EVALUATING EVALUATIONS: A META-EVALUATION CHECKLIST

Michael Scriven
Claremont Graduate University

What are the criteria of merit for an evaluation in any field, including program evaluation? Any professional meta-evaluator—someone who frequently and professionally evaluates evaluations, e.g., mid-level managers in research or evaluation centers, or editors who publish evaluations—and perhaps even every evaluator, has a list of these, although it may be implicit in their practice rather than an explicit part of it. Making it explicit facilitates evaluation of it, and that facilitates improving it, the aim of this effort.

Moreover, such a list is very useful, not just for evaluators and meta-evaluators, but for their clients (and prospective clients), critics, and audiences; clients, including editors, are of course very important meta-evaluators in practice, since their conclusions pay the bills for evaluators—or make their name, which helps towards paying the bills. Several suggestions have been made for such a list, some by me e.g., in *jmde.com*, and most famously by Michael Quinn Patton with his utilization-focused evaluation. But I think we might be able to do a little more, at least in terms of detail. Here's my latest effort, in the hope it will inspire corrections and other suggestions.

This issue is of considerably broader significance than the title might suggest; for the criteria of merit for evaluations heavily overlap with those for any reports of applied scientific work, so the checklist below could be useful for editors and clients in those fields, too.

Note that this approach differs from MQP's in that it does not treat utilization as a *necessary* criterion of merit, although it's nevertheless heavily utilization-focused, i.e., aimed at maximizing utilization. This apparent paradox, which MQP avoids by making utilization a defining criterion of merit, is not paradoxical since it simply allows for the fact that poor utilization can be the fault of the client: it may be due to suppression, or careless misinterpretation, or deliberate misuse of the evaluation by the client. Its absence can only be blamed on the evaluator *if* the evaluator was responsible for it via a weakness in the evaluation e.g., its lack of relevance to the client's questions, or its lack of clarity. Also, it seems to me that one should divorce the merit of an evaluation from its utilization in order to avoid giving *any* credit to an evaluation that is immediately utilized, although it's invalid; and not much credit to one that cost far more than was necessary. So I believe that although utilization is an essential *goal* for a good evaluation, it is not a defining feature, just as I believe that democracy is an essential *goal* for a justifiable political revolution (e.g., in Libya today) but *achieving* a democratic government is not a defining feature of a justifiable revolution, since it may be aborted by ruthless countermeasures.

The most useful list of defining features of a good evaluation depends on the level of the inquiry. Within a subfield of evaluation, for example program evaluation, there are some good checklists of specific matters that have to be covered by good evaluations, with some guidance as to how they should be covered. These include the Program Evaluation Standards, the GAO Yellow Book, and the Key Evaluation Checklist (the latest version of the latter is available elsewhere on this site). There are also many such lists in subsubfields, e.g., for the evaluation of computer hardware and software within product evaluation. The meta-evaluator can always proceed by simply using one or more of these as setting the standards for the matters that must be covered by—and to some extent, how they must be covered by—a good evaluation. It's almost essential to refer to them in order to cover the matter of validity, which is the first criterion of merit. But it's also useful, in both teaching evaluation and in its practice, to have a higher-level list that will apply to any subfield. It may also be useful to have this in order to evaluate the subfield lists themselves—e.g., in order to pick the best set of program evaluation criteria against which to measure designs for a particular assignment. In fact, perhaps a little surprisingly, it can be very helpful for non-evaluators to have such a list, couched in general terms they understand, when they are trying to judge the merit of an evaluation they are reading and may in fact have commissioned. We'll call this attempt at such a list the Meta-evaluation Checklist (MEC).

With, or even without more sophistication about evaluation in a particular field of evaluation, the next step after using the MEC is to apply one of the checklists of required-coverage items mentioned above. For program evaluation, my preference is for the shortest, the KEC, since the five core checkpoints in that list (listed later here) are reasonably comprehensive and still make sense to non-professional audiences or clients.

