Evaluation Revolutions

Journal of MultiDisciplinary Evaluation
Volume 11, Issue 25, 2015

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Michael Scriven
Claremont Graduate University

The everyday practice of evaluation has continued for millennia, only recently sprouting an academic branch that became more sophisticated and transformed into an important discipline. More precisely, it developed into a family of sub-disciplines—product evaluation, program evaluation, personnel evaluation, policy analysis, etc. At the metalevel, the perception of it in most of the academic world has undergone some highly significant shifts, separated by what can fairly be described as revolutions. That sequence is what I try to describe below, starting with a description of the pre-revolutionary baseline state. The list extends from the past, through the present, and into the future, the latter including some suggestions about revolutions I think we need to kickstart. Most evaluators will continue to spend most of their time on applied evaluation in some specialty field, but I'm hoping they will help out part-time with the revolutions, at least by thinking and arguing about them. Better still if they are intrigued and challenged by the suggestions for future revolutions, and add their own experience to the revolutionary task—or to a counter-revolutionary reaction.

0. Practical Evaluation from Prehistory to Present

We can see the modern version of this practical process by watching a young language-learning child begin, with your help, to get a grasp on 'good move' and 'bad move' in the context of card and video games and building blocks. We'll call that 'good1' and 'bad1.' Concurrently, she's also trying to get a grasp on 'good2' and 'bad2', meaning 'what mommy thinks is good/bad' i.e., 'nice/naughty', and eventually she connects the two concepts, but that's hard. Hard enough that I bet you have to think for a minute to be able to explain what the connection is. [1] Essentially they start as two different concepts for her, like two definitions of temperature for a student taking a first science course (e.g., what the thermometer reads and what thermodynamics is about), and much confusion and frustration results, since parents often think they're the same for her since *they're* using the same term for each, or even that it's stupid of her not to understand that both concepts are really one. And of course, the parents often argue about what *is* good/bad or right/wrong, which doesn't help her climb the learning curve.

She's essentially struggling with the information-processing precursor of *understanding theoretical concepts*. It would be much easier for her if her parents were always very careful to give reasons for identifying actions as good or bad, e.g., unfair to others, or forbidden by God. However, we all do get hold of some general concept of evaluation, sooner or later, though not with great clarity, and, perhaps surprisingly for something so untidy, the good1/bad1 version at least is an enormously useful concept. It is not only crucial for improving our decisions and plans across the whole span of our life, but for improving our designing/making/modifying of artifacts and environments, and our perspectives on ourselves, others, and the world's story, from all of which we learn about our mistakes and triumphs, a kind of learning that can greatly improve our actions, futures, and reflections.

Direct instruction in the logic and limits of the evaluation process, in both its formative and summative—and even its ascriptive—roles, is arguably a serious omission from the standard curriculum of public education. That gap is partly due to the negative reaction that ostracized evaluation from the legitimate sphere of scientific concepts for half a century, courtesy of the positivist philosophy of science, but also perhaps due to excessive anxiety about evaluation because of its ego-threat, and no doubt also to the lack of any standardized, systematic approach. The nearest approach is perhaps the teaching of critical thinking, and the frequent opposition to that, especially because it can lead to doubting some parental edicts, is poorly matched by the limited or non-existent training for teachers to defend it or instruct in it.

At this point, another important role of evaluation needs to be mentioned, one that certainly required the development of language in order to appear. It is in fact a role that contributes to the evolutionary payoffs from having a language and supports advanced cognitive development. It's clear this function has been poorly understood and appreciated, and I'm going to preface my list of revolutions that we've endured or need to kickstart by pointing it out, because it's homely, useful, and relatively non-threatening—and needs to be included in explicating evaluation's functions. I have suggested names for the first three functions

[1] She may begin with an implicit grasp of the idea that 'good' is what brings the speaker closer to what they want to do, and she only gradually comes to see that it's more than this, because sometimes what her little brother, or even herself, wants to do is not good, at least not according to her parents. When is that so; and who is right when her parents disagree, or disagree with the government? Now, *that's* what makes it a tough concept to grasp.

