The Calorie Count of Evaluation

Michael Scriven
Claremont Graduate University

We all know one kind of answer to the question of the value of evaluations; they can help to improve the evaluands, and/or establish accountability for the expenditure that created or supports or buys the evaluands, and/or increase our knowledge about the evaluands' merit, worth, or significance, etc. Those answers are about the useful *functions* of evaluation; they're like saying that food can help build bone structure and/or muscle mass and/or brain mass; these are all functions that food performs. But there's a more fundamental and more specific level at which we say that a particular food has a certain calorie count, a certain fat content, a certain sugar content, has some iron or peanut oil or salt in it, etc. That level of analysis—a kind of nitty-gritty or *component* level—is very valuable for the nutritionist and for many consumer concerns. What is the equivalent of this kind of answer for evaluations? Is this a level of analysis we have been overlooking?

What brought up this line of thought was a four-letter meta-evaluation of the five-year impact evaluation of Heifer International's efforts at poverty reduction in 20 countries that I recently designed and directed. A student in the evaluation doctoral program at Western Michigan referred to this effort, in a written comment she may or may not have intended me to see, as 'crap.' While that's not the sort of meta-evaluation that evaluators hope to get, it's a hell of a stimulus to take a second look at what one's doing, and whether it's worth anything. And it's admirable for at least one reason—its brevity—although it doesn't make the Guinness Book of Records since the letter grade F makes it look rather longwinded. Of course, this meta-evaluation is a little short on documentation, so we'll have to speculate about that. I hope she'll respond here—she can do it anonymously if she prefers—if we misconceive her reasons. In any case, I think we can learn something useful from looking at possible defenses against this dismissive meta-evaluation, and in doing so, learn something interesting about evaluation.

The Five-Year Study (5YS, for short) was supposed to be, and claimed to be, an impact evaluation, which means that the central feature of the design had to be a method of solving the attribution problem, i.e., the problem of establishing the causal implications of an intervention and/or the causation of observed phenomena. The method I developed for this purpose was an elaboration of the standard scientific procedure in forensic, epidemiological, and geological sciences

(and others), and the standard procedure in scholarly history, and in the law, and in almost all of technology. I call it the 'general elimination method' (GEM) and discuss it in some detail in this journal's Summer, 2008 issue; it is a qualitative method. Now, in the current war about causation, most readers will be aware that this kind of approach is regarded as unsound by a substantial group of reputable social scientists, who favor some version of a control group study; many of them think that only a control group study with random allocation of subjects between experimental and control groups—the randomized controlled design (RCT)—is entirely acceptable. Someone from that group, perhaps including our meta-evaluator from the doctoral program at Western Michigan University, might well regard the 5YS as hopelessly flawed and hence as crap. Now, don't worry, I'm not going to just go over *that* argument again here! Let's try a more radical approach; let's just suppose that the RCT group are right. Does that leave anything worthwhile in this or any other non-RCT impact evaluation?

The problem with what we can call 'the RCT approach' as a basis for a dismissive meta-evaluation of a GEM evaluation is that *it ignores the calorie count*. It's like someone who says vegetarian meals aren't 'real food,' meaning that such meals exclude a component the critic regards as essential. If we suppose, for the moment, that the carnivore is right, that a good diet really has to include meat, let's just look for a moment at what a merely vegetarian diet *does* include that *must* be counted as valuable in the currency of good nutrition, or analogously in evaluation, even if it's missing a key ingredient. The results may surprise you, because—I suggest—we have been too inclined to regard evaluations that fail to include that key ingredient as worthless, when we should have just said they are flawed. The flaw, *even if fatal,* does not justify consigning the evaluation to the trash can, only to the recycling can—from which we can recover much of value. The difference in nuance between those two conclusions is one to which evaluators, including meta-evaluators, should be very sensitive, and it suggests a reclamation process that we have been ignoring *whether or not we think that our standards for establishing causation have been met*. So now it's time to ask how much of value, if any, there is to recover?

We can introduce the answer via a list of the questions that need to be answered in any good thorough evaluation, and that can be answered without any debatable assumptions about whether the intervention caused the results claimed. This will be our calorie counter. I think it's clear that *these* questions, and more, were in fact answered at length in the 5YS case, and in order to give specific examples, I'll draw on that case to add some realism, but what I'm mainly after here is a general conclusion. The first point or two below are covered in more detail than the others, but even with them, many of the details are omitted; we used a 96 item checklist for evaluating impact in each of the two hundred or so recipient villages we visited, and there were other components to the evaluation besides the village visits.

1. *Was the intervention under study actually delivered—or to what degree was it delivered—to the alleged recipients? Was the intervention, if delivered, a match to the description of it used by the organization that invented and/or supports its use? These major questions should be interpreted as requiring coverage of*

the 'dosage' problem that bedevils every drug study, i.e., (i) checking for a gap between ordering and delivery; (ii) checking for a gap between delivery and use; and (ii) checking for variability in types of use and not just quantity of use. In 5YS, this meant talking to the families—in their own homes—about the age and quality of the livestock they had received from HI, the effort and costs involved in its maintenance and care, how long it was kept, the number and health and disposition of its progeny, the income it generated through sale of progeny or products such as milk; and the relevance and completeness of the training in its care and in managing its offspring, and their impregnation and birthing, that was also provided.[1] A great deal was learned from these exchanges, for example that in a few cases the livestock provided was sick, or of poor quality, indicating serious failures in the quality control (QC) system that were at first denied by the country staff, a denial we repudiated with evidence, leading to full replacement and improvement of local QC.

