

# EVALUATION AS A REVOLUTIONARY DISCIPLINE

Michael Scriven  
Claremont Graduate University

**8<sup>th</sup> Evaluation Conference**  
**Warsaw 13.11.2012**

It is sometimes important to stand back and look at what we often call 'the big picture' in which our work fits as a small component. In our case, this means looking at the whole field of evaluation, which is made up of about 20 subfields, rather than just at policy analysis, program evaluation, or personnel evaluation, where most of us work. Three reasons for doing this are: (i) doing so often suggests connections that do not occur to us when working on problems within our own domain or sub-domain. It may also provide (ii) some sense of pride in what we do, a motivation that helps in defending our approach from the usual attacks on its cost, or its allegedly unscientific nature, or lack of utility. And it may provide (iii) a sense of the limitations and boundaries of the field, which makes us sensibly a little cautious about remarks that one often hears concerning the 'nature of evaluation.' I thought that an effort at such an overview might be appropriate for the beginning of the methodological discussions in this conference.

We can usefully look for new perspectives from either what might be called a geographical/spatial perspective (where the space refers to a map of the divisions of knowledge a.k.a., disciplines), or from a historical/temporal one of the usual and more familiar kind we've all seen in writings about the history of thought. I'm going to make a few comments from each of these points of view, and hope they may inspire you to some reactions to which I will have a chance to respond briefly in the question period for which I am going to leave a few minutes at the end of my talk.

Let's begin with a helicopter view of the geography of what I think is now established as the discipline of evaluation.<sup>1</sup> Following the dictionary definitions, this discipline concerns the entire zone of applications of the terms '**good**' and '**bad**,' and '**right**' and '**wrong**.' More specifically and usefully, as the dictionaries indicate, evaluation refers to the determination of **merit, worth, and significance**, three terms that are approximately the same, respectively, as at least one sense of **quality, value, and importance**. For those of you brought up with some training in the social sciences, this domain may strike you as illegitimate because you were inspired or taught by someone still committed to the world of positivist or neo-positivist philosophy of science, according to whom science cannot include evaluative claims since they are (allegedly) irretrievably subjective, vague, and/or untestable. But that view is completely wrong and was based on a superficial analysis of evaluative lan-

---

<sup>1</sup> Scriven, M., Evaluation as a discipline. *Studies in Educational Evaluation*, (Elsevier, Summer, 1994), 147-166

guage. While it is true that some evaluative language merely expresses a matter of taste (e.g., in arguments about football teams or modern painting), a great deal of it is testable and often objectively true—for example, its usage within science to express (and establish) ratings of good and bad hypotheses, theories, instruments, and data quality, and its use within professional test practice to rate the quality of student test answers and essays. Detailed professional work in the areas where evaluation is objectively supportable has led to the contemporary development of eight or ten sub-domains of professional evaluation which have good name recognition amongst most of you: program evaluation, policy analysis, product and personnel evaluation, performance evaluation (e.g., in athletics), proposal evaluation, and portfolio evaluation. Recent work, especially by Chris Coryn, has uncovered serious flaws in some of these areas, notably federal research proposal evaluation, but these flaws are fixable and in some countries have been fixed.

I have added three other types of evaluation to that list, of which two have become quite active research areas. The three are: meta-evaluation (the evaluation of evaluations), intradisciplinary evaluation (e.g., the examples from science mentioned above), and metadisciplinary evaluation (the evaluation of disciplines and would-be disciplines, e.g., the recent evaluation by the US National Research Council of the forensic sciences). In all three of these important sub-areas there is some very good and very important work. For example, Coryn has been called in by the governments in New Zealand, Canada, Switzerland, and Russia, to improve their national methods for evaluating the requests for federal funding in those countries, an activity that involves billions of dollars, where overdue methodological improvements can save millions.

