July 24th, 2013

KEY EVALUATION CHECKLIST (KEC)

Michael Scriven
Claremont Graduate University
& The Evaluation Center, Western Michigan University

• For use in the professional designing, managing, and monitoring or evaluating of:
proposals (or plans), projects, programs, processes, and ***policies***;
• for assessing the evaluability of these;
• for requesting *proposals* (i.e., writing *RFPs*) to do or evaluate them;
& for evaluating proposed, ongoing, or completed ***evaluations*** of them.[1]

INTRODUCTION

Why a checklist? Because professional evaluation is typically a complex business, and documented experience in those complex fields that have taken the trouble to investigate the use of valid checklists—most notably engineering and medicine—has demonstrated that using them produces huge benefits, probably more than any of the famous discoveries in science and technology.[2] There are general-purpose evaluation checklists, most notably the excellent one for program evaluation that has been elevated to the very distinguished status of ANSI (American National Standards Institute) recognition:[3] it is well worth checking out for comparative or complementary utility against this one. Of the detailed ones, this one is the oldest and the most recently and frequently revised, and more adaptable to fields beyond program evaluation—in fact to almost the whole domain of applied science as well as the disciplined humanities and arts such as classical dance and calligraphy (see details in General Note 1 below). It is also more easily used in ultra-short form, using the headings of Parts A through D, as an 18-line checklist; see General Note 2 below. (Some people think the version you're reading now is too long to be a checklist). But all these thoughts are of course just an author's views… If what you're really looking for is a how-to-do-it guide to doing an evaluation, you'll find that is covered in detail in this document, and summarized in a Note at the very end of it, under the heading ***General Note 8***.

This Introduction section now takes the form of a number of 'General Notes,' a few more of which will be found in the body of the document, along with many checkpoint-specific Notes… Punctuation note: the ellipsis (three periods in a row) is here used not only for missing letters or words but to signify a break in theme that is one notch below a paragraph break and one above a period. The context clearly disambiguates the two uses.

General Note 1: APPLICABILITY The KEC can be used, with care, for evaluating more than

---

[1] That is, for what is called meta-evaluation, i.e., the evaluation of one or more evaluations. (See D5 below.)

[2] See the excellent discussion under 'Medical Errors' in Wikipedia, and Atul Gawande's *The Checklist Manifesto,* or the New Yorker article on which the latter is based: http://www.newyorker.com/reporting/2007/12/10/071210fa_fact_gawande

[3] *Program Evaluation Standards* (Sage, 2010 edition).

the half dozen evaluands[4] listed in the sub-heading at the top of this page, just as it can be used, with considerable care, by others besides professional evaluators. For example, it can be used for most of the task of: (i) the evaluation of products;[5] (ii) the evaluation of organizations and organizational units[6] such as departments, research centers, consultancies, associations, companies, and for that matter, (iii) hotels, restaurants, and for that matter mobile food carts; (iv) services, which can be treated as if they were aspects or constituents of programs, i.e., as processes (covered below under C1); (v) many processes, policies, practices, or procedures, which are often implicit programs (e.g., "Our practice at this school is to provide guards for children walking home after dark"), hence evaluable using the KEC; for (vi) habitual or peak patterns of behaviour i.e., performances (as in "In my practice as a consulting engineer, I often assist designers, not just manufacturers"), which is, strictly speaking, a slightly different subdivision of evaluation; and, (vii) with some use of the imagination and a heavy emphasis on the ethical values involved, some tasks or major parts of tasks in the evaluation of personnel. Still, it is often worthwhile to develop somewhat more specific checklists for evaluands that are not *exactly* programs, or are specialized *types* of program: an example is organized training, for which a specialized checklist from this author is available at michaelscriven.info.

So, in this edition (about its 90th) the KEC is a kind of ~50-page/30,000 word mini-textbook or reference work for a wide range of professionals working in evaluation or management—with all the limitations of that size (it's too long for some needs and too short for others), and surely more that I hope you will point out. It is written at an intermediate level of professional analysis: many professionally done evaluations make mistakes that would be avoided by someone taking account of the points covered here, but there are also many sophisticated techniques, sometimes crucial for professional evaluators in a particular subfield, that are not covered here, notably including statistical and experimental design techniques that are not unique to evaluation, and cost-analytic techniques from the audit field.

General Note 2: TABLE OF CONTENTS

---

[4] 'Evaluand' is the term used to refer to whatever is being evaluated. Note that what counts as a program is often also called an initiative or intervention or project; sometimes even an approach or strategy, although the latter are perhaps better thought of as *types* of program.

[5] For which it was originally designed and used, c. 1971—although it has since been completely rewritten for its present purposes, and then revised or rewritten (and circulated or re-posted) more than 85 times. The latest version can currently be found at michaelscriven.info and can be identified by examining the date in the heading, running footers, and word count. It is an example of 'continuous interactive publication' a type of project with some new significance in the field of knowledge development, although (understandably) a source of irritation to some librarians and bibliographers and of course publishers. It enables the author, like a landscape designer and unlike a traditional painter, architect, or composer, to steadily improve his or her specific individual creations over the years or decades, with the help of user input. It is simply a technologically-enabled extension towards the limit of the stepwise process of producing successive editions in traditional publishing, and arguably a substantial improvement, in the cases where it's appropriate.

[6] There is of course a large literature on the evaluation of organizations, from Baldrige to Senge, and some of it will be useful for a serious evaluator, but much of it is confused and confusing (e.g., about the differences and links between evaluation and explanation, needs and markets, criteria and indicators, goals and duties)—and too often omits ethics.

PART A: PRELIMINARIES: A1, Executive Summary; A2, Clarifications; A3, Design and Methods.

PART B: FOUNDATIONS: B1, Background and Context; B2, Descriptions & Definitions; B3, Consumers (Impactees); B4, Resources ('Strengths Assessment'); B5, Values.

PART C: SUBEVALUATIONS: C1, Process: C2, Outcomes; C3, Costs; C4, Comparisons; C5, Generalizability.

PART D: CONCLUSIONS & IMPLICATIONS: D1, Synthesis; D2, Recommendations, Explanations, Predictions, & Redesigns; D3, Responsibility and Justification; D4, Report & Support; D5, Meta-evaluation.[7]

General Note 3: TERMINOLOGY Throughout this document, "*evaluation*" is taken to refer to the process of determination of merit, worth, or significance (abbreviated m/w/s)[8]; "*an evaluation*" is taken to refer to a declaration of value, possibly but not only as the result of such a process; and "*evaluand*" to mean whatever is being evaluated… "*Dimensions of merit"* (a.k.a., *"criteria of merit*") are the characteristics of the evaluand that definitionally bear on its m/w/s (i.e., could be used in explaining what 'good X' means), and "*indicators of merit*" (the status of many characteristics is borderline) refers to factors that are empirically but not definitionally linked to the evaluand's m/w/s… *Professional evaluation* is simply evaluation requiring specialized tools or skills that are not in the everyday repertoire; it is usually systematic (and inferential), but may also be merely judgmental, even perceptual, if the judgment skill is professionally trained and maintained, or is a (recently) tested advanced skill (think of livestock judges, football referees, saw controllers in a sawmill)… The KEC is a tool for use in systematic professional evaluation, so knowledge of some terms from evaluation vocabulary is assumed, e.g., formative, goal-free, ranking; their definitions can be found in my *Evaluation Thesaurus* (4e, Sage, 1991), or in the *Evaluation Glossary*, online at evaluation.wmich.edu. However, every conscientious program manager (or designer or fixer) does evaluation of their own projects, and will benefit from using this, skipping the occasional technical details… The most common reasons for doing evaluation are (i) to identify needed *improvements to* the evaluand (*formative* evaluation); (ii) to support *decisions about* the program, including deciding whether it was a proper use of the funds employed (*summative* evaluation[9]); and (iii) simply to enlarge or refine our body of evaluative knowledge (*ascriptive* evaluation, as in 'best practices' and 'lessons learned' studies, and almost all evaluations by historians).[10] Keep in mind that an evaluation may serve more

---

[7] It's not important, but you can remember the part titles from this mnemonic: A for Approach, B for Before, C for Core (or Center), and D for Dependencies. Since these have 3+5+5+5 components, it's an 18-point checklist.