that have achieved some currency, namely summative evaluation, formative evaluation, and ascriptive evaluation, so I'm going to suggest we add a fourth role to those—the *compressive* role, i.e., the function of evaluation as an invaluable and extremely powerful information-compressor of a particularly useful kind of information. The example I've often given of this because it's right under our teaching and learning noses much of our time in school and outside it, is the well-tested one of the academic grade. At the end of three months of talking to the students in one of our classes, reading their writings, and considering their questions and their answers to our questions, we have, let's hope, acquired a sense of their ability in the particular domain of skills or knowledge that we are covering. We need to record that conclusion, so that others, with a need to know and a right to know, can combine the summaries of performance provided by the range of instructors that cover the curriculum into an overall summative evaluation of the student's competence that will facilitate decisions by those selecting students for further education or awards or jobs. We may also hope that this summary, the single course grade or—at the more general level—the Grade Point Average, will *also* serve as a formative evaluation of some use to the student, for his or her own decision-making and improvement-making. In order to perform these worthy functions, each instructor condenses their evaluative summary into—amazingly, although with notable costs—a single letter, usually chosen from a small set of five or less, (A, B, C, D, F (a.k.a., E, in some countries); or P/NP (a.k.a., S/U (Satisfactory/Unsatisfactory)), or perhaps from a few more of dubious marginal validity if we add pluses and minuses.[2]

Now, isn't this just a case of convenient text design for our old friends formative, summative, and ascriptive? Well, 'convenient' is a bit mean-spirited for this characteristic: success in compressing valuable information by a factor of thousands or millions is a survival characteristic in itself, as with good graphic illustration and good meta-analysis, because it means that when time is short or your memory is overburdened, you can get or have handy in your head the information you need, not diluted with relevant but not critically helpful noise. For example, your medical needs probably don't include knowing *details* about the relative merits and side-effects of *all* the analgesics on the market—you just want to know the best one for you in your present condition, and that's exactly what your physician, if you are fortunate enough to have such a person available, can tell you. And s/he can tell you this just because—and only because—s/he has *evaluated* them, with skills based on available reports and experience.

Program evaluations, or executive summaries, that are too long for the client or his/her execs to read carefully, like oral medicines that are too large to swallow, must be recorded by the meta-evaluator as *failures*, because pragmatic considerations are just as important as content validity in the pragmatic enterprise of the evaluation consultant.[3] I think that evaluation theorists should start paying more attention to systematic analysis of these pragmatic criteria of merit—Michael Quinn Patton's initial shove in that direction got us thinking and acting much better in this zone, but I think we have not extended the logical analysis much further after these many years. One way of approaching that extension might be to say that in evaluating evaluation consulting and reports, one should be focusing heavily on the values of significance, outcomes, and cost-effectiveness rather than just on merit (which mostly means validity and coverage). Another dimension to cover is the informatics one suggested above in discussing the concision requirement (a.k.a., compressive function) of evaluation. I have elsewhere used that approach quite successfully for the related problem of the logical analysis of explanation, by applying informatics to it.[4] And I think evaluation can *only* be understood thoroughly (e.g., defined, or applied to itself) when we get it into that framework; so I put that task on the front burner for heating up later in this article.

[2] At least two distinguished institutions—St. John's and the University of California at Santa Cruz—have long refused to accept this 'ruthless bondage,' instead convening a committee to write a text account of the student's achievements. But what often happens, perhaps normally happens—apart from a large increase in labor costs—is that the next decision-maker in the chain of users (a role I have often had to play) executes the compression for them, without their permission or preference, converting the text to a grade (in order to get comparability), almost certainly less well than they could have done.

[3] Not necessarily true in the more academic world of ascriptive evaluation—e.g., the historian's usual world—since the old pragmatic constraints on length of books hardly apply to the cloud.

[4] "The Psycho-Logic of Modern Science" pp. 47-79 in *The Metaphysical Foundations of Modern Science*, Noetic Sciences Press, 1994, also at michaelscriven.info. After doing a doctoral thesis on the logic of explanation and forty years more work on various approaches to it, I only found an adequate solution in terms of informatics.

