important and general ones—in science as well as in common experience, are “cluster concepts” that are learnt, explained, and defined in terms of a cluster of properties, few if any of which are *necessarily* connected to the concept. These properties are more commonly just criteria for the concepts, correctly said to be part of their meaning, but not necessarily connected to it. In some scientific cases, temperature being a well-known example, the concept at one point may have been “defined” in a certain way and thus briefly appear to be fully specified in terms of a set of necessary and sufficient conditions, but subsequent events clearly demonstrate that its full meaning extended far beyond the supposed definition, which consequently had to be revised— several times. Thus, coming to understand the significance of Wittgenstein’s language games overpowered Moore’s attempted proof of fallacy. Hence, ‘good’ might be naturalistically definable even though not logically equivalent to any determinate (finite[8]) set of naturalistic properties.

And there is a more complex point involved. The main reason that there is no naturalistic equivalent of ‘good’ is not because of the fallacy of thinking that Moore’s hypothetical equation was the condition for meaning equivalence, but because the term is learnt only in context: we learn what makes X good, for a thousand Xs, not by learning what good is and combining that with our understanding of what X is. Good, by itself, is a function word, not an abbreviation for a set of properties; it simply reminds us of that second set of properties for each X, the set that we have learnt increases its merit or quality or value or worth, A good apple, for example, is one that lacks worms and bruises—but worms and bruises are not part of the definition of apple. (However, it’s true that we tend to learn at least a little about what makes a good X along with learning what makes an X, so the notions are often genetically connected, at a simple level).

This logical point is of great importance in the practice of evaluation, since the first step in many evaluation projects is to unpack the meritorious variant of the concept of which the subject of the evaluation[9] is an instance—a health clinic, a curriculum innovation, a drunk-driving policy—into its constituent criteria, which can then be tied to empirical indicators for field investigation. And in doing this, we often need highly specialized knowledge, because knowing what makes a good clinic, for example, requires knowing what equipment it needs, and you don’t learn much about that when you learn what a clinic is. That’s partly because the needed equipment changes as medicine changes, and part of the utility of ‘good’ is that it’s a pointer term that points us towards getting these updates.

[8] A finite list seems to be the relevant consideration, since Moore’s argument loses plausibility if the list is not finite.

[9] The term for whatever is being evaluated is ‘evaluand.’

Hence, it's perfectly possible for 'good' to refer only to naturalistic properties even though no finite list of them is possible in the abstract because it's a different list for each evaluand and the list of evaluands is indefinitely extensible, hence not determinate.

However, since many criteria of merit are not themselves naturalistic—for example, the criteria of merit for arguments—the project is impossible for other reasons. The bottom line is that 'good' is a concept that is analyzable into everyday functional concepts, not a referent to some transcendental property.[10] So the naturalistic fallacy provides, for more than one reason, no argument against the possibility of rational evaluation. Note that many people were and are influenced into abandoning or even ridiculing the value-free doctrine for a completely different reason, based on a total misunderstanding of the doctrine. It was argued, notably by political radicals of the 1960s and 1970s, who were rightly incensed by the discovery that social scientists had been providing technical help to dictatorships in South America, that the idea of value-free science was absurd because scientists were obviously affected by their own values in their choice of careers, and their choice of applications of their work. But of course no one arguing for the value-free doctrine had ever suggested the contrary; the doctrine is about keeping the inner workings of science—its inferences, its data, etc.—free from value judgments. Having motives for doing or applying science is obviously universal, and would not be denied by anyone with a college degree; the idea that science could not justify evaluative conclusions was entirely different and supported by arguments that looked good to good scientists.