[8] In the most abstract terms, 'evaluation' refers to identifying what's good and bad, right and wrong, about something; but in order to develop an applied discipline, it's more useful to take one step towards specificity and identify merit, worth, and significance as the macro-dimensions of goodness etc. that are of interest.

[9] Major decisions about the program include: refunding, defunding, exporting, replicating, developing further, and deciding whether it represents a proper or optimal use of funds (i.e., evaluation for accountability, the same perspective as an audit (although audits usually only concern money)).

[10] It is possible that we should add preformative evaluation to this list, i.e., evaluation of the precursor effects of a program, its design, and its evaluability (see jmde.com for June, 2012). But keeping

than one purpose, or shift from one to the other as time passes or the context changes… Merely for simplification, we talk throughout this document about the evaluation of 'programs' rather than 'programs, plans, or policies, or evaluations of them, etc…' as detailed in the sub-heading above.

General Note 4*: *TYPE OF CHECKLIST This is an *iterative* checklist, not a *one-shot* or *knockdown* checklist, i.e., you should expect to work through it several times when dealing with a single project, even for design purposes, since discoveries or problems that come up under later checkpoints will often require modification of what was done or entered under earlier ones (and no rearrangement of the order will completely avoid this).[11] For more on the nature of checklists, and their use in evaluation, see the author's paper on that topic, and a number of other papers about, and examples of, checklists for evaluation by various authors, under the listing for the Checklist Project at evaluation.wmich.edu.

General Note 5: EXPLANATION & JUSTIFICATIONS Since it is not entirely helpful to simply list here what (allegedly) needs to be covered in an evaluation when the reasons for the recommended coverage (or exclusions) are not obvious—especially when the issues are highly controversial (e.g., Checkpoint D2)—brief summaries of the reasons for the position taken are also provided in such cases.

General Note 6: CHECKPOINT FOCUS The determination of merit, or worth, or significance (a.k.a. (respectively) quality, value, or importance), the triumvirate value foci of most evaluations, each rely to different degrees on slightly different slices of the KEC, as well as on a good deal of it as common ground. These differences are marked by a comment on these distinctive elements with the relevant term of the three underlined in the comment, e.g., worth, unlike merit (or quality, as the terms are commonly used), brings in Cost (Checkpoint C3).

General Note 7: THE COST AND COMPLEXITY OF EVALUATION: IS THERE A SHORT FORM OF THE KEC? The KEC is a list of what *ideally* should be covered in an evaluation, but in the real world, the budget and the timeline for an evaluation are often not enough to cover the whole list thoroughly. People sometimes ask what checkpoints could be skipped when one has a very small evaluation budget. The answer is, "None, *but*…", i.e., none should be skipped *completely*, but only *a very light* level of coverage of each of the checkpoints is (virtually always) a necessary condition for validity. More precisely, (i) sometimes the client, including you if you are the client (often true in ascriptive evaluation), can show that one or two are not relevant to the information need *in a particular context* (e.g., cost may not be important in some cases); (ii) the fact that you shouldn't skip any checkpoints doesn't mean you have to spend *substantial time or money* on each of them. What you *do* have to do is *think through the implications of each checkpoint* for the case in hand, and consider whether an economical way of coping with it—e.g., by relying on current literature for the needs assessment required in most evaluations—would *probably* be adequate for an *acceptably probable* conclusion. In other words, focus on robustness (see Checkpoint D5, Meta-evaluation, below). Or you may have to rely on a subject-matter expert for an estimate based on his/her experience about one or more preferably minor checkpoints in a half-day

---

things simple is a big advantage, so just keep in mind that it's not a contradiction to say that formative evaluation can occur (or refer to a time/period) before the evaluand exists.

[11] An important category of these is identified in Note C2.5 below

of consulting; or on a few hours of literature search by you on the relevant facts about e.g., resources, or critical competitors[12]. That's sometimes all that this client and the audiences involved want and need. But reality sometimes means that a professionally adequate evaluation simply cannot be done;[13] that's the cost of integrity for evaluators and, sometimes, excessive parsimony for clients… Don't forget that honesty on this point can prevent some bad situations later—or maybe should lead to a change of budget, up or down, that you should be considering before you take the job on… A common question about the cost of evaluation asks what percentage of program costs should be spent on evaluation. There is no possible answer to this question: it's underspecified. It's like asking how much you should spend on clothes during your life. One can say that for very large programs, less than 1% is sometimes more than enough; on very small programs, 20% will sometimes not be enough; but even these figures are misleading without a discussion of the type of cost. In terms of net long-term cost (i.e., treating evaluation as an investment) good evaluation will often pay for itself in cost-savings and improvement of quality, quite often in a year or less. In terms of this year's payables, the cost depends on the client's requirements for *robustness* (especially with respect to specific criticisms the client expects to encounter) and the level of *detail* needed, and on the geographic distribution of the evaluand, the need for interpreters in data-gathering and report translation, the need for new tests and approaches, the need for an emergency fund, and several other factors (many of which enter into analogous issues like the total cost of the insurance package for this organization, if you're looking for analogies to explain the absence of a simple formula). The only good answer is a reasonably detailed and carefully justified budget. See Note A3.2 and the Costs checkpoint below for more details.

PART A: PRELIMINARIES

These preliminary checkpoints are clearly essential parts of an evaluation *report*, but may seem to have no relevance to the design and execution phases of the *evaluation itself*. That's why they are segregated from the rest of the KEC checklist: however, it turns out to be quite useful to begin all one's thinking about an evaluation by role-playing the situation when you will come to write a report on it. Amongst other benefits, it makes you realize the importance of: describing context; of settling on a level of technical terminology and presupposition; of clearly identifying the most notable conclusions; and of starting a log on the project as well as its evaluation as soon as the latter becomes a possibility. Similarly, it's good practice to make explicit at an early stage the clarification steps you take, and the methodology array and its justification

A1. Executive Summary

The most important element in this section is an overview that is usually thought of as a kind of postscript: it's a summary of the results, and not (or not just) the investigatory process. We put this section up front in the KEC because you need to do some thinking about it from the very beginning, and may need to talk to the client—or prospective readers—

[12] A critical competitor is an entity that looks as if it might be better, overall, than the evaluand. More about these in C4 below.