Now that's the conventional scene in professional evaluation, and the achievements in these areas are what I would appeal to as the evidence that evaluation is certainly a profession and in fact a discipline. But there's *much* more to evaluation than that, since everyone does it every day, often with a high degree of skill that took years to acquire. Think, for example, of the evaluation of produce by the expert chef visiting the market before the sun comes up, to pick the fresh fruit and vegetables for the day's dishes. Indeed, there is a vast range of professional evaluators working in a non-academic context: the chef comes close, but the highly trained diamond buyer is a clearer example where precise measurement is inherent. The diamond buyer balances the cost of an uncut or cut stone against its score on the classic 4 dimensions of color, cut, clarity, and caratage (weight), and often does it within seconds—and then bets a fortune—and his or her future—on the evaluation. This is an interesting case because it can be both professional and virtually instantaneous; although the professional skills take years of supervised training to acquire, just like the skill of a policy analyst or program evaluator, they take virtually no time to apply. They are in fact largely perceptual skills, like many that the hunter or timber cruiser or tracker acquires..

So there are four divisions amongst evaluators: think of a 2x2 matrix, with professional vs. amateur headings on the columns, and inferential vs. perceptual on the rows. The helicopter view of the domain shows them up clearly, and it is just intel-

lectual snobbery to think that the non-academic evaluators are doing something less difficult or important than us.

Now there's another class of professional evaluators that has been omitted from most of our thinking about the scope of evaluation, and since it's a very respectable class, it's good to keep it in mind when defending one's turf from superficial attacks from people still haunted by the specter of 'value-free science.' In the elite set of classical disciplines—the subjects that were studied by the 'educated people' in the Greek and Roman empires—when one comes to think about them carefully, it turns out that half or more of each of them is evaluative. I'm thinking of logic, medicine, mathematics, ethics, and engineering (civil and weapons especially). Logic is half devoted to the evaluation of arguments, ethics is half about the evaluation of acts and attitudes, engineering is centrally concerned with the evaluation of designs and construction, and medicine with human health and illness. These disciplines never took the doctrine of value-free science seriously; it would have completely destroyed them. And they have been safely considered and practiced for some thousands of years, so the doctrine that science must be value-free is clearly false.

The value-free doctrine was based on a recommendation carelessly put forward by a group of people trained in physics, chemistry, and biology as a core element in the methodology of science as they understood it. Because those sciences were so successful around the turn of the 19<sup>th</sup> century into the 20<sup>th</sup>, exactly the time when the social sciences were beginning to scabble for legitimacy, it was understandable that the latter group picked up what were said to be key elements in the success of physics etc., warranted valid by Mach and the Vienna Circle. But they were misled, and the results were disastrous, not only methodologically, i.e., for the future of the social sciences, but also ethically, since the ethical side of human behavior and thought was ruled out as a legitimate domain for scientific investigation. That meant that huge policy issues were dominated by the biased and sometimes immature value system of the political figures and parties in power.

As I suppose one must expect, once the value-free doctrine had been adopted, it stuck hard and was very hard to dislodge, despite the evidence to the contrary from the more ancient sciences and other subjects that stood as sturdy counter-examples, including the evidence from half a dozen sturdy new branches of evaluation like our own, all of which have been applied to a dozen or a score of areas like health and education and defense work with good and verifiable results. Instead, the ghost of that wretched doctrine still echoes through the halls of the social sciences, so that we find recent handbooks of applied social science that contain nothing about evaluation, although 90% of the questions that applied social science tries to answer are evaluative questions. A remarkable persistence in a doctrine with so many obvious flaws!

It now appears to me that we are not going to get the full respect that evaluation deserves until we provide a completely reconstructed philosophy of science, and I have almost completed my work on that, although it's a good deal too long to insert in this talk. Instead, I want to turn to the second half of my task today, which is to review the historical perspective on the development of evaluation and its huge im-

plications for our many disciplines of concern. This review, although too brief to prove every point I mention, may be enough to encourage some of you to rethink your overall picture of the relation of the disciplines—to each other, and to evaluation—and in particular to understand what I mean when I talk about the respect that evaluation deserves. And it will do part of the job of providing secure foundations for our discipline, including avoiding the trap of thinking that being value-free is part of good scientific methodology.