The argument against objectivity is even less formidable. It is normally based on the claim that no one is without bias, often supported by the previous argument that scientists often have evident political biases. However, that premise does not imply that no group, e.g., a group of scientists or historians in a particular field of expertise, including those with a wide range of political and ethical preferences, is incapable of generating claims that we can have good reason to believe to be objective. The counterargument, which is an induction from the historically common discovery of bias even in such bodies of claims, does no more to support skepticism than the analogous argument supports it about the truth of similarly supported claims such as the claim that vaccination against smallpox works very well. Of course, the skeptic's argument does support constant vigilance and avoidance of overconfidence, but these are already ensconced as primary imperatives in the scientific attitude. We are often wrong, even when we had been certain, but 'often', although it

[10] Or, if you like this phrasing better, it refers to a non-naturalistic concept that is epistemologically on a par with the concept of a dimension, i.e., a property that is mildly abstracted from a set of naturalistic properties (length, width, etc.).

refers to a numerically large number of cases, refers only to a minute fraction of the cases where due diligence supports a firm conclusion. So this argument is simply an illicit exaggeration of a bad statistical inference; or, if you prefer, a misunderstanding of the meaning of claims of truth or objectivity.

Which brings us to the topic-specific logic of evaluation, that is, to the logic underlying practical reasoning in the field. It involves two levels of generality, corresponding roughly to the levels distinguished in ethics as normative ethics and metaethics. We'll begin with the more abstract level, which is the effort to convey the general nature of (professional) evaluation. Historically, this has mostly meant program evaluation, since that is the sub-area in which most of the theorizing has been done, but we'll look at the answers (a few of them) in terms of their accuracy for all areas.

4. Conceptions and misconceptions of evaluation

In the development of the various branches of evaluation, most of them intensely practical in their aims, very little crosspollination has occurred. This is partly due to the bizarre episode in the history of thought we have already discussed, where evaluation was tabooed by the mistaken ideal of "value-free social science". This ban on the legitimacy of any kind of evaluation of course deterred any efforts to develop a general theory if taken seriously in areas like medicine and engineering, where many of the workaday tasks were largely evaluative—no-one supposed that the doctor should not make recommendations about your diet or medications, even though they were of course evaluative. And it was not taken seriously in, for example, personnel and performance evaluation, where the name of the game was to a large extent to provide better ways to select the best applicants or candidates for promotion or discharge. But it was taken seriously enough in mainstream sociology, where it originated with Max Weber, in political science, in much of social psychology, as well as in many departments of psychology, to scare almost everyone away from serious treatment of serious social problems, which of course is heavily dependent on the evaluation of the proposed solutions. This made them largely useless where needed most; but since the practical demand was great, the opponents of evaluation did propose what amounted to a behaviorist surrogate for serious evaluation, one that could be found in almost every text for many decades in mid-twentieth century social science. The grip of this surrogate was so strong that even today one frequently hears or sees it proposed online as the right way to go about (social) program evaluation. The suggestion was simple enough, and goes like this: (i) identify the goals of the program; (ii) convert them to behavioral objectives (or as near as you can get to that); (iii) find or create tests for measuring the behaviors specified in those objectives; (iv) test the program participants to see if they have achieved these objectives; (v) analyze the results to see if they have, in which case the program is successful; if not, it has failed.

The absurdity of this suggestion would have been obvious if anyone had been thinking about the general logic of evaluation, because everyone does product evaluation all the time and everyone knows that you don't evaluate products against the intents of their makers but against the needs of their users or prospective users (see any issue of *Consumer Reports*). But the surviving evaluators were hiding out in their own silos, so the absurdity was not apparent for a long time. There are half a dozen flaws in the "goalachievement model" of evaluation. Getting to the goals may only show they were set too low, not that the program was worthwhile. Not getting to them may only show they were set too high, not that the program wasn't marvelous. The goals may be immoral, or they may be irrelevant to the needs of those served. Getting to goals that are perfectly matched to needs may still represent a trivial achievement; or it may be important, but overshadowed by side-effects, whose existence is not mentioned in this model for doing evaluation. Doing wonderful things is boring from the practical point of view if it costs far too much; or if the costs were moderate, if it could have been done another way for less. Doing even a little good may be worthwhile if it cost very little. And so on, and on.[11]

Goals are for guiding action; for planning and managing. They are not the key issue for evaluators—as ethicists have long noted in their territory, "the road to hell is paved with good intentions"—although they are something that must be commented on at some point in an evaluation. The key issue for program evaluators is the effects the program actually has, measured in terms of what they mean to those you affected (and those you did not reach), whether or not you meant to have those effects; plus how much it costs to do what you did do; and what alternatives there were; and how you got to where you got (since the end does not justify all means); and other things, e.g., the potential uses of the program by contrast with the uses under examination.