1. Fifty years an Outcast

The first revolution was an attempted genocide. When the emerging social sciences, at the turn of the 19th into the 20th century, sought an account of the scientific method that had been followed by the immensely successful physical and biological sciences, they made a very bad choice. They bought into logical positivism, as preached by the Vienna Circle. That meant condemning all evaluation to the status of the untouchables—part of the huge range of essentially unscientific propositions, along with non-propositional utterances like commands, exclamations, and questions. The positivists gave various reasons for doing this, but the leading group of these reasons was based on the claim that evaluative propositions were essentially mere statements of personal preference or expressions of taste, and hence entirely subjective and of no scientific status or interest. This is of course a generalization (and of course an evaluative one, hence self-refuting), and so the first question to be asked is whether the sample on which it was based was a representative sample of the population about which the conclusion is drawn. The answer is that it was an absurdly biased sample, since not only the two largest sub-populations of existing evaluative claims, but also (and distinctly) the most important one within the social scientists' fields, are *not* mere matters of taste.

The largest sub-population is of course the set of practical evaluations from everyday life in the produce markets, work, and home, which continued their valuable services unruffled by the blunders of bad social science methodology. Of course, there *were* plenty of 'mere expressions of personal (or family) taste' amongst these justified evaluations of produce, products, or projects, but these were treated correctly, mostly as *valuable information* for the food or financial shopper (i.e., needs assessment data). Of course, quite a few of them, e.g., claims about the best football team or sculptor or beer, were claims to truth that could not be substantiated, hence mere opinion masquerading as factual claims. But there were millions or perhaps billions of overlooked well-substantiated evaluative claims there.

However, the truly tragic error was the failure to recognize that the very sciences on whose supposed methodology the ostracism was based were themselves imbued from top to bottom with evaluative claims for which good evidence was provided and accepted—claims about data *quality*, the *merit* of theories, the *quality* of journals, the *superiority* of certain instrument-makers, the *brilliance* of scientists and students. The only good news is that evaluation eventually survived this attempt at assassination.

2. The Return to Respectability

Long before the editors of social science journals lifted their ban on evaluation articles, the work of Ralph Tyler and other renegades—comprehensive, quantitative, and careful—began to make the 'value-free science' doctrine look like a bias rather than a benefit. As the AEA and now a hundred similar organizations in other countries moved to a level of multi- thousands of members, as evaluation-specific journals moved up the usual library rating scales, as the number of models and checklists and training centers increased, it became clear that we had a new profession on our hands. But did we have a discipline?

3. The Alpha Discipline role

At first it appeared so, and I began to work on some of the metaquestions that have to be answered in any overview of a sprouting discipline. In particular, I was interested in the way that evaluation could be and should be applied to itself—a process I called meta-evaluation. Then I got interested in the way in which evaluation turned up in other disciplines, and activities like sports and business, in the form of some kind of quality control mechanism. It was encouraging to find that evaluation, like statistics and communications, had a powerful service function in all disciplines; to refer to this feature I introduced the term 'transdiscipline.'

However, it was somewhat depressing to discover that one of the three big organizations providing this service commercially, for businesses, had managed to completely exclude ethical considerations from its checklist of criteria of merit.[5] Also depressing was the discovery that the great classic disciplines, although they *thought* they had a quality control system, in fact the procedure that everyone immediately put forward as performing that function—peer review—turned out to have been hardly ever studied for simple but essential virtues like reliability and validity, and when Chris Coryn got down to details and actually did study it seriously, he found that, as this is commonly done, even at the national level, e.g., for

[5] They have now attempted to remedy this flaw, perhaps partly because of the roasting they got when they presented at the doctoral evaluation seminar at Western Michigan University, Kalamazoo.

funding research, it is a bad parody of the real thing, almost completely worthless in most cases. A huge Stanford study by John Ioannidis, recently concluded, addresses the other famous test of scientific validity—replicability—with equally astonishing, and depressing, results, roughly that virtually no-one, including some original discoverers of the most important medical breakthroughs in recent years, could replicate their famous results in a repetition of the original experiment.[6]

One moral of all this is that evaluators have a public responsibility to continue this kind of work—checking that the transdisciplinary function of evaluation is really serving other disciplines, not just having the name of evaluation taken in vain. And of course suggesting ways to improve the way it's done, when we find fault.

But there is another lesson to be learnt, and it counts heavily against the status of evaluation as a discipline. Not that we're likely to be run out of town, given the shortcomings of the town's senior citizens, but we really must get more serious about meta-evaluation (i.e., having our own work evaluated by external evaluators), which means not only doing this as part of standard procedure—as the ANSI standards, i.e., the US standards, require—but doing it well *and publicizing the results*. We can't persuasively talk the talk unless we walk the walk, and we must realize that professional evaluation is, by its very nature, particularly vulnerable to the kind of flaw we've been finding in other disciplines. This is because our real world practice is largely in the role of consultant, and consultants' work does not normally undergo peer review. We need to tighten up the trashy way peer review is done in other disciplines *and* use serious meta-evaluation to fill the gap in our own emerging discipline with respect to the job that we say (and can prove) that peer review ought to be done in the other disciplines.