[13] Use the *Program Evaluation Standards, 2e,* as the basis for deciding.

about it early on. Doing that is a way of forcing you and the client to agree about what you're trying to do; more on this below. Typically the executive summary should be provided without even mentioning the process whereby you got the results, unless the methodology is especially notable. In other words, take care to avoid the pernicious practice of using the executive summary as a 'teaser' that only describes what you looked at or how you looked at it, instead of what you found. Throughout the whole process of designing or doing an evaluation, keep asking yourself what the overall summary is going to say, based on what you have learned so far, and how directly and adequately it relates to the client's and stakeholders' and (probable future) audiences' information and other needs[14], given their pre-existing information; this helps you to focus on what still needs to be done in order to find out what matters most. The executive summary should usually be a selective summary of Parts B and C, and should not run more than one or at most two pages if you expect it to be read by executives. Only rarely is the occasional practice of two summaries (e.g., a one-pager and a ten- pager) worth the trouble, but discuss this option with the client if in doubt—and the earlier the better. The summary should also (usually) convey some sense of the *strength* of the conclusions—which combines an estimate of both the weight of the evidence for the premises and the robustness of the inference(s) to the conclusion(s), more details in D5—and any *other notable limitations* of the study (see A3 below). Of course, the final version of the executive summary will be written near the end of writing the report, but it's worth trying the practice of re-editing an informal draft of it every couple of weeks during a major evaluation because this forces one to keep thinking about identification and substantiation of the most important conclusions. Append these versions to the log, for future consideration.

Note A1.1 This Note should be just for beginners, but experience has demonstrated that others can also benefit from its advice: *the executive summary is a summary of the evaluation results not of the program's characteristics.* (Checkpoint B2 is reserved for the latter.)

A2. Clarifications

Now is the time to clearly *identify and define* in your notes, for assertion in the final report—and resolution of ambiguities along the way—the answers to some key questions, such as: (i) exactly who the *client* is (a.k.a. 'evaluation commissioner'), if there is one besides you[15]: this is the person, group, or committee who officially requests, and, if it's a paid evaluation, pays for (or authorizes payment for) the evaluation, and—you hope—the same entity to whom you first report (if not, try to arrange this, to avoid crossed wires in communications). (ii) Who are the prospective (i.e., overt) *audiences* (for the report)? This cannot be answered by simply asking the client who they want it to go to. There may also be audiences who have a right to see it e.g., because of Freedom of Information or Rights of Human Subjects legislation. (iii) Who are the *stakeholders* in the program (those who have or will have a substantial vested interest—not just an intellectual interest—in the outcome of the evaluation, and may have important information or views about the program and its

[14] "Other" may include needs for reassurance, insight, empathy or sympathy, justice, etc.

[15] People whose business is evaluation nearly always have a client in the usual sense; but the search for truth (and perhaps the hope of fame or improved help for the needy) is sometimes what drives researchers to do evaluations, as for any other research, and for convenience these cases (called ascriptive evaluations) are treated as having the investigator as client.

situation/history). They are usually a subset of the audiences. And there be others who (probably) will see, or should see: (a) the results, and/or (b) the raw data—these are the covert audiences. (iv) Get clear in your mind, and with your client, your actual role or roles—internal evaluator, external evaluator, a hybrid (e.g., an outsider on the payroll for a limited time to help the staff with setting up and running evaluation processes), an evaluation trainer (sometimes described as an empowerment evaluator), a repairer/'fixit guy,' redesigner, visionary (or re-visionary), etc. Each of these roles has different risks and responsibilities, and is viewed with different expectations by your staff and colleagues, the clients, the staff of the program being evaluated, et al. You may also pick up some other roles along the way—e.g., counsellor, therapist, mediator, decision-maker, inventor, redescriber, advocate—sometimes as a role you'll play for everyone but sometimes for only part of the staff/stakeholders/others involved. It's good to formulate and sometimes to clarify these roles, at least for your own thinking (especially watching for possible conflicts of role), in the project log. The project log is absolutely essential; and it's worth considering making a standard practice of having someone else read it regularly and initial entries in it that may at some stage become very important.

And (v) most importantly, now is the time to pin down the question(s) you're trying to answer and the kind of answer you're expected to provide. This means getting down to the nature and details of the job or jobs, as the client sees them—and encouraging the client (who may be you) to clarify their position on the details that they have not yet thought out. Note that some or all of the questions that come out of this process are often *not* evaluative questions, but 'questions of fact;' this doesn't mean you should dismiss them, but simply identify them as such. The big problem arises when the client has an essentially evaluative question but thinks it can be answered by a merely factual inquiry; this issue must be addressed, if not immediately then before heavy commitment by either party. This fifth process may require answering some related questions of possibly less critical importance but nevertheless important, e.g., what's going to count as proof, for this client or these audiences—if you're going to evaluate a teacher training program, will it be enough to quote results from the research literature to support the procedure used, or will you have to replicate those studies with the particular teachers in this study; and if you have to do that, will the changes in those teachers count as proof of success, or will you have to test the future students of those teachers for improved learning by contrast with previous cohorts? Getting tougher about what counts as evidence of success can translate into doubling the time frame or more; and quadrupling the cost, or more, so it's not a mere academic quibble. Other possibly important questions include: can you determine the nature and source of the request, need, or interest, leading to the evaluation. For example, is the request, or the need, for an evaluation of plain merit, or of *worth*—which usually involves really serious attention to comparative cost analysis—rather than just of merit; or of *significance* which always requires advanced knowledge of the research (or other current work) scene in the evaluand's field; or of more than one of these? Is the evaluation to be formative, summative, or ascriptive[16]; or for more than one of these purposes? (If formative or summative, make

[16] Formative evaluations, as mentioned earlier, are usually done to find areas needing improvement of the evaluation: summative are mainly done to support a decision about the disposition of the evaluand (e.g., to refund, defund, or replicate it); and 'ascriptive' evaluations are done simply for the record, for history, for benefit to the discipline, or just for interest i.e., for the sake of the knowledge gained.

clear that this normally means both an analytic and a holistic assessment.) Exactly what are you supposed to be evaluating (the evaluand alone, or also the context and/or the infrastructure?): how much of the context is to be taken as fixed; do they just want an evaluation in general terms, or if they want details, what counts as a detail (enough to *replicate* the program elsewhere, or just enough to *recognize* it anywhere, or just enough for prospective readers to know what you're *referring* to); are you supposed to be simply evaluating the effects of the program as a whole (holistic evaluation); or the dimensions of its success and failure (one type of analytic evaluation); or the quality on each of those dimensions, or the quantitative contribution of each of its components to its overall m/w/s (another two types of analytic evaluation); are you required to rank the evaluand against other actual or possible programs (which ones?), or only to grade it;[17] and to what extent is a conclusion that involves generalization from this context being requested or required (e.g., where are they thinking of exporting it?)… And, of particular importance, is the main thrust to be on ex post facto (historical) evaluation, or ex ante (predictive) evaluation, or (the most common, but don't assume it) both? (Note that predictive program evaluation comes very close to covering (almost all varieties of) policy analysis.)… Are you also being asked (or expected) either to evaluate the client's theory of how the evaluand's components work, or to create/improve such a 'program theory'—keeping in mind that the latter is something over and above the literal evaluation of the program, and especially keeping in mind that this is sometimes impossible for even the most expert of field experts in the present state of subject-matter knowledge? [18]… Another issue: is the required conclusion simply to provide and justify grades, ranks, scores, profiles, or (a different level of difficulty altogether) an optimal distribution of funding?... Are recommendations (for improvement or disposition), or identifications of human fault, or predictions, requested, or expected, or feasible (another level of difficulty, too—see Checkpoint D2)?... Is the client really willing and anxious to learn from faults or is this just conventional rhetoric? Your contract or, for an internal evaluator, your job, may depend on getting the answer to this question right, so you might consider trying this test: ask them to explain how they would handle the discovery of extremely serious flaws in the program—you will often get an idea from their reaction to this question whether they have 'the right stuff' to be a good client. Or you may discover that you are really expected to produce a justification for the program in order to save someone's neck; and that they have no interest in hearing about faults… And, have they thought about post-

[17] Grading refers not only to the usual academic letter grades (A-F, Satisfactory/Unsatisfactory, etc.) but to any allocation to a category of merit, worth, or significance, e.g., grading of meat, grading of ideas and thinkers.