NOTES: The term 'evaluand' is used here to refer to whatever is being evaluated… The key criteria *and sub-criteria* involved are initial-capitalized… The first five criteria have non-zero 'bars,' i.e., levels of achievement each of which *must* be cleared (i.e., shortfalls on bars cannot be offset by any level of superior performance on other dimensions.)… The level of detail, particularly under Validity, is for the professional, and can be skipped over by the general reader…

THE META-EVALUATION CHECKLIST (MEC)

1. **Validity** This is the key criterion—the matter of truth. There are several major topics to be addressed under this heading, of which the first two are the dominant ones. (i) The first determines what might be called the rules of the game, that is, it determines what *kind* of meta-evaluation is required—the *contextual constraints*. We'll also need the same for its target, i.e., we'll need to know what kind of evaluation was originally required.[1] These needs assessments are largely a matter of pinning down: (a) the *focus* of the evaluation required—for the meta-evaluation, this means answering the questions, Exactly what *is* the evaluand, and What *aspect(s)* of it should you be evaluating—an evaluation's *conclusions*, or its *process*, or its *impact*, or all of these—and should the meta-evaluation be designed for use as summative, formative, or simply ascriptive[2]; (b) what about the *function or role* of the original evaluation, particularly whether *it* is/was supposed to be formative, summative, or ascrip-

[1] This criterion therefore refers to a double *evaluation* needs assessment, not to be confused with the original *evaluand's* needs assessment, i.e., the needs assessment for whatever the original object of investigation was, e.g., if it was an educational program, it will have needed an educational needs assessment.

[2] Evaluations done simply to increase our evaluative knowledge are ascriptive; examples include most evaluations done by historians of the work or life of historical figures or groups.

tive; (c) what *level of analysis* is required on the macro/micro scale—holistic or analytic; (d) what *logical type* is required—ranking or gap-ranking,[3] vs. grading vs. profiling, vs. scoring vs. apportionment; (e) the *level of detail/precision* required (virtually all meta-valuations ever done were partial e.g., because they did not go back to examine the original evaluation's data-gathering process and its error rate) so you have to settle on what counts as adequate vs. excessive detail for the present context, especially since this massively impacts cost; and (e) what, if any, are the other *contextual factors*, i.e., assumptions about the environment of use of the evaluand, probable audiences, maximum time and cost restrictions, etc.

(ii) The second component of validity is the matter of the probable truth of the conclusion(s), given the parameters established in the first component. This involves two main dimensions: coverage and correctness. Coverage is the one for which having an area-specific checklist becomes important: in program evaluation, the KEC tells you that there must be a correct Description of the evaluand and sub-evaluations of Process, Outcomes (including unintended outcomes), Costs (non-money as well as money), Alternatives, and Generalizability.

Correctness means, in general, that the relevant scientific (or other disciplinary) standards are met—in other words, the adequacy of the evidence and the inferences that are provided to support the proposed evaluative conclusions in the target evaluation. Still in general terms, this part of a meta-evaluation is particularly focused on: (a) logical soundness (including statistical soundness, where statistics is involved), and (b) the usual requirements of adequacy of scientific evidence within a domain; (c) evidence of confirmation or at least confirmability, for the evaluation as a whole. This is conventionally established via 'triangulation' (which may of course involve only two or more than three sources) of its conclusions from *independent* sources—which may be of one type, but is strengthened if it comes from more than one logical type, the list including: direct observation, reported observation, test/measurement data, document data, theoretical, logical, analogical, and judgmental sources. In the case of evaluations, there is another element that also has to be examined for validity, meaning Coverage and Correctness, namely the values component. Doing this means: (d) checking whether all relevant values were *identified*, and whether they were *specified* in the detail needed for this evaluation, *scaled* appropriately, *measured or estimated* reliably, and finally *integrated* in a defensible way with the empirical findings in an inference to the appropriate sub-evaluative and overall conclusions. (Don't forget to begin with the simplest value of all, truth, and the simplest case of its relevance—the description of the evaluand. Often enough, the *client's* description of the evaluand is question-begging, i.e., assumes merit, e.g., is entitled 'an external evaluation' when in fact it's only partially external.)

[3] In gap-ranking, an estimate of the intervals between, as well as the order of merit, of the evaluands is provided. It may be only a qualitative estimate or, as in horse-racing, a rough quantitative estimate ("by a head/neck/nose/3 lengths"). Gap-ranking is often an extremely useful half-way case between bare ranking and scoring.