2. *Were the recipients in fact those targeted, i.e., did they match the description of them used by those funding and/or those delivering the intervention; and if not, who were the actual recipients and which targeted recipients were missed?* Information gathered in dealing with this question for 5YS covered the extent to which the recipients were the 'poor farmers' the program described as being helped—in rare cases, they were rather well-off by local standards, and the reasons for the exceptions were uncovered and the explanations for inclusions evaluated (in most cases they were satisfactory). These matters are very important for the credibility and merit of the program and for suggestions for improvement.
3. *Were the plans and processes of delivery and training ethically, culturally, and politically appropriate? Were they effective and efficient?* For example, was bribery or skimming involved, or discrimination against women or natives in the personnel management of the organization in charge of field operations? In particular, for 5YS, how consistent were field practices with the Twelve Cornerstone values to which Heifer International has always been committed (which include gender equity), and subsequent modifications of and additions to them by the governing board? How good were the Cornerstones? And how good was the instruction about them? These questions required some extensive value analysis to answer, as well as empirical work and knowledge of the state of the art in training/teaching methodology (including knowledge of ultra-low-cost computer-assisted training). We observed considerable variability in the quality of the training, which led us to generate a general checklist for evaluation of training—an extensive development of Kirkpatrick's groundbreaking effort at this. (Some overlap with Question 1 here.)
4. *How well did the staff from the intervention organization, and the people they hired: (i) do their job; (ii) treat recipients (and others they observed), apart from doing their job?*

[1] Yes, there are a couple of causal claims in there, but I'm not cheating on my conditional assumption that the GEM approach is invalid; see the discussion under 6 below.

What other suggestions could be elicited from recipients or other players or evaluators about how the intervention effort could be more helpful or do more good with (more or less) the same expenditure of time and money? Input from recipients (and their neighbors) and other stakeholders, on this set of questions, is arguably an ethical obligation and certainly a valuable source of good ideas. We made several recommendations based on such input. Of course, we also looked carefully at the large slice of the whole management and logistical process to which we had access, and made many suggestions ourselves about how we thought this could be improved e.g., that in the general management system, there should be a Plan B, for which staff were trained, for handling various possible disasters. While non-trivial recommendations are not part of the logical process of evaluation, and are not deducible from it, they are clearly a useful option that an evaluator is often in an excellent position to produce, and there are many evaluations where a good recommendation from the evaluator, all by itself, produces gains for the client that far outweigh the cost of the evaluation. Keep in mind that you do not need a program theory of the evaluand to produce recommendations, only a few scattered causal links; this is true even if your recommendation is for a completely different program.

5. *How did life for those who did not receive the intervention differ from or match the life of those who did?* This would presumably be a requirement for any competent evaluation, whenever it is possible. For example, we regularly interviewed non-recipients in the same village to answer this question, so we had a crude comparison group (differing in eligibility factors, or choice, or time in village), from whom we got valuable information about droughts, floods, changes in market prices of feed, range, and vet meds, and taxes and local roads and other infrastructure changes, and of course information about direct interventions by other NGOs and GOs aimed at improving conditions. All of these alter the quality of life, for some, in some ways, and must be factored into any comprehensive causal story; but, without any reference to causation, they give us a context that can greatly alter the significance of the intervention, and hence its phenomenological impact, which is not just a quality of life issue but a driver of morale and resilience or despair.
6. *What were the immediately perceptible benefits of the provably delivered version of the intervention?* We started on this earlier (in question 1) by asking what it took to feed and care for the donated livestock—a cost of the intervention to the recipients. That cost was immediately perceptible to the recipients. Of course, that question and the present ones are causal questions, but they are not a violation of our condition, which only excludes 'debatable assumptions about causation.' Not even an RCT enthusiast would deny that if you give your nephew the money to buy a bicycle he has longed for, and he uses it to buy the bike, that your gift caused this increase of his material possessions. Similarly, the Heifer recipients can indubitably identify an

> increase in the consumption of milk by their children, or of money in the household budget from selling the milk, as due to the gift of a pregnant heifer. One might as well deny that the crowd at a baseball game is not entitled to their conclusion that they just saw the player at bat hit the ball into the stands. The plain fact is that many causal claims can be established beyond reasonable doubt by direct observation, and it was a bad influence of positivism to teach the contrary, an influence we need to grow up about and move on from. Of course, there are many causal claims that direct observation cannot possibly establish; for example, that the ingestion of steroids was the cause of this hitter's remarkable statistics for the season, or that *all* the improvements in a village family's quality of life over the period following a Heifer gift, were due to the gift. To establish those claims requires a more sophisticated design; sometimes RCT, but, in my view, sometimes also GEM. But I'm not assuming that GEM is valid in 1 through 6; for those conclusions, I don't need it any more than I need RCT, to establish the key result—that this intervention produced very large sustainable[2] benefits to the quality of life of very poor farming families.

So the bottom line is a double claim: (i) a serious and extensive evaluation will generate a mass of valuable evidence that well serves the many functions of evaluation—see 1 through 5 and perhaps also 6—*even if its core design is flawed*; (ii) some serious evaluations can establish many or most of their conclusions, including causal ones, using *observably verifiable causal claims* as in 6. So don't judge a book by its cover—what's on the cover are just headlines. Read it carefully for valuable content. And don't judge a diet or an evaluation by the flag it flies—count the calories.

[2] The sustainable part requires detailed further evidence not discussed here, but of the same kind.