A. What are the criteria of merit for an evaluation in any field, including program evaluation? That’s probably the most important question to which a professional meta-evaluator—someone who is evaluating an evaluation—and perhaps even every evaluator, needs an answer in order to operate professionally, although the answer may be implicit in their practice rather than an explicit part of it. Moreover, it’s a very important question, 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 practical meta-evaluators, 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 may be able to do a little better. Here’s my latest effort, in the hope it will inspire suggestions and corrections.

And 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 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 (defining) 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 actual utilization is the responsibility of the client, and its absence can’t be blamed on the evaluator unless the evaluator is responsible for it by a weakness in the evaluation e.g., its lack of relevance to the client’s questions, or its lack of clarity, especially if the lack or abuse of utilization is due to suppression or careless misinterpretation or deliberate misuse of the evaluation by the client. Also, it seems important not to give any credit to an evaluation that is immediately utilized, although invalid; and not much to one that cost far more than was necessary. So I believe that utilization is an essential goal for a good evaluation, but not a defining feature, just as I believe that democracy is an essential goal for a worthwhile 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 quest for an answer to this question requires some care with defining the level of the inquiry. Within a subfield of evaluation, for example program evaluation, there are some good checklists of matters that have to be covered by evaluations, with some guidance as to how they should be covered. These include the Program Evaluation Standards, the GAO Yellow Book, and the KEC. 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 using one or more of these as setting the standards for the matters that must be covered by—and to some extent, how covered by—a good evaluation. But in order to evaluate these lists in their turn—e.g., in order to pick the best set of program evaluation criteria against which to measure designs for a particular assignment—we need, in the logic of evaluation, some more general standards to which these more specific standards should adhere. 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. With more sophistication about program evaluation, the three sets of standards-for-coverage mentioned will be more useful; but for someone with less of the area-specific knowledge, a higher-level checklist may make more sense. We’ll call it the Meta-evaluation Checklist (MEC) although it probably won’t be the only one for long.

NOTES: The term ‘evaluand’ is used to refer to whatever is being evaluated... The key concepts involved are 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 META-EVALUATION CHECKLIST (MEC)

1. Validity 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, and 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 of it should you be evaluating—an evaluation’s conclusions, or its process, or its impact, or all of these; (b) what about the function or role of the original evaluation, particularly whether it is/was supposed to be formative, summative, or ascriptive; (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,2 vs. grading vs. profiling, vs. scoring; (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, 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 generally means the extent to which the relevant scientific (or other disciplinary) standards were 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. 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

---

1 This amounts 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 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.

first version circulated March 3, 2010, the fourth revision was dated 3/13, this is 6th ed., 2.22.11
of confirmation or at least confirmability, for the evaluation as a whole, which 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 being: direct observation, recorded data, theoretical, logical, analogical, or judgmental sources. But in the case of evaluations, there is another element that also has to be examined for validity, namely the values component. Doing this means checking whether all relevant values were identified, and whether they were specified in the detail needed for this evaluation, scaled appropriately, and measured or estimated reliably, and finally integrated in a defensible way with the empirical findings in an inference to the appropriate evaluative conclusion.

(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 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. 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.

(There are three or four ways to strengthen it, so an improved version of it could well be satisfactory.)

(iv) Another major requirement in determining the validity of complex evaluands is comprehensive coverage of all the relevant considerations: in many subdivisions of evaluation, these include the identification and application of all relevant values as previously mentioned, plus the validity of non-evaluative descriptions of the evaluand as well as of evaluative conclusions (and descriptions), and in many subdivisions of evaluation, including policy analysis, product evaluation, and program evaluation, they include the search for side-effects, not just intended effects. If the sub-type of evaluation needed is ranking, then both cost analysis and comparisons are also essential. Usually, one needs a field-specific checklist to ensure CompCover: in program evaluation this could be one of those mentioned above.

2. Clarity (i.e., a combination of Comprehensibility to the client/audiences/stakeholders/staff; this is to be combined with Brevity, a factor that reduces effort and the frequency of errors of interpretation, and can improve acceptance); the PES and the KEC and the Yellow Book are all subject to some criticism on this account, and it may be the leading factor in determining their relative merit.)

---


first version circulated March 3, 2010, the fourth revision was dated 3/13, this is 6th ed. 2.22.11
3. **Credibility** (to client/audiences/stakeholders/staff, meaning a combination of low level of Apparent Bias e.g., from COI (conflict of interest) with Expertise.) The focus here is on matters of credibility not covered by direct validity considerations.

4. **Propriety** meaning ethicity, 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) at least 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 to the extent possible 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 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.

6. **Generalizability** is not a requirement—it’s not a 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 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 changes in program context or program variations of the usual relatively minor kind (e.g., fatigue of evaluation or program staff, 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 second being distinguished by attention to costs (or the lack of need to consider them—e.g., by the first version circulated March 3, 2010, the fourth revision was dated 3/13, this is 6th ed., 2.22.11
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 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 most projects in applied science and a good many in technology.

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, besides its use in evaluating evaluations by others.

B. Now a word about reasons for doing meta-evaluation by contrast with how meta-evaluation is or should be done.

The main reasons are what we might call the face-valid ones, that are the main reasons for all evaluation: (i) decision support or accountability (i.e., summative evaluation), (ii) improvement (i.e., formative evaluation), and (iii) for the sake of the knowledge gained (which covers most historians' reason for most of their work), i.e., ascriptive evaluation. Some further or 'deep 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.

Acknowledgements. The first version in this series was significantly improved in response to comments and suggestions from Leslie Cooksey and Michael Quinn Patton (and much reconsideration by me), 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.

[2,693 words]