Evaluation is a cognitive process; for most of us, in our professional life as evaluators, it is a conscious process, an inferential process; but for some of us, now and far into the past of the hominid species, it has been routinized into a perceptual process, by lengthy learning and training. We know from the archaeological evidence from the middens that serious product evaluation has been going on for more than a million years, as the flint-chippers gradually perfected and extended their craft. But we can reasonably infer that it was going on long before the stone age, although the relics have rotted away; there were wooden bowls and spears, thatched huts, and fish-nets—and clothing of which we do have a few specimens. But there can surely be no doubt that personnel evaluation was going on even amongst our *pre-linguistic* ancestors as they chose leaders and mates and teachers to help their children master the skills of tracking and fishing and food-plant finding. Wherever there is teaching, which began at least a million years before there was a spoken language, there has to be evaluation, because teaching is judged as good or bad by the outcomes in the learners, and that judgment requires evaluation of their achievements, just like the distinction between good and bad performances in the pupils. Once language emerged and developed to a modest level, plans and proposals would be possible and of course evaluated, certainly a thousand years before the colossal engineering projects like the Pyramids and the Great Wall. The early hominids were solving the problems of survival 3.5 million years ago, and one of their essential tools was the cognitive process of evaluation. Sometimes it resulted in explicit knowledge, sometimes in tacit knowledge, but very often it was evaluative knowledge.

Of course there is a temptation to think of all these evaluative activities as very primitive, but the more we study what our ancestors made and did, and how hard it is for our contemporaries to survive in the recent spate of 'reality' TV shows where people are put into tropical settings without modern technology, the more we come to realize that what was done was very hard to do and represented great achievements. It seems clear that millennia before anything like science emerged, homo sapiens had build up a very large repository of hard-won knowledge, much of it non-verbal, but much that was verbal, and that a great deal of this knowledge was crucial for survival. And much of *that* knowledge was *evaluative knowledge* about the best way to do things, or cook things, or the best things to eat and avoid, and the pitfalls to avoid in getting them. This included much idiosyncratic knowledge about who was the best fisherman to go out with, or the best warrior to be on the side of, or the best leader to follow; but it also included masses of (at least regional ) generalizations, e.g., about the general characteristics of ripe mangos and poisonous snakes. Of course, a great deal of non-evaluative knowledge also had to be mastered, e.g., the route to the best hunting or gathering grounds, and who had to be obeyed or avoid-

ed. But the bottom line was that much of the hard-earned and invaluable basic knowledge that contributed to survival was *verifiable evaluative knowledge*. So a look at the history of our species adds extra weight against the view that such knowledge is merely an expression of preferences or taste and hence completely subjective, as the positivists argued. Refining our evaluative knowledge—not rejecting it all, just refining it—is one of the functions of science (and technology) and we evaluators are the scientists who do just that, standing four square upon the hard work of our ancestors.

Now we have segued into the history of evaluation, and I want to put some structure into that history by separating it loosely into periods during which very different paradigms of evaluative knowledge were dominant. And I want to use that sequence as a springboard from which to dive into the future of evaluation—or at least its possible futures, partly depending on whether I can persuade you of the way I think we ought to be moving.

Paradigm 1. From about 3.5 million years ago until about 1900 A.D. **The commonsense paradigm** prevailed, i.e., the view that evaluative knowledge is important, verifiable, and at least on a par with non-evaluative knowledge in these respects.

Paradigm 2. From about 1900 to 1950. **The paradigm of evaluation as scientifically worthless/untouchable**. No leading social scientific journals accept evaluations for publication, or articles using evaluative terminology. This is the first of the great paradigm revolutions involving evaluation.

Paradigm 3. (About) 1950 to now. The counterrevolution begins to build: the commonsense paradigm returns, albeit only at the fringes of the social sciences, most seriously in educational research. It has a new twist, however, for program/personnel/policy **evaluation starts moving beyond mere acceptance towards professional status**. Unfortunately, the idea persists amongst high-prestige social scientists that ‘real science’ or ‘quality science’ avoids evaluation, or that there is a difference between facts and values, or evaluation and research, or evaluation and description, the three false dichotomies that are signs of an (often unconscious) commitment to the value-free doctrine.