[18] Essentially, this is a request for decisive non-evaluative explanatory (probably causal) research on the evaluand and/or context. You may or may not have the skills for this, depending on the exact problem; these are advanced topic-specific skills that you certainly didn't acquire in the course of your evaluation training. It's one thing to determine whether (and to what degree) a particular program reduces delinquency: any good evaluator can do that (given the budget and time required). It's another thing altogether to be able to explain why that program does or does not work—that often requires an adequate theory of delinquency, which so far doesn't exist. Although 'program theory' enthusiasts think their obligations always include or require such a theory, the standards for acceptance of any of these theories by the field as a whole are often beyond their reach; and you risk lowering the value of the evaluation field if you claim your evaluation depends on providing such a theory, since in many of the most important areas, you will not be able to do so.

report help with interpretation and utilization? (If not, offer it without extra charge—see Checkpoint D2 below.)

Try to get all of (i) to (v) into a written contract if possible (essential if you're an external evaluator, highly desirable for an internal one.) And get it cleared by legal counsel, certainly that of your employer if there is one, and probably also one that is devoted exclusively to your own interest, since your employer's counsel is primarily interested in saving its skin from public harm. If you're using language from your client's attorneys, look out for any non-disclosure agreement (NDA) that may prevent you from replying in public even if the client publicly misrepresents your findings… It's best to complete the discussion of these issues about what's expected and/or feasible to evaluate, and clarify your commitment (and your cost estimate, if it's not already fixed) only after doing a quick pass through the KEC, so ask for a little time to do this, overnight or 24 hours at least (see Note D2.3 near the end of the KEC)… Be sure to note later any subsequently negotiated, or imposed, changes in any of the preceding in the project log and get them signed off if possible… And it's good practice to do the same 'running log' approach for acknowledgments/thanks/etc., so it will also be almost completed by the time you come to write the final report.

A3. Design and Methods

Now that you've got the questions straight—as far as you can at this stage—how are you going to find the answers? You need a plan that lays out the aspects of the evaluand you have to investigate in order to evaluate it (including relevant values), and a set of investigative procedures to implement each aspect of your plan—i.e., the design of the evaluation—based on some general account of how to investigate these aspects of the design; in other words a methodology. To a considerable extent, the methodologies now used in evaluation originate in social science methodology, and are well covered elsewhere, in both social science and evaluation texts. In this section, we just list a few entry points for *that* slice of evaluation methodology, and provide rather more details about the *evaluative slice* of evaluation methodology, the neglected part, not covered in the social science texts. This part is introduced with a few comments here, and then mostly covered, or at least dealt with in more detail, under the later checkpoints that refer to the key evaluative aspects of the investigation—the Values, Process, Outcomes, Costs, Comparisons, Generalizability, and Synthesis checkpoints. Leaving out this slice of the methodology of evaluation is roughly the same as leaving out any discussion of inferential statistics from a discussion of statistics.

Three orienting points to start with: (i) Program evaluation is usually about a single program rather than a set of programs. Although program evaluation is not as individualistic—the technical term is *idiographic* rather than *nomothetic*—as dentistry, forensic pathology, or motorcycle maintenance (e.g., since most programs have large numbers of impactees that are treated similarly or simultaneously rather than just one), it is more individualistic than most social sciences, even applied social sciences. So you'll often need to be knowledgeable about case study methodology[19]. (ii) Program evaluation is nearly always a very complex task, involving the investigation of a number of different aspects of program performance—even a number of different aspects of a single element in that such as impact or cost—which means it is part of the realm of study that requires extensive use of *checklists*. The humble checklist has been ignored in most of the literature on research methods, but

[19] This means mastering at least some books by Yin, Stake, and Brinkerhoff (check Amazon).

turns out to be more complex and also far more important in the field than was generally realized, so look up the online Checklists Project at http://www.wmich.edu/evalctr/checklists for some papers about the methodology of checklists and a long list of specific checklists composed by evaluators (including a considerably earlier version of this one). You can find one or two others, and the latest version of this one, at michaelscriven.info which may or may not be the same as this one; (iii) There are obviously many techniques from social science methodology that you need to have in your repertoire, along with the ubiquitous methodology of checklists, but there is also a set of *evaluation-specific* ones that you must master. These include some devoted to the handling of values (identification of the ones relevant to the particular evaluand and its context, and their definition, validation, and measurement);x and their integration with the empirical data you locate.

Now for some entry points for applying social science methodology, and some examples of the kind of question that you may need to answer. Do you have adequate domain expertise (a.k.a. subject-matter, and/or local context knowledge) for what you have now identified as the real tasks? If not, how will you add it to the evaluation team (via consultant(s), advisory panel, full team membership, sub-contract, or surveys/interviews)? More generally, this is the time to identify, as soon as possible, all investigative procedures for which you'll need expertise, time, equipment, and staff—and perhaps training—in this evaluation: skilled techniques like process observation, participant observation, logging, journaling, audio/-photo/video recording, testing, simulating, role-playing, surveys, interviews, statistics,[20] experimental design, focus groups, text analysis, library/online searches/search engines, etc.; and data-analytic procedures (stats, cost-analysis, modelling, topical-expert consulting, etc.), plus reporting techniques (text, stories, plays, graphics, freestyle drawings, stills, movies, etc.), and their justification. You probably need to allocate time for a lit review on some of these methods.... In particular, on the difficult causation component of the methodology (establishing that certain claimed or discovered phenomena are the effects of the interventions), can you use and afford separate control or comparison groups to determine causation of supposed effects/outcomes? If not, look at interrupted time series designs, or the GEM (General Elimination Methodology[21]) approach, and some ideas in case study design… If there is to be a control or quasi-control (i.e., comparison) group, can you and should you try to randomly allocate subjects to it (and can you get through IRB (the Institutional Review Board))?... How will you control differential attrition; cross-group contamination; other threats to internal validity? If you can't control these, what's the decision-rule for declining/aborting the study?... Can you double- or single-blind the study (or triple-blind if you're very lucky)?... If the job requires you to determine the separate contribution to the effects from individual components of the evaluand—how will you do that?... If a sample is to be used at any point, how selected, and if stratified, how stratified?… Will/should the evaluation be goal-based or goal-free, or (the ideal) a hybrid?[22]... To what extent participatory or collaborative; if to a considerable extent, what standards and choi-

[20] For serious statistics, keep in mind that you can't use Excel's statistics, you must use specialist software like SPSS. (There are several online articles about major errors in Excel statistics.)