The introduction of "goal-free evaluation" and the demonstration that it was feasible and in significant ways superior to "goal-attainment evaluation" finally established—but only in the community of professional evaluators who keep up with the subject's development—that the goal-based approach was just a reflection of the fact that program evaluation mainly began as an effort funded by program managers and planners, whose interest was of

[11] See "Hard-Won Lessons in Program Evaluation" by the present author in *New Directions in Program Evaluation*, Summer 1993, (Jossey-Bass), whole issue.

course mainly in whether their goals had been met.[12] Ethically, and to an increasing extent politically, the consumer's point of view must be taken as primary.

However, the main threat to establishing evaluation as a respectable subject, or profession, or discipline, was much more fundamental: it was the claim that for basic logical reasons, there could never be a scientific or indeed a rational basis for evaluative conclusions, and we've already looked at those arguments.

Once program evaluation got past the barrier of the value-free doctrine, following personnel and performance evaluation (the latter being the category into which all student testing falls), and the cartoonlike effort of the goal-achievement model, it began to generate many justificatory accounts of the real nature of evaluation. These began with the relatively simple checklist models like the CIPP (Context, Input, Process, Product) model and went on in three directions. One of them led to more sophisticated checklists, e.g., the still-primitive DAC model used widely in international aid evaluation, and the Key Evaluation Checklist from me (on line). Another led to the formulation of national standards for program evaluation—a formidable and impressive project. The third was less structured, and led to a long and still-expanding list of original and interesting "models" ranging from the goal-free approach, through the "advocate-adversary" model based on the legal metaphor, to the more recent participatory, collaborative, and empowerment models (involving increasing amounts of involvement of the evaluees as partners or authors in evaluating themselves or their programs), and the "appreciative inquiry" approach (which strives to emphasize a more constructive approach than the common critical one seen as leading to much of the negative reaction to the use of evaluation in many organizations). Many articles and several books have been devoted to taxonomies and critiques of these models, so I will omit almost all of them, but recommend those discussions for those interested in how a new discipline seeks to rationalize and reform its own processes.[13]

[12] See *"Prose and Cons About Goal-Free Evaluation", in *Evaluation Comment*, December, 1972, Center for the Study of Evaluation, University of California in Los Angeles. Reprinted in *Evaluation Studies Review Annual*, Vol. 1, Gene Glass, ed., Sage Publications, August, 1976; reprinted in Hamilton, D., Jenkins, D., King, C., MacDonald, B., and Parlett, Michael., eds., *Beyond the Numbers Game: A Reader in Educational Evaluation*, Macmillan, 1977, pp. 130–171.

[13] A recent and one of the best of these is: *Evaluation Theory, Models, and Applications* by Daniel L. Stufflebeam and Anthony J. Shinkfield, Jossey-Bass, April, 2007, and Dan's 2018 update of the CIPP model.

We have already covered two such models (three if one counts the transdisciplinary account as a model) and it is worth looking at what is probably the most popular common model today, since it is based on a misconception that may be as serious as the goal-based error. According to this model, referred to in the United States as the logic model approach, and in the United Kingdom often called the realist or realistic approach, evaluations should begin with and to some extent (or a large extent, or completely, depending on the author) be based on an account of how the program works (or is supposed to work, according to other authors). To cut to the bottom line of a refutation, this looks as if it's based on a simple confusion of the aim of evaluation with the aim of explanation. The strong counter-examples are from the evaluation of medicine, where we were evaluating aspirin for fifty years without the slightest idea of how it worked; or from the evaluation of claimed faith-healing, where the supposed explanation is supernatural intervention, but the evaluation can proceed entirely without commitment for or against that view. However, these counter-examples have not proved persuasive so far, and it is possible that this is because there are more subtle benefits to be reaped from the logic model approach. It goes without saying that it would be a notable benefit if the evaluator could in fact provide explanations of whatever it is that s/he finds—success or failure.