It's also true that we have that duty because we are by definition the only discipline whose job-description includes the evaluation of evaluation procedures; hence we need to be especially careful, when judging others, that we are innocent of the crimes we blame them for. Just in case you think this is overkill, talk to anyone who has worked in international aid evaluation in the last 30 years, and get enlightened. Billions of dollars have been misspent there, and billions of people cheated of help they should have received, because of improper practice in evaluation and its use, from poor selection through poor design to lack of follow-up. I spent a couple of years as the external evaluator for the Bill and Melinda Gates East Africa project (working pro bono), and for all my admiration and indeed affection for what they are doing in terms of intentions, they would still be one of my prime examples in talking about how not to use evaluation.

What I've been talking about here is the role of the discipline of evaluation in studying how evaluation is done in other disciplines—important because without a satisfactory answer, they would not qualify as disciplines, since disciplines by definition have to meet high standards of methodological merit. I call this function the role of evaluation as *the alpha discipline*, because a would-be discipline must pass the standards for which we are responsible, nominally at least. This watchdog role also must, of course, include inspections, and instructions on how to apply and enforce these rules in order to be classified as a discipline. The need here is evidenced by the ease with which I was able to think of ten inexpensive ways to improve the evaluation of research proposals so that it becomes a reasonably reliable procedure.[7]

But there's a worse problem, methodologically speaking, not politically or sociologically or ethically speaking— that we have to fix in order that we ourselves can qualify as a respectable discipline. Earlier on I said that my first candidate for a 'revolution we should be leading' is better understanding of the pragmatics of the concept of evaluation. Now, I think some of you may be supposing that by reference to pragmatics I'm just talking about practical or real world consider-ations. After all, that's what 'pragmatic' means. But that's not the essence of what I'm talking about. The plural word, 'pragmatics' has a stricter meaning from the singular one (what fun it must be to learn English when you're not a kid!). And it's the plural word I'm talking about: here's the definition in the 'mother of all dictionaries'—the Oxford English Dictionary: "(The study of) the use of linguistic signs (especially sentences) in actual situations."

I think we can treat 'situations' as including both the signer's circumstances and his or her (or its) context. The use of the term "pragmatics" implies a contrast with alternatives in a particular

[6] It should be mentioned here that Ernie House was an inspired leader of this line of criticism with his own fine work on the corruption of the FDA review panels on which the commercial distribution of new drugs depends.

[7] In Scriven, M., & Coryn, C.L.S. (2008). The logic of research evaluation. *New Directions for Evaluation,* (*118*), 89-106.

triad of analytic points of view: syntactics, semantics, and pragmatics. Classical formal logic—from Aristotle through Godel and Carnap—was about *syntactics*. It morphed into *semantics* with the post-positivists, and made the big shift to *pragmatics* with Wittgenstein (not that he used or would have liked that name for his approach). It is also close to a strand in what is often called critical thinking, informal logic, rhetoric, or argumentation. I am afraid we won't have a secure discipline of evaluation until we have done a full analysis of the pragmatics of that concept, which today certainly includes looking at the informatics aspects of its use. Why is this important?

First, notice that despite the weakness in their quality control system, thermodynamics (like most of the classical disciplines) does meet this standard for clear, albeit implicit, definition of its core concept. We should meet it too, even though it's harder for us to do because our primary concept can't be defined via a network of governing laws in the way that physical concepts can; it has to be defined via pragmatics (e.g., ostensive definition—definition by pointing at examples—to use the old term for a notably early procedure in pragmatics). We have to adopt this somewhat unusual way to define it, because it's that kind of concept, although this elusive nature was one of the reasons it was cast into outer darkness for half a century, with its specter still haunting social scientists even today. And we are still some way from completing that project, although Jane Davidson and I are making a start with trying to identify what we're calling 'evaluation-specific methodology.' I hope we can get some help with that project from some of you.[8]