Paradigm 4. (About) 1990 to now. The concept of evaluation as a discipline begins to crystallize, this being a step beyond the status of profession, one that is marked by clarification of the limits and core of the field of study, and its special methodology and concepts, if any. From this discussion, the special features of evaluation in the domain of the disciplines emerge, led by the idea of evaluation as a *transdiscipline*, i.e., it is one of a small group of disciplines that include statistics and communications, that have a role inside other disciplines as well as a standalone role as an autonomous discipline. The essentially unique feature of evaluation is that is a key part of *every* other discipline including even the physical disciplines like gymnastics, ballet, and marathon training, since every discipline, by definition, has a set of standards for data quality, inferential validity, acceptability for publication, significance for various honors, etc., and these standards are subject to the requirements of the dis-

cipline of evaluation. This situation can be put by saying that **evaluation is the alpha discipline**. The need to take this role, and hence evaluation, very seriously emerges as: (i) studies of the precarious situation of peer review—the key quality control mechanism of all sciences—make apparent; and (ii) major scandals in e.g., anaesthesiology and the forensic sciences make it clear that the sciences are not running an even marginally adequate quality control system, or funding dispersal systems. Quality control systems in all disciplines should simply be treated as an applied field of evaluation, and work has now begun by evaluators in collaboration with leading scientists on designing improvements for them. Evaluation, in short, is the keeper of the keys to the kingdom of the disciplines, hence the term ‘alpha discipline.’

This approach is naturally treated by many scientists as an invasion of their territory, which qualifies it as revolutionary, but the transdisciplinary role is complementary not dominance. It will incidentally—from the purely logical point of view—bring to bear and deal with the issue of bringing social payoffs into the calculus of evaluating research projects and results, something truly revolutionary in the traditional sense.

Paradigm 5. 2012. Now we come to the emergence of the paradigm of **evaluation as the exemplar discipline for all applied disciplines**. Supporters of this paradigm recommend that the mainstream applied disciplines, from the social sciences to engineering, adopt the model used in sophisticated evaluation studies (program, product, policy, etc.), which treats the evaluation element as equally important and distinctive from the non-evaluative element. One problem about implementing this recommendation will be to shake up the leaders, especially of the applied social sciences, enough to convince them that this is necessary; the alternative is clearly that they will become implausible candidates for public funding. The other problem is to clarify the logic of values, the tool that evaluation studies bring to the party, and in particular the process of validating and ranking values. Of course the idea that the social sciences—and other applied disciplines—should convert their attitude towards the discipline that they have long regarded as junk into deference or even respect, is revolutionary, so will likely take half a century, the duration of their previous ill-justified dismissal.

Paradigm 6. 2012. The paradigm of **evaluation as the warden of the alpha value**. The alpha value is of course ethics,<sup>2</sup> and its validation in that role spins off from paradigm 5. As the alpha role approach gets traction, the need to control the clearly righteous push for cultural sensitivity so that it stops short of ethical relativism becomes a focus of attention. Evaluation has to undertake not only the validation of the ethical value, but the justification for treating it as the top-weighted value when conflicts of value emerge. Previous would-be owners and disowners of ethics, e.g., religions and neo-positivistic scientists, will regard this paradigm as an intolerable

---

<sup>2</sup> More exactly, the ethical axiom of prima facie equal rights for all people. The usual additions to this in the moral premises of any particular faith, can be obtained from this one, shared by all, along with some non-evaluative premises about culture-specific preferences e.g., an institution of private property or monogamy.

shift of power, but it has long been clear that their positions have been destroyed by progress in game theory, evolutionary theory, comparative psychology, theological critique, cosmology, meta-ethics, and the logic of evaluation, so this revolution is overdue.

And so I conclude this brief excursion into a relatively fast-moving battlefield in the history of thought, with the following reflection. I expect the most common reaction to this approach will be that it exhibits an absurdly exaggerated sense of the importance of a relatively new discipline. To that I reply: it has been science that committed the sin of pride in this dispute. To begin with, bad science led scientists to dismiss evaluation as merely the expression of preference, an absurd result when they themselves were evaluating the elements of their own science all the time and they had two million years of hard-earned evaluative knowledge in every archaeological textbook pointing the other way. And then today, worse science has forced us to turn to evaluation to correct the blunders made in preventing fraud and carelessness from ruining good science. So my message is not just that evaluation's return to respect is obligatory but the obligation was created by those now complaining about those waving its banner to rally real scientists, people who really value finding out the truth and not protecting their position of unjustified eminence.

The bottom line is that these revolutions are the best hope for salvation from the bad effects of bad science.