[21] See "A Summative Evaluation of RCT methodology; and an alternative approach to causal research" in *Journal of Multidisciplinary Evaluation* vol. 5, no. 9, March 2008, at jmde.com.

[22] That is, at least partly done by evaluators who are not informed of the goals of the program.

ces will you use, and justify, for selecting partners/assistants? In considering your decision on that, keep in mind that participatory approaches improve implementation (and sometimes validity), but may cost you credibility (and possibly validity). How will you handle that threat?... If judges are to be involved at any point, what reliability and bias controls will you need (again, for credibility as well as validity)?... How will you search for side-effects and side-impacts, an essential element in almost all evaluations (see Checkpoint C2)?

Most important of all, with respect to all (significantly) relevant values how are you going to go through the value-side steps in the evaluation process, i.e., (i) identify, (ii) particularize, (iii) validate, (iv) measure, (v) set standards ('cutting scores') on each value scale, (vi) set weights for them, and then (vii) incorporate (synthesize, integrate) the value-side with the empirical data-gathering side in order to generate the evaluative conclusion?... Now check the suggestions about values-specific methodology in the Values checkpoint, especially the comment on pattern-searching… When you can handle all this, you are in a position to set out the 'logic of the evaluation,' i.e., a general description and justification of the total design for this project, something that—at least in outline—is a critical part of the report, under the heading of Methodology.

Note A3.1: The above process will also generate a list of needed resources for your planning and budgeting efforts—i.e., the money (and other costs) estimate. And it will also provide the basis for the crucial statement of the limitations of the evaluation that may need to be reiterated in the conclusion and perhaps in the executive summary.

Note A3.2: Evaluation cost guidelines. Legal advisers recommend that consultants make no comments about this in published materials, to avoid charges of price-fixing. But it's clear that they are severely biased in giving this kind of advice, since they are vulnerable to suit if they don't warn us to keep off this topic, and safe if they do. So we must act in terms of what our professional and ethical obligations are first, and in terms of what caution is appropriate second. In particular, we must be able to discuss whether or not any formula makes sense in costing evaluation: for example, can we at least say that a competent evaluation typically or commonly costs N% of a program's budget? I think we can reject this possibility, for the reasons mentioned earlier. No percentage formula has any general validity, and it's borderline unprofessional to suggest the contrary. Why? Because it's easy to immediately give counter-examples to any suggestions. For example, surely one might suggest that evaluation should be allocated at least 1% of a program's budget? Definitely not; very large programs (e.g., those—there are many of them—with multi-million dollar annual budgets) that have long ago worked out that their task is extremely closely defined and repetitive, and have paid for all the test and design development costs, and have computerized data collection, with those computer costs already amortized, can easily get below 1%, especially after the start-up years. In the other direction, it's obvious that a highly innovative small program of great complexity, working in a highly unstable and fast-changing environment (two different things), may need to develop novel methodology and test designs with the attendant field trial costs for both of these, the total of these easily surpassing their projected annual costs, i.e., more than 100%. Conclusion: Clarify the specific evaluation request, and explain the costs that would be involved in getting the job done, and ask for their feedback. i.e., justify your proposal.

Note A3.3: By now you've covered a good many sheets of paper with notes and calculations. If you have not already started to use a program management 'tool' (i.e., software),

this is the time to do so, when you've got to produce a guide to managing the whole operation. The fancy versions of these are for engineers in charge of building a skyscraper or dam, and they have a very steep learning curve, but there are plenty that will harness your complexities without costing a day's pay. (Be sure to read the negative reviews of whatever you're considering before you buy it.) You can also try to do the job using a top end idea processor, or just lists and diagrams. But the software automatically prevents some common mistakes that arise in situations like this, where you have to juggle not only tasks but time, money, and equipment. Perhaps the most common mistake for those not using software is assigning someone or some equipment to two different tasks at the same time (called 'Did Parallel, Needed Serial').

PART B: FOUNDATIONS

This is the set of investigations that establishes the context and nature of the program, and some of the empirical components of the evaluation that you'll need in order to start specific work on the key dimensions of m/w/s in Part C. That is, they specify important elements that will end up in the actual evaluation report, by contrast with preparing for it, i.e., Part A, as well as providing foundations for the core elements in C. These are all part of the content of the KEC, and each one is numbered for that purpose.

B1. Background and Context

Identify historical, recent, concurrent, and projected settings for the program; start a list of contextual factors that may be relevant to success/failure of the program; and put matched labels (or metadata tags) on any that look as if they may interact. In particular, identify: (i) any 'upstream stakeholders'—and their stakes—other than the clients (i.e., identify people or groups or organizations that assisted in creation or implementation or support of the program or its evaluation, e.g., with funding or advice or housing or equipment or help); (ii) any enabling legislation/mission statements, etc.—and any other relevant legislation/policies—and log any legislative/executive/practice or attitude changes that occur after start-up; (iii) the underlying rationale, including the official program theory, and political logic (if either exist or can be reliably inferred; although neither are necessary for getting an evaluative conclusion, they are sometimes useful/required); (iv) general results of a literature review on similar interventions, including 'fugitive studies' (those not published in standard media), and on the Internet (consider checking the 'invisible web,' and the latest group and individual blogs/wikis with the specialized search engines needed to access these); (v) previous evaluations, if any; (vi) their impact, if any.

B2. Descriptions & Definitions

What you are going to evaluate is officially a certain program, but actually it's the total intervention made in the name of the program. That will usually include much more than just the program, e.g., it may include the personalities of the field staff and support agencies, their modus operandi in dealing with local communities, their modes of dress and transportation, etc. So, record any official descriptions of the program, its components, its context/environment, and the client's program logic, but don't assume they are correct, even as descriptions of the actual program delivered, let alone of the total intervention. Be sure to develop a correct and complete description of the first three (not the fourth), which

may be very different from the client's version, in enough detail to recognize the evaluand in any situation you observe, and perhaps—depending on the purpose of the evaluation—to replicate it. You don't need to develop the *correct* program logic, only the *supposed* program logic, unless you have undertaken to do so and have the resources to add this—often major, and sometimes suicidal—requirement to the basic evaluation tasks. Of course, you will sometimes see, or find later, some obvious flaws in the client's effort at a program logic and you may be able to point those out, diplomatically, at some appropriate time. Get a detailed description of goals/mileposts for the program (if not operating in goal-free mode). Explain the meaning of any 'technical terms,' i.e., those that will not be in the prospective audiences' vocabulary, e.g., 'hands-on' or 'inquiry-based' science teaching, 'care-provider.' Note significant patterns/analogies/metaphors that are used by (or implicit in) participants' accounts, or that occur to you; these are potential descriptions and may be more enlightening than literal prose; discuss whether or not they can be justified; do the same for any graphic materials associated with the program. Distinguish the instigator's efforts in trying to start up a program from the program itself; both are interventions, only the latter is (normally) the evaluand. Remember, you're only going to provide a summary of the program description, not a complete description, which might take more space than your complete evaluation.