But that's absurdly ambitious since in many cases where an evaluation is called for, e.g., in investigating a new approach to reducing melanoma or juvenile delinquency, the best experts in the field have been unable to determine an acceptable explanation. And explanation is at best a side-effect of the main goal that, in this case, should dominate the investigation, which is to determine whether success or failure occurred (more accurately, to determine the kind and extent of the benefits or damage wrought). Moreover, it is not true that having an explanation of X's effects makes a substantial contribution to the evaluation of X in a particular context, since we may know how X works but not know whether its effects were net positive or negative, or even how large they are, in this context or even in general. So the explanation is not only not necessary for evaluation, but not sufficient for—not even substantially contributory to—evaluation.

The potential costs of this approach are considerable. They run from simply wasting some resources on explanation-hunting to spending so much time and/or other resources on this hunt that one fails to do the evaluation task at all. It's hard to avoid the conclusion that the evaluators favoring this approach are driven by the urge to make a significant contribution to explanatory science, perhaps because many of them were originally trained in a scientific field, and a lack of understanding of the difference between the two enterprises.

There are other interesting efforts at conceptualizing evaluation, of which perhaps the most interesting ones now popular refer to the spectrum encouraging some degree of participation of the staff of the program or design team responsible for a program or product. The benefits of this approach are access to data and insights not always made available to the external evaluator, increased “buy-in” and hence better implementation of findings, and reduced anxiety/opposition to the evaluative effort. Costs of this approach depend on the degree of shared responsibility and co-vary with that, along the dimensions of credibility, utility, validity, and ethicality (the latter comes in since the targeted consumers, and other impacted groups, are still not involved in the usual versions of these models).

But now it’s time to move on to the more specific level of evaluation: procedural logic.

5. The basic logic of evaluation

5.1 Types of value claim

It is important to distinguish half a dozen of these types since each of them requires somewhat different data to support it, and each legitimates somewhat different conclusions. We begin with: *(i) Personal preferences (wants).* Often offered as the paradigm of value claims, in order to denigrate them, these are in fact the least important from the point of view of professional evaluation. Nevertheless, it should be absolutely clear that, although paradigmatically subjective in the purely descriptive sense, these claims are entirely objective in the evaluative sense that they make an intersubjectively testable and hence scientifically respectable claim. The evidence we all use to verify or falsify them is of two main kinds: evidence bearing on the veracity of the claimant, and behavioral evidence of the obvious kind. If, for example, someone claims to prefer classic literature to contemporary fiction, we expect to see their library and their movie-going reflect that taste, of course as a norm from which not only exceptions but general deviations can be justified without refuting their claim—and we know the difference between a justification and a cover-up for a lie. This type of claim is extremely important in the logic of evaluation because it can provide the main value premise required for support of evaluative conclusions with exactly the same circle of applicability as the premise, namely the person or people said to have the preferences. In order to give people advice about, for example, the best car for them to buy, or the best investments for them to make, expert advisers need input about their preferences. Now, when groups of people share certain preferences, we can often move to an even more objective form of value premise, namely…

(ii) Market value. This kind of value has a specific legal status and definition[14] that references the verification procedure, so it is not only objective but quantitative.

But the market value is not the most important value; in many cases, an expert adviser will say, “Well, there’s no doubt that the market value of this property is X, but it’s seriously overpriced—we can find you a much better value, in fact a better property for less money.” That is, there’s something that we refer to as the “real” value that is sometimes, not always, different from market value. This has other names, e.g., …

(iii) Real, true, essential value. This is a theoretical construct based on abstracting from the properly established criteria of merit for an evaluand. When we talk of a change as “really significant” or a doctoral thesis as “truly excellent,” we are stressing that a truly careful evaluation will reveal the result we are claiming. This sense of value is the one that professional evaluation seeks to uncover, and it is the one that evaluation as a discipline is all about, just as the “real truth” is what the professional journalist or scientist seeks. To reach it, we usually have to assemble a number of values and a number of facts, and integrate (synthesize) them in a way that calls on a couple more types of value. The first of these is:

(iv) Public value. Public values are the kind to which we appeal when we are trying to establish the m/w/s of some evaluand (a program, product, or person) in a way that will be intersubjectively acceptable, as for example when it is to be paid for with public money, or is intended to receive public acknowledgment for its m/w/s.

There are many species of these, including legal, ethical, cultural, logical, scientific, educational, historical, professional, and expert values, and the context determines which one(s) it will be relevant to consider. (The name for these is intended to suggest the contrast with private or misleading values, e.g., personal preferences.) However, public values are general in nature, and in order to pin down an evaluative conclusion of the required specificity, we often—although not always—need a refinement of them, as follows:

(v) Standards and requirements. Standards are specific levels or amounts of public values that are set for certain evaluative or practical purposes. Some public values are so essentially tied to action that they are themselves referred to and formulated as standards—a good example is safety standards, which are normally expressed in exact terms, e.g., “A fire extinguisher of type X must be located and visible in standard working conditions within five feet of

[14] In the real estate market, one definition is: The most probable price (in terms of money) which a property should bring in a competitive and open market under all conditions requisite to a fair sale, the buyer and seller each acting prudently and knowledgeably, and assuming the price is not affected by undue stimulus.

each machine operator's normal location." But in many cases, the standards will be context dependent, and stated for specific evaluative tasks or contexts, for example: "An A on this test will require a mark of 85%." (The formulation of a set of standards like this one, covering all grades, is called a rubric.)

The preceding five categories are the main types of value, but there are a couple more that should be mentioned to avoid confusion.

(vi) Contextual values. There are many contexts in which a statement of bare empirical fact has evaluative significance. (This is one reason why the facts vs. values dichotomy is misleading.) It is a "mere" statement of fact that someone has just broken the world record for the hundred-meter dash, but of course it's also a statement about the m/w/s of the event and the sprinter. Hence, in an argument, it may be entirely appropriate to address the evaluative component of the statement rather than the factual one, e.g., by pointing out that the track was an exceptionally favorable one and the new record was only a hundredth of a second better than the previous one, set on an ancient track.

(vii) Illustrative and exemplary value. When someone talks of the "perfect crime" or the "perfect storm" they are not suggesting that the crime or the storm are ideal events in themselves, but only that they are ideal examples of their kind; Aristotle discussed the case of the "good thief" in the same way.

Now we need to look at the basic evaluative operations, and distinguish them.

We are all loosely familiar with them, but the exact definitions of the terms when used precisely are rarely quoted correctly, and once again, they need to be carefully distinguished because the data and analysis required to answer an evaluative request for any one of them is essentially different from that required for any other. (Note that the same names as below are used for analogous non-evaluative operations, and sometimes it's not clear which is intended.)

5.2 Types of evaluative operation

The following material may seem to duplicate what is found in educational measurement texts, but a careful reading will reveal that the focus here on evaluation leads to significant differences. Educational measurement is still taught as closely analogous to measurement in the physical sciences, but it is crucially different in that educational measurement is usually supposed to be measuring something of value, and that task requires specific and more complex consideration. We distinguish four basic operations, and add two refinements.

(i) Ranking or ordering (the term 'ordering' is more commonly used for non-evaluative tasks such as placing in order of height). Placing evaluands in order of m/w/s: strict ranking allows no ties (evaluands with the same rank), while weak ranking allows some ties, sometimes allowing only a small number of them (this number should be specified), and more commonly allowing only a small number of evaluands at the same rank (this number should also be specified). An evaluand's rank has no meaning except for its relative ordering against the other ranked evaluands. Ranking is the key operation for evaluative tasks like choosing which alternative to adopt in decision-making, which product to buy, who it to get prizes or scholarships or assistance. Note that the top-ranked evaluand in a set may have negligible merit, and the lowest-ranked may be superb (not in the same set, however). Note also that a ranking may be partial, i.e., the best or worst sub-set may be ranked and the others simply listed as "of lower/higher rank"; they are not of course tied. (This is not the same as an incomplete grading, where the remainder is of unknown rank, and therefore the ranks of those ranked so far are only tentative; contrast this with incomplete grading (below). Note also that ranking is automatically exhaustive, cf. grading.

Gap-ranking. This is a useful variant of ranking where an indication of the size of the interval between evaluands is given in addition to their rank order. This is common in horse-racing ("X is the winner by three lengths, Y next by a head from Z, who is third by a nose", etc.). Gap estimates can often be provided with considerable practical benefit in other cases, telling the buyer for example, that taking the runner-up at a much lower price would be a good buy, or that in an election where our candidate loses, that it was "very close". When scores are involved, we can often give precise quantitative measures of the intervals.

(ii) Grading. This is the operation of placing evaluands in one of a set of categories (the grades) which are defined in three ways: (a) each evaluand in a particular grade is superior in terms of X (the evaluative property within the large group of properties indicating m/s/w that is being used as the basis for grading) to all in a lower grade, and inferior to all in higher grades; (b) each grade is named with a term that has some degree of independent meaning in the vocabulary of m/s/w, a meaning that must of course be consistent with condition (a); (c) the set of grades is exhaustive, i.e., all evaluands are included in one or another. Of course, there can be unlimited ties in a graded group, by contrast with (all strict and most weak) rankings. The most familiar example of a set of grades is perhaps the set A through F (in the U.S.) A though E (in the U.K.), or the roughly equivalent sets such as Excellent / Very good / Good / Marginally acceptable / Unacceptable). Note that condition (b) is violated by the so-called "grading on the curve" procedure, which is log-

ically self-contradictory and ethically abhorrent (since those “failed” by the curve’s definition may in fact have done acceptable work)[15]. Note also that a set of grades must be exhaustive, i.e., so defined as to cover all probable cases.

Note also that a partial and incomplete grading should not require any reconsideration when it is completed, since grades are determinate predicates, all of whose relevant conditions are considered when the grade is awarded, unlike ranking.

The essential difference between ranking and grading is that neither can do the job of the other. Each adds something to the story the other tells: the grade adds the “absolute reference” to a set of external standards; the rank adds the inter-evaluand comparison element absent in grading (if the ranking is not too weak). Each provides answers to questions the other cannot answer: ranking answers the quest for the best and the better; grading answers the quest for true quality, good enough quality, etc.

Profiling. A variant of grading, with great evaluative utility. A profile can be constructed of either an evaluand, in which case it shows strengths and weaknesses, a good way to express the results of an evaluation in a graphically direct way (or a simple table); or it may be the profile of a consumer or market or community, where it will show the relative importance of their wants or needs. The graphical version is a vertical bar graph of an evaluand, one on which each bar represents one of the criteria of merit. (More details in the following section.)

Part of the importance of a profile is that it avoids the need, when evaluating something, to convert the result into a single overall scale, in order to give a single grade to the evaluand. To do that, one must determine the rela-

[15] Grading on the curve was and is an exa*m*ple of the extremes to which the value-free adherents (or quantitative enthusiasts) were prepared to go in order to avoid actually making a judgment of quality. It is often defended as an acceptable way of avoiding the errors of “subjective” judgments of quality in large classes, but of course its validity depends on three weak assumptions—that classes do not vary significantly from time to time, and that the test is of unvarying difficulty, and impregnable security. Granted there are substantial interinstructor variations in the definitions of grade quality amongst random samples of faculty, it has long been known how to reduce them to minor dimensions, so continuing to use grading on the curve is a confession of inadequate training of faculty. And ultimately it was a futile gesture, since it depended on scoring, which depends on the fundamental judgment of quality in the allocation of every single point on the scale, e.g., in the judgment that a correct answer to the nth question on the test is worth m points. When the test is multiple-choice and each question is allocated one point, it’s obvious that this judgment is false for almost every test, so this is hardly a valid foundation on which to build a test grade.

tive importance of the dimensions of merit, a requirement that is sometimes extremely difficult to justify. While a profile thus provides a multi-dimensional portrayal of a multi-dimensional reality, it isn't good enough if what the evaluative question requires is a ranking, a single winner for the best car to buy or the best proposal to fund or the valedictorian. Then the evaluator must do the reduction to a single scale, usually by appeal to the profile of the relevant consumers/users/voters.

(iii) Scoring. The application of a m/w/s measurement scale to the evaluand(s).