But that's not the only reason we need to put some work into cleaning up our own field. It's also partly because we won't get clear about the status of the highly prominent current theory-based approach to evaluation until we're entirely clear about the relation of explanation to evaluation, which requires pragmatics. More importantly, in fact crucially, it's important because we cannot give a thorough account of evaluation's function, its limitations, and its prerequisites, without analyzing its pragmatics. In other words, the core of evaluation, the backbone of our discipline, the key to what distinguishes it from traditional social science, cannot be set out without covering its pragmatics. For those of you not excited about getting the logical analysis right, let me segue into another line of reasoning for the same conclusion, by moving to the next era. But as a Parthian shot, let me say that one of the key elements in the pragmatics of evaluative reasoning is uncovering and explicating a third mode of valid reasoning beside deduction and induction, a quest for the Holy Grail that has been tried without success for a thousand years or so. I think we're now close to having found the 'Third Way' of reasoning; and that will make evaluation itself much clearer.

4. The Exemplar Role

At least we've made a start on everything mentioned so far, which make possible the first needed revolution. We'll call this necessary toolbox for the alpha role, the advanced logic of evaluation, which has to be able to deal with the evaluation of disciplines and not just the usual evaluands like programs and products.[9] Although we do have a good slice of that toolkit in working shape, we still have substantially more to do, and then we have to compress it into good teaching materials. But the next two items are really only hopes at this point, although a little easier to explain than the first.

The second needed revolution is a role reversal. Some leading texts on evaluation describe evaluation as a branch of social science, which is surely not true since the social sciences have no methodology for dealing with evaluative propositions, and are only partially recovered from the days when they put all evaluation in the untouchable class. I would like to see the exact reverse of this relationship come to pass. That is, I would like to see all standard social science texts being rewritten to treat the logic of evaluation as one of the key methodologies, like statistics and experimental design, that must be mastered in order to do all applied (and some pure) social science. In that way, good evaluation research designs will be the exemplar for much of social science, instead of social science treating personnel or program evaluation as something they can do with their current resources, albeit conceding that there are some specialists in these sub-areas. That's the view of evaluation fields as being satellites circling social science's sun. I think a more appropriate model is the reverse. Social

[8] And there is a little bit of funding support available for research related to that project from the Faster Forward Fund, which will be laid out in a future issue of JMDE, with space for questions and answers.

[9] The logic of evaluation means the process for identifying, defining, measuring, weighting, and—the key point—validating evaluative claims. The advanced logic of evaluation includes the process of setting up and applying a value *system*, which includes separating the axioms from the theorems, and using some inferences that go beyond deduction and statistical induction as valid schemas; and which can evaluate knowledge structures like disciplines, and metastructures like science and the humanities or arts.

science is special, by comparison with the biological and physical sciences, just because it tackles the problems of humans in groups (by definition), and the key difference between humans in groups and molecules or planets in groups, is that the humans are at least partly driven by values, that is—according to social science—by internal phenomena unlike physical or biological phenomena where the phenomena as well as their effects are directly observable. Taking this step would incidentally narrow the gap between social science and history, already hard to define, but inaccurately located by the positivists' map as reflecting social science's lack of interest in particular cases, a non-existent crevasse.

5. The Omega Role: Ethics as a Branch of Evaluation

The third 'necessary revolution' proposal in my wish list is another territorial imperative move by evaluation. Making that move is a necessary consequence of the recognition that ethics is by definition just an applied field of evaluation, because it is about determining the right and wrong, good and bad of something, mainly human behavior and values. Ethics is closely related to policy analysis but with intertwined strands of personnel, program, and portfolio evaluation (e.g., time management). I'm pretty sure that the lessons we have learned in those four existing sub-divisions of evaluation, when brought to bear on ethical problems, including what is often called 'the,' or 'the great,' ethical problem ("Why should anyone be ethical?") will yield new and improved results. Certainly we have to try out this kind of new approach, because we don't seem to be making great progress under the present management i.e., departments of philosophy, while the ethical problems in our society persist and multiply. And, or instead, depending on your predilections, we may find the reverse effect, i.e., that the so-called 'good reasons' approach to ethics (e.g., Toulmin and Baier) can improve our performance in one or more of those four subdivisions of evaluation. I hope others will join in the search for serious advances in this critical area.

References

"pragmatics, n." *OED Online*. Oxford University Press, September 2015. Web. 4 October 2015.