B3. Consumers (Impactees)

Consumers, as the term is used here (impactees is a less ambiguous term), comprise (i) the recipients/users of the services/products (i.e., the downstream *direct* impactees) PLUS (ii) the downstream indirect impactees (e.g., recipient's family or co-workers, and others who are impacted via ripple effect[23]). Program staff are also impactees, but we usually keep them separate by calling them the midstream impactees, because the obligations to them, and the effects on them, are almost always very different and much weaker in most kinds of program evaluation (and their welfare is not the raison d'être of the program). The funding agency, taxpayers, and political supporters, who are also impactees in some sense and some cases, are also treated differently (and called upstream impactees, or, sometimes, stakeholders, although that term is often used more loosely to include all impactees), except when they are also direct recipients. Note that there are also upstream impactees who are not funders or recipients of the services but react to the announcement or planning of the program before it actually comes online (we can call them anticipators); e.g., real estate agents and employment agencies. In identifying consumers remember that they often won't know the name of the program or its goals and may not know that they were impacted or even targeted by it. (You may need to use tracer[24] &/or modus operandi methodology (GEM[25]).) While looking for the impacted population, you may also consider how others could have been impacted, or protected from impact, by variations in the program: these define alternative possible (a.k.a. virtual) impacted populations, which may suggest some ways to expand, modify, or contract the program when/if you spend time on Check-

[23] Usually this includes some members, perhaps all members, of the communities to which the direct impactees belong. In some cases, this will also include members of the research and evaluation communities, and government officials, who read or might read the evaluation report.

[24] Tracer methodology is a proactive evaluation techniqe where, for example, the evaluator marks the subjects that receive the treatment; the term comes from medical research.

[25] See earlier reference to the *Journal of Multidisciplinary Evaluation*.

point D1 (Synthesis)[26], and Checkpoint D2 (Recommendations); and hence some ways that the program should perhaps have been redefined by now, which bears on issues of praise and blame (Checkpoints B1 and D3). Considering possible variations is of course constrained by the resources available—see next checkpoint.

Note B3.1: Do not use or allow the use of the term 'beneficiaries' to refer to the impactees, since it carries with it the completely unacceptable assumption that all the effects of the program (or all the important effects) are beneficial, when of course the unintended effects may be the dealbreakers. It is also misleading, on a smaller scale, to use the term 'recipients' since many impactees are not receiving anything but merely being affected, e.g., by the actions of someone who learnt something about flu control from an educational program. The term 'recipient' should be used only for those who, whether as intended or not, are directly impacted.

B4. Resources (a.k.a. "Strengths Assessment")

This checkpoint is important for answering the questions (i) whether the program made the best use of resources available, not resources used (i.e., an *extended* kind of cost-effectiveness: see Note C3.4 for more), and (ii) what it might *realistically* do to improve (i.e., within the resources available). It refers to the financial, physical, and intellectual-social-relational assets of the program (not the evaluation!). These include the abilities, knowledge, and goodwill of staff, volunteers, community members, and other supporters. This checkpoint should cover what *could now* (or *could have been*) used, not just what *was* used: this is what defines the "possibility space," i.e., the range of what could have been done, often an important element in the assessment of achievement; in the comparisons, and in identifying directions for improvement that an evaluation considers. This means the checkpoint is crucial for Checkpoint C4 (Comparisons), Checkpoint D1 (Synthesis, for achievement), Checkpoint D2 (Recommendations), and Checkpoint D3 (Responsibility). Particularly for D1 and D2, it's helpful to list specific resources that were not used but were available in this implementation. For example, to what extent were potential impactees, stakeholders, fund-raisers, volunteers, and possible donors not recruited or not involved as much as they could have been? (As a crosscheck, and as a complement, consider all constraints on the program, including legal, environmental, and fiscal constraints.) Some matters such as adequate insurance coverage (or, more generally, risk management) could be discussed here or under Process (Checkpoint C1 below); the latter is preferable since the status of insurance coverage is ephemeral, and good process must include a procedure for regular checking on it. This checkpoint is the one that covers individual and social capital *available to* the program; the evaluator must also identify social capital *used by* the program (enter this as part of its Costs at Checkpoint C3), and, sometimes, social capital benefits *produced by* the program (enter as part of the Outcomes, at Checkpoint C2).[27] Remem-

[26] A related issue, equally important, is: What might have been done that was not done?

[27] Individual human capital is the sum of the physical and intellectual abilities, skills, powers, experience, health, energy, and attitudes a person has acquired. These blur into their—and their community's—social capital, which also includes their mutual relationships (their 'social networks') and their share of any latent attributes that their group acquires over and above the sum of their individual human capital (i.e., those that depend on interactions with others). For example, the extent of the trust or altruism (or the opposite) that pervades a group, be it family, sports team, army platoon, corporation, or other organization, is part of the value the group has acquired, a survival-

ber to include the resources contributed by other stakeholders, including other organizations and clients.[28]

B5. Values

The values of primary interest in typical professional program evaluations are usually needs, not mere personal preferences of the impactees, unless those overlap with their needs and the community/society's needs and committed values, e.g., those in the Bill of Rights and the wider body of law. Preferences as such are not irrelevant in evaluation, especially the preferences of impactees, and on some issues, e.g., surgery options and luxury spas, they are often definitive; it's just that they are generally less important—think of food preferences in children—than dietary *needs* and medical, legal, or ethical *requirements*, especially for program evaluation by contrast with product evaluation. While there are intercultural and international differences of great importance in evaluating programs, most of the values listed below are highly regarded in all cultures; the differences are generally in their precise interpretation, the contextual parameters, the exact standards laid down for each of them, and the relative weight assigned to them; and taking those differences into account is fully allowed for in the approach here. Of course, your client won't let you forget what they value, usually the goals of the program, and you should indeed keep those in mind and report on success in achieving them; but since you must value every unintended effect of the program just as seriously as the intended ones, and in most contexts you must take into account values other than those of the clients, e.g., those of the impactees and usually also those of other stakeholders, and the needs of the larger society, and the planet, you need a repertoire of values to check when doing serious program evaluation, and what follows is a proposal for that. Keep in mind that with respect to each of these (sets of) values, you will usually have to: (i) define and justify relevance to this program in this context; (ii) justify the relative weight (i.e., comparative importance) you will accord this value; (iii) identify any bars (i.e., *absolute* minimum acceptable performance standards on each value dimension) that you will require an evaluand to meet in order to be considered at all in this context; (iv) specify the empirical performance levels that will justify the application of each grade level above the bar on that value that you may wish to distinguish (e.g., define

---

related value that they (and perhaps others) benefit from having in reserve. (Example of additively limited social capital: the skills of football or other team members that will only provide (direct) benefits for others who are part of the group, e.g., a team, with complementary skills.) These forms of capital are, metaphorically, possessions or assets to be called on when needed, although they are not directly observable in their normal latent state. A commonly discussed major benefit resulting from the human capital of trust and civic literacy is support for democracy; a less obvious one, resulting in tangible assets, is the current set of efforts towards a Universal Digital Library containing 'all human knowledge'… Human capital can usually be taken to include natural gifts as well as acquired ones, or those whose status is indeterminate as between these categories (e.g., creativity, patience, empathy, adaptability), but there may be contexts in which this should not be assumed. (The short term for all this might seem to be "human resources" but that term is now widely used to mean "employees," and that is not what we are directly talking about here.)… The above is a best effort to construct the current meaning: the 25 citations in Google for 'human capital' and the 10 for 'social capital' (at 6/06/07) include many oversimplified and erroneous as well as other and inconsistent uses—few dictionaries have yet caught up with these terms (although the term 'human capital' dates from 1916).