This is the quantitative approach to evaluation, sometimes the best approach, often impossible, quite often absurd. The best case is exemplified in the evaluation of athletic performances where the stopwatch is the measuring instrument, and can be used, in conjunction with suitable rubrics and databases,[16] to yield both ranking and grading. The absurdity shows up in the faculty handbooks used in some major universities where the faculty are instructed to award grades based on scores in some specific way, e.g., "The grade of A is to be given if and only if the student's mark on a course is between 85% and 100%." But as every faculty member in the humanities knows very well, this is an absurd suggestion since essay tests are the norm there, and it is much more appropriate— and is the standard practice, regardless of what the handbook says—to give the answers a grade rather than a score and calculate or estimate a weighted average of the grades in order to get the test grade (which are in turn cumulated to get the course grade). It is also invalid where it is normally used, in science and math courses, since it rests on the assumption that all items (and other scored assignments) are of equal difficulty, a highly unrealistic assumption. Getting 86% on a test may be a sign of a middle B, even a C, or an A , depending on the test's difficulty and the response's originality, presentation, etc. as everyone knows. But some administrator with a quantitative obsession wrote the handbook, so—if the faculty (or students, or alums) knew anything about the logic of evaluation—this regulation would make the institution look absurd. But, since it's often been there for half a century, one should probably conclude that the conditional does not apply.

(iv) Apportioning. This is the fourth and last of the basic evaluative operation we'll distinguish. It involves the distribution or allocation of valuable resources to a set of evaluands in a way that reflects their relative m/w/s. A good example is the distribution by a department head between the members

[16] And, in serious athletic competition, a recently calibrated altimeter, since the international athletic associations have determined the extent to which altitude must be used to discount athletic performance that is to be considered for world or country record status.

of the department of a sum of money, S, allocated to the department for merit raises. If one can find a valid measure of their relative merit on a scale of, say, 1 to 10, possibly the most plausible way to do this would be to add up the total of all their scores on this scale—let's call the result M—and allocate the nth member of the department, whose own score is N, N/M x S. This is a simple case of the general problem of portfolio construction. In the more difficult case of constructing a solution when your investment is to be into a number of stocks or bonds or other investments, each of which has different profiles of probable performance in different respects (safety, short-term income, long-term growth, etc.) and given your own profile of values for each of these aspects of performance, there is no publicly known algorithm of any real plausibility, although investment managers at large funds undoubtedly have their own strategies based on massive trials of computer models. So the portfolio evaluator has no ideal against which to compare actual cases.

However, there are a number of ways not to do it, and one can at least learn to avoid them. A simple example of these, on a related problem, is this: if the total organization budget is cut from S to 90% of S in a later year, a common error is to think that an across the board cut of 10% in the submitted department budgets would be an appropriate, perhaps even the most "fair" response. It is a bad response because it simply rewards those who have padded their budgets the most, and will encourage that response in future years from smart department heads. The correct first step in a response to the budget cut problem is to trim all budgets to the same standards of economical use of resources, albeit imprecisely.

6. Refining the evaluative tools

The vertical scales of a profile can just show units of measurement for each dimension (criterion) but more usefully, after conversion using the standards (rubrics) for each dimension (these are available since it's usually a value), it shows their grades on a single common scale, the vertical axis. We may take these to be A through F, for discussion purposes; here can add two or three useful refinements to this representation.

First, we can mark what are called "bars" on these scales, whether in their raw or evaluative form. The most important bar is a horizontal line showing any absolute minimum value (on the raw scale) or grade (on the converted scale) that must be attained.

For example, if we are profiling the qualifications of a candidate for a staff position where fluency in Spanish is absolutely necessary, we put a bar under B, or A (for fluency equivalent to a native speaker) across the scale referring to Spanish competency. Of course only some scales have bars, and in a particular case all or none might be barred.

Bars must be very carefully applied, since failing to clear the bar on any scale eliminates the candidate from any further consideration, no matter how well they score on other dimensions. (For this reason, dimensions without bars are sometimes called “compensatory”.) Because of the power of bars, when evaluating a new candidate it is usually good practice to start by checking the barred dimensions (called “autonomous”), since it will save time (and favors) to find failings there first.

In addition to bars on particular scales, there may be a holistic bar. For example, in admitting a candidate to graduate school in a doctoral program, one may require certain minimum scores on the GRE math scale, a masters degree from a certified college, and a 3.5 grade point average across their courses in that degree program; but one may also be unwilling to admit a candidate who just scrapes over all the bars. This bar might translate as requiring a B average on all dimensions considered, and can be shown on the evaluative profile as a dotted horizontal line at that grade level across the whole chart, meaning that average is required.

For some (rare) purposes, it may be useful to also use a “high bar” on some scale meaning one that if cleared, implies that the evaluand is to be admitted, appointed, or purchased, regardless of their score on all other scales. This has been called the “Einstein bar” meaning that if Einstein applied for a position at your university in any year after 1905 (the “anna mirabilis” when he published several papers each of which is often said to have been deserving of a Nobel prize in itself) he would be judged as clearing the high bar on the scale of research competency and thereby bypass any consideration of his performance on the other two bars on which faculty candidates are assessed (teaching and service).

Second, one can also show differential weights on a profile, representing the differential importance of various dimensions if this is something that needs to be represented. For example, an undergraduate college might consider teaching to be twice as important as research or service, and show this by doubling the width of the teaching column. Using this device, the best candidate then shows up as the one with the largest area under their profile (the line through the points representing their achievement on each dimension) rather than the one with the highest average score (as long as the profile does not dip below the minimum bar on any scale). It should be noted however, that validating differential weights is extremely difficult (it requires deciding whether teaching should be twice as important as research, 50% more important, three times as important, etc.).

In evaluating programs, the five key bars—analogous to the three used in evaluating faculty—used in the Key Evaluation Checklist approach are process, outcomes, cost, comparisons, and generalizability, and the first four have minimum bars on them. Of course, the details turn out to be where the professional skills come in, and a minimalist expansion of these runs 12,000 words at the Checklists in Evaluation website, i.e., 50% more than this paper (under Tools and Resources at evaluation.wmich.edu).[17]

By giving details on the profiling refinements, we have in fact done a good deal of work on one of the great methodological problems unique to evaluation, the synthesis problem—how to combine the "factual premises" with the value premises.

7. Conclusion

It is perhaps best to conclude by mentioning something interesting still on the table, so here is one example of material not approached here but falling under the article's title.

There is currently a major war going on in evaluation circles over the single question of how outcomes are to be determined in program evaluation, that is, it concerns only one sub-topic of one of the five main criteria for one subdivision of evaluation. This dispute has brought in the entire literature on causation, and the outcome of it will determine the use of at least five billion dollars a year of direct expenditure.[18] So this article is only an introduction to its subject, but it may be enough to suggest to people seriously interested in the domains of critical thinking/informal logic, and the philosophy of the social sciences, the serious possibility that evaluation now constitutes an area in which some mapping has been done, some value has resulted, and important work remains unfinished.

The title of this essay refers to the "unruly" logic of evaluation. Why not just call it "mixed methods"? Because none of the methods in the mix is tidy; and because there is something else involved that's quite different from what's involved in the mixed methods as currently conceived. That "something more" is the method of linguistic microanalysis that is buried in the Wittgensteinian approach, as refined by the Oxbridge efforts of the 1950s and thereafter. To formulate and validate an account of this approach and its link to the process of reasoning to evaluative conclusions, I believe one

[17] There are a dozen texts on the market that expand on evaluation at greater length, of which the one by E. Jane Davidson (*Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*) sticks closest to a serious logic of evaluation approach and is not exclusively focused on program evaluation.

[18] Some further discussion of this topic will be found in the reference in note 5 above.

has to reanalyze the concepts of definition and translation and thence reconsider abductive and probative inference—tasks for another day. Or two. The results, if successful, might be called the "paralogic" of evaluation. Till then, "unruly".

About the author:
Michael Scriven was born in the United Kingdom, went to Australia from age 12 through a BA & MA at Melbourne in honors math. Returning to Oxford, he received a D.Phil. in philosophy (thesis on explanation theory). He then held appointments (about four years each) at Minnesota, Swarthmore and Indiana; did summer teaching at Wesleyan and Harvard; did research in primatology in Austria; and two Institutions for Advance Study (on democracy and on the behavioral sciences). Next he spent 12 years at Berkeley, with eventually a split appointment between philosophy and education. He left UCB to start an institution for evaluation research at the University of San Francisco; ran an evaluation consulting business; was a Professor of Psychology at Claremont Graduate University; served was co-head of the School of Education at the University of Western Australia for twelve years; now back at Claremont as co-director of the Claremont Evaluation Center.