(iii) Note that Validity at least requires Reliability i.e., a reasonable level of inter-source (including test-retest) consistency. But validity requires more than mere consistency between several sources: it requires some evidence of 'real' value, which usually (not quite always) means visible or directly testable evidence somewhere along the line of implications of the evaluation. For example, we expect drugs or programs identified as 'better' to result, sooner or later, in measurable, visible and/or felt benefits to patients. The lack of this 'reality connection' is what makes mere agreement amongst wine or art critics unconvincing as to the validity of their evaluations, since the validation process never gets outside the circle of opinions. Nor is science immune to this mistake (think of the essentially universal agreement up to 1980 that antibiotics were useless for treating ulcers, an agreement which turned out to be completely unfounded). At a more fundamental level, almost all scientific evaluation of research proposals rests on peer review. According to the above generally accepted claim, peer review must, to be acceptable, at least meet the requirement of reliability; but it fails to meet that requirement at any acceptable level, in the few studies that have been done, so the current peer review approach to proposal evaluation is invalid[4]. (There are ten affordable ways to strengthen it, so an improved version of it could well be satisfactory.[5])

(iv) A validity-related consideration is Robustness, i.e., the extent to which the conclusion depends on the precise values of the variables it involves. A meta-evaluation, like any evaluation, is *more* valuable if it's *less* dependent on small variations or errors of measurement in the factors involved. We consider this issue in more detail under checkpoint 6, Generalizability, below.

2. **Credibility** to client/audiences/stakeholders/staff, meaning a combination of low level of Probable Bias e.g., from COI (conflict of interest) with a high level of Expertise. (The focus here is on matters of credibility *not* covered by directly checkable validity considerations. Obviously, the big issues here are *independence* and *relevant experience*. Certainly ties of *friendship, finance, or family* compromise independence, and must be checked; but less obviously and more commonly, it is corrupted by the totally pernicious *over-specification of the design or management* of the evaluation by the client, via the RFP or the final contract. It seems so reasonable, indeed mandated by accountability concerns, for the client to supervise the way their money is being spent that one often finds a requirement of frequent checks with a client liaison person or committee included in the contract; it is absolutely impermissible, as mistaken as requiring mid-operational checking with your surgeon. As to design control, even this checklist seems to encourage it under Validity; but propriety requires an absolute separation of those necessary coverage considerations, which can be included in the contract, from major technical design issues, which must not be specified. This is a hard distinction to draw, and open discussion of it is essential, possibly including appeal to external specialists on this issue. Again, recurrent for-

[4] Coryn, C. L. S., & Scriven, M. (Eds.). (2008) Reforming the evaluation of research. *New Directions for Evaluation, 118*.

[5] For details, see the footnotes in my "Conceptualizing Evaluation" in *Evaluation Roots,* education. M. Alkin (Sage, 2011).

mative evaluation (MQPatton's 'developmental evaluation') makes so much sense that one tends to overlook the inevitable role-switching it involves, from independent external evaluator to co-author of the program (or spurned wannabe co-author): there must be scrutiny of this, and probably one should require that the developmental evaluator requires the client to inject irregular formative evaluation by another evaluator.

3. **Clarity** i.e., a combination of Comprehensibility to the client/audiences/stakeholders/staff, with Concision, both factors that reduce effort and the frequency of errors of interpretation, and can improve acceptance and implementation. (The PES, the KEC, and the Yellow Book are all deserving of some criticism on this account, and it may be the leading factor in determining their relative merit for a given purpose.)
4. **Propriety** meaning ethicality, legality, and cultural/conventional appropriateness, to the extent these can be combined: this must include consideration of respect for contractual obligations (e.g., timelines), privacy, informed consent, and the avoidance of exploitation of social class/gender/age/religious/ethnic/sexual orientation groups.
5. **Cost-utility** means 'being economical' in commonsense terms, but it also covers costs and benefits analysis that includes context and environmental/personal/social capital gains and losses. This normally requires: (i) that the original evaluation *at least* included a careful **cost-feasibility** estimate (which of course is a barred dimension, i.e., cost-unfeasible is a deal breaker), and typically also (ii) *at least an estimate* of **comparative cost-effectiveness**, a property that should be maximized within the constraints of 1-4.

   NOTES: (i) The comparisons here should include at least competent judgmental identification and cost-effectiveness estimates of other ways of doing the original evaluation, from professional-judgment-only up through using something like the Program Evaluation Standards. (ii) Costs covered here must include: (a) the costs of disruption and reduction of work (amount and quality) by the process of evaluation itself, and the time spent reading or listening to or discussing the evaluation; (b) stress caused by the evaluation process, as a cost in itself, over and above its effects on job performance; (c) the usual direct costs in money, time, opportunities lost, space, etc. (iii) Benefits here must include (a) gains in efficiency or quality of work and savings in costs due to the evaluation's content or its occurrence; (b) gains in morale of staff from favorable reports; (c) gains in support from stakeholders e.g., donors, legislators, due to report as a demonstration of accountability; (d) the usual area-specific gains such as knowledge gains, improved decision-making, improved program quality, resource conservation, etc. (iii) The complexity of the prior notes must not detract from the core issue here, which is: did the evaluation pay for itself (or show a profit to the client), or did it merely discharge an obligation (legal or ethical); *or* was it, de facto, an unnecessarily expensive gesture?
6. **Generalizability** is not a *requirement*—it's not a *necessary or defining* criterion of merit, so there is no 'bar' that has to be cleared on this dimension—but it is a bonus-earning factor, so evaluation designers and practitioners should try to score on this.

> This has three facets of particular importance: (i) utility/merit of this evaluation design (and procedures for its use) in the evaluation of this evaluand at other times, or when using other evaluation staff, etc. (a.k.a. **reuseability**); (ii) utility/merit of the particular evaluation design and/or implementation procedures *or the conclusions from their use* in the evaluation *or estimation of the results from evaluating* other programs (a.k.a. **exportability**); (iii) **robustness** i.e., the extent to which the evaluation results are immune to a degree of changes in program context or program variations of the usual relatively minor kind (e.g., fatigue of evaluation or program staff characteristics, variations in recipient personality or environmental e.g., seasonal variations), *or* minor errors in data values or inferential processes. If sustainability really meant 'resilience to risk'—which is sometimes proposed as its definition—this would be close to **sustainability**. In any case, evaluators should always *try* for robust designs. (This consideration could be included under Validity, but is placed here because this is in a way the repository of variability considerations.)

The sum (actually the synthesis) of 1-6 provides an estimate of Merit or Value, the latter quality being distinguished by attention to costs (or the lack of need to consider them—e.g., by the Gates Foundation), vs. the usual need to consider them carefully (costs are covered in Cost-Utility above). Note that Merit and Value can reach the highest grades available with or without any significant rating on Generalizability—that's what's meant by saying Checkpoint 6 is only a 'bonus dimension' of merit/value.

In most practical contexts, Value is approximately the same as Utility (with slightly different emphases, depending on context). And Utility is the property that maximizes Utilization, insofar as the evaluator can control it, i.e., under the constraints of rationality and propriety on the part of the client and audiences. NOTE: a small editing job will take out of the MEC the references to research that is specifically evaluative, and the residual checklist will work quite well for evaluating many reports in applied science and technology.

Now a word about *reasons for doing meta-evaluation* by contrast with *how meta-evaluation is or should be done*, our topic so far.

The main reasons are what we might call the face-valid ones, that is, the main reasons for all evaluation: (i) for decision support, accountability, transparency (i.e., summative evaluation), (ii) improvement of the evaluand (i.e., formative evaluation[6]), and (iii) for the sake of the knowledge gained (which covers most historians' reason for much of their evaluative work), i.e., ascriptive evaluation. Some further, perhaps 'deeper' reasons are: (iv) practicing what you preach, as a marketing strategy, for you and for evaluation in general, (v) as a professional imperative (self-improvement), and (vi) as an ethical imperative.

Final note. As a matter of empirical fact, do you think that a quick scan through the MEC would sometimes lead you to think of things you've left out or underemphasized? (It has had that result for me on a current evaluation.) If so, that's another reason to try to get it improved, and to use it yourself, besides its use in evaluating evaluations by others. Send in your suggestions to me at mjscriv1@gmail.com with MEC in the subject line.

[6] In formative meta-evaluation, the meta-evaluator can't be around long, or s/he will become a co-author.

Acknowledgements. The first memo in this series was significantly improved in response to comments and suggestions from Leslie Cooksy and Michael Quinn Patton, although they did not suggest the doubling in length it took me to produce this effort at improvement, and may not be pleased with the result. After letting it sit for a year, and getting some good suggestions from Chris Coryn, I've made some further extensive changes and additions and now hope for more criticism.

[3,337 words @ 2011-08-16]

first version circulated March 3, 2010, the fourth revision was dated 3.13.11, this is 9th ed., 8.16.11