[28] Thanks to Jane Davidson for the reminder on this last point.

what will count as fair/good/excellent. And one more thing, rarely identified as part of the evaluator's task but crucial; (v) once you have a list of impactees, however partial, you must begin to look for groups or patterns within them, e.g., pregnant women, because they have greater calorific requirements (i.e., needs) than those who are not pregnant. If you don't do this, you will probably miss extremely important ways to optimize the use of intervention resources.[29]

To get all this done, you should begin by identifying the relevant values for evaluating this evaluand in these circumstances. There are several very important groups of these. (i) Some of these follow simply from understanding the nature of the evaluand (these are sometimes called *definitional criteria of merit* (a.k.a. definitional *dimensions* of merit). For example, if it's a health program, then the criteria of merit, simply from the meaning of the terms, include the extent (a.k.a., reach or breadth) of its impact (i.e., the size and range of the demographic (age/gender/ethnic/economic) and medical categories of the impactee population), and the impact's depth (usually a function of magnitude, extent and duration) of beneficial effects. (ii) Other primary criteria of merit in such a case are extracted from a general or specialist understanding of the nature of a health program, include safety of staff and patients, quality of medical care (from diagnosis to follow-up), low adverse eco-impact, physical ease of access/entry; and basic staff competencies plus basic functioning supplies and equipment for diagnostic and minor therapeutic services. Knowing just what these values are is one reason you need either specific evaluand-area expertise or consultants who have it. (iii) Then look for particular, site-specific, criteria of merit—for example, the possible need for one or more second-language competencies in service providers. You will probably need to do or find a valid needs assessment for the targeted population, notable sub-populations, and perhaps also for any other probably impacted populations. Here you must almost always include representatives from the impactee population as relevant experts, and you may need only their cognitive and affective expertise for the needs assessment, but probably should do a serious needs assessment and have them help design and interpret it. (iv) Next, list the explicit goals/values of the client if not already covered, since they will surely want to know whether and to what extent these were met. (v) Finally, turn to the list below to find other relevant values. Validate them as relevant or irrelevant for the present evaluation, and as contextually supportable.[30]

---

[29] And if you do this, you will be doing what every scientist tries to do—find patterns in data. This is one of several ways in which good evaluation requires full-fledged traditional scientific skills; and something more as well (handling the values component).

[30] The view taken here is the commonsense one that values of the kind used by evaluators looking at programs serving the usual 'good causes' of health, education, social service, disaster relief, etc., are readily and objectively supportable, to a degree acceptable to essentially all stakeholders and supervisory or audit personnel, contrary to the doctrine of value-free social science which held that values are essentially matters of taste and hence lack objective verifiability. The ones in the list here are usually fully supportable to the degree of precision needed by the evaluator for the particular case, by appeal to publicly available evidence, expertise, and careful reasoning. Bringing them and their supporting evidence into consideration is what distinguishes evaluation from plain empirical research, and only their use makes it possible for evaluators to answer the questions that mere empirical research cannot answer, e.g., Is this the best vocational high school in this city?, Do we really need a new cancer clinic building?, Is the new mediation training program for police officers who are working the gang beat really worth what it costs to implement? In other words, the most impor-

Now, for each of the values you are going to rely on at all heavily, there are two important steps you will usually need to take, after the supremely important step of identifying all that are relevant. First, you need to establish a scale or scales on which you can describe or measure performance that is relevant to (usually one dimension of) merit. On each of these scales, you need to locate levels of performance or nature that will count as qualifying for a named value level (these are called the 'cut scores' if the dimension is measurable). For example, you might measure knowledge of first aid on a certain well-validated test, and set 90% as the score that marks an A grade, 75% as a C grade, etc.[31] Second, you will usually need to stipulate the relative importance of each of these scales in determining the overall m/w/s of the evaluand.

A useful basic toolkit for this involves doing what we call identifying the "stars, bars, and steps" for our listed values. (i) The "stars" (usually best limited to 1–3 stars) are the weights, i.e., the relative or absolute *importance* of the dimensions of merit (or worth or significance) that will be used as premises to carry you from the facts about the evaluand, as you locate or determine those, to the evaluative conclusions you need. Their absolute importance might be expressed qualitatively (e.g., major/medium/minor, or by letter grades A-F); or quantitatively (e.g., points on a five, ten, or other point scale, or—often a better method of giving relative importance—by the allocation of 100 'weighting points' across the set of values[32]); or, if merely relative values are all you need, these can even be expressed in terms of an ordering of their comparative importance (rarely an adequate approach). (ii) The "bars" are *absolute minimum standards* for acceptability, if any: that is, they are minima on the particular scales, scores or ratings, each of which must be 'cleared' (exceeded) if the candidate is to be acceptable, no matter how well s/he scores on other scales. Note that an F grade for performance on a particular scale does not mean 'failure to clear a bar,' e.g., an F on the GRE quantitative scale may be acceptable if offset by other virtues, for selecting students into a creative writing program[33]. Bars and stars may be set on any relevant properties (a.k.a. dimensions of merit), or directly on dimensions of measured (valued) performance, and may additionally include holistic bars or stars[34]. (iii) In serious

---

tant practical questions, for most people—and their representatives—who are looking at programs (and the same applies to product, personnel, policy evaluation, etc.)

[31] This difficult process of identifying 'cutscores' is a specialized topic in test theory—there is a whole book by that title devoted to discussing how it should be done. A good review of the main issues is in Gene Glass' paper (ref to follow).

[32] Better because it forces constraint on the grader, thereby avoiding one major barrier to comparability.

[33] If an F is acceptable on that scale, why is that dimension still listed at all—why is it relevant? Answer: it may be one of several on which high scores are weighted as a credit, on one of which the candidate must score high, but not on any particular one. In other words the applicant has to have *some* special talent, but a wide range of talents are acceptable. This might be described as a case where there is a 'group' or 'holistic' bar, i.e., a '*floating' bar* on a group of dimensions, which must be cleared by the evaluand's performance on at least one of them. It can be exhibited in the list of dimensions of merit by bracketing the group of dimensions in the abscissa, and stating the height of the floating bar in an attached note. Example: "Candidates for admission to the psychology grad program must have passed one upper division statistics course."

[34] Refocused example: The candidates for admission to a graduate program—whose quality is one criterion of merit for the program—may meet all dimension-specific minimum standards in each

evaluation, it is often appropriate to locate "steps" i.e., *points or zones on measured dimensions of merit where the weight changes*, if the mere stars don't provide enough precision. An example of this is the setting of several cutting scores on the GRE for different grades in the use of that scale for the two types of evaluation given above (evaluating the program and evaluating applicants to it). The grades, bars, and stars (weights), are often loosely included under what is called '*standards*.' (Bars and steps may be fuzzy as well as precise.)

Three values are of such general importance that they receive full checkpoint status and are listed in the next section: cost (minimization of), superiority (to comparable alternatives), and generalizability/exportabillity. Their presence in the KEC brings the number of types of values considered, including the list below, up to a total of 21.

At least check all these values for relevance and look for others: and for those that are relevant, set up an outline of a set of defensible standards that you will use in assessing the evaluand's performance on that standard. Since these are context-dependent (e.g., the standards for a C in evaluating free clinics in Zurich today are not the same as for a C in evaluating a free clinic in Darfur at the moment), and the client's evaluation-needs—i.e., the questions they need to be able to answer—differ massively, there isn't a universal dictionary for them. You'll need to have a topical expert on your team or do a good literature search to develop a draft, and eventually run serious sessions with impactee and other stakeholder representatives to ensure defensibility for the revised draft, using focus groups, surveys, and/or interviews. The final version of each of the standards, and the set of them, is often called a rubric, meaning a table translating evaluative terms into observable or testable terms and/or vice versa.[35] These rubrics are invaluable when we come to Checkpoint D1, the Synthesis checkpoint.

(i) Definitional values—those that follow from the definitions of terms in standard usage (e.g., breadth and depth of impact are, definitionally, dimensions of merit for a public health program), or that follow from the contextual implications of having an ideal or excellent evaluand of this type (e.g., an ideal shuttle bus service for night shift workers would feature increased frequency of service around shift change times). The latter draw from general knowledge and to some extent from program-area expertise.

(ii) Needs of the impacted population, via a needs assessment (it is essential to distinguish performance needs (e.g., need for mobility, health, skills) from treatment needs (e.g., need

---

respect for which these were specified (i.e., they 'clear these bars'), but may be so close to missing the bars (minima) in so many respects, and so weak in respects for which no minimum was specified, that the selection committee feels they are not good enough for the program. We can describe this as a case where they failed to clear a holistic (a.k.a. overall or group or floating) bar that was implicit in this example, but can often be made explicit through dialog. (The usual way to express a quantitative holistic bar is via an average grade; but that is not always the best way to specify it and is often not strictly defensible since the grading scale is not an interval scale.)

[35] The term 'rubric' as used here is a technical term originating in educational testing parlance; this meaning is not in most dictionaries, or is sometimes distinguished as 'an assessment rubric.' A complication we need to note here is that some of the observable/measurable terms may themselves be evaluative, at least in some contexts. That's just a reflection of the fact that much of our basic knowledge, back to the dawn of hominid existence, is evaluative knowledge e.g., of the best and bad way to do things, the best people to follow or not to fight.

for specific medication, education, or delivery systems); and needs that are currently met from unmet needs;[36] and meetable needs from ideal but impractical or impossible-with-present-resources needs (consider the Resources checkpoint))… Needs for X are the levels of every factor below which the subject(s) will be unable to function *satisfactorily* (not the same as *optimally, maximally, or ideally)*; and of course, what functions are under study and what level will count as satisfactory will vary with the study and the context… The needs are matters of fact, not values in themselves, but in any context that accepts the most rudimentary ethical considerations (i.e., the non-zero value of the welfare of all human beings, including avoidance of unnecessary harm, pain, loss of capacities, skills, and knowledge), those facts are value-imbued… Needs may have macro as well as micro levels that must be considered; e.g., there are local *community* needs, *regional* needs within a country, *national* needs, *geographic regional* needs, and *global* needs (these often overlap, e.g., in the case of building codes (illustrated by their absence in the Port-au-Prince earthquake of 2010)… Doing a needs assessment is sometimes the most important part of an evaluation, and in much of the literature is based on invalid definitions of need, e.g., the idea that needs are the gaps between the actual level of some factor (e.g., income, calories) and the ideal level… Of course, the basic needs for sustainable supplies of food, water, shelter, medical care, and clothing are known universals, but a good case can be made for some others, especially basic knowledge about the world and survival skills (including ethics); justice; and resilience… In our times, when economic values have been a major focus, often to excess (and often too little), the major indicator of need by major international institutions such as the World Bank has often been the GDP (Gross Domestic Product). The oversimplification involved in using a single value as the dependent variable has often been remarked, and, the OECD, in May, 2011, released a more comprehensive compound indicator, the Better Life Index (BLI), which includes measures of health, education, and 9 other factors (see Note 5.2 below).

Final note; check the Resources checkpoint, a.k.a. Strengths Assessment, for other entities valued in that context and hence of value in this evaluation.

(iii) Logical requirements (e.g., consistency, sound inferences in design of program or measurement instruments e.g., tests).

(iv) Legal requirements (but see (v) and (vi) below).

(v) Ethical requirements (overlaps with others, especially legal, and overrides them when in conflict), usually including (reasonable) safety for all participants, and confidentiality (sometimes anonymity) of all records, for all impactees. (Problems like conflict of interest and protection of human rights have federal legal status in the US, and are also regarded as scientifically good procedural standards, and as having some very general ethical status.) In most circumstances, needs such as health, shelter, education, equable treatment, voting rights, and other welfare considerations such as reduction of severe harm, hurt, and risk, for impactees and potential impactees, are obvious values to which ethical weight must be

[36] A very common mistake—reflected in definitions of needs that are widely used—is to think that met needs are not 'really' needs, and should not be included in a needs assessment. That immediately leads to the 'theft' of resources that are meeting currently met needs, in order to serve the remaining unmet needs, a blunder that can cost lives. So, identify all needs, *then* identify the ones that are still unmet.

given. Never forget that ethics trumps all other cards, including respect for the ethics of others, and if you don't see how that can avoid contradiction, you're not alone—but you probably need to do more homework in ethical theory.[37]

(vi) Cultural values (not the same as needs or wants or ethical codes, although overlapping with them) held with a high degree of respect (and thus distinguished from matters of manners, style, taste, etc.), of which an extremely important one in many cultures is honor; another group, not always distinct from that one, concerns respect for ancestors, elders, tribal or totem spirits, and local deities. These, like legal requirements, are subject to override, in principle at least, by basic ethical values, although often mistakenly taken to have the same and sometimes higher status (especially sacrilege). Sometimes they do have the same status because they are sometimes just versions of basic ethical values.

(vii) Personal, group, and organizational goals/desires (unless you're doing a goal-free evaluation) if not in conflict with ethical/legal/practical considerations, including reduction of harm, hurt, and risk—and gratification of aesthetic preferences. Example: the professional values of the American Evaluation Association. These are usually less important than the needs of the impactees, since they lack specific ethical or legal backing, but are enough by themselves to drive the inference to many evaluative conclusions about e.g., what recreational facilities to provide in community-owned parks, subject to consistency with ethical and legal constraints. They include some things that are often but not always wrongly claimed as needs rather than mere desires, e.g., convenience, recreation, respect, earned recognition, excitement, and compatibility with aesthetic preferences of recipients, donors, or rulers. There are borderline cases between needs and desires such as joy, security, and interest, and these can be allocated to needs or desires as seems most appropriate in the particular case; but double counting must be avoided.

(viii) Environmental needs, if these are regarded as not simply reducible to 'human needs with respect to the environment,' e.g., habitat needs of other species (fauna or flora), and perhaps Gaian 'needs of the planet.'

(ix) Fidelity to alleged specifications (a.k.a. "authenticity," "adherence," "implementation," "dosage," or "compliance")—this is often usefully expressed via an "index of implementation"; and—a different but related matter—consistency with the supposed program model (if you can establish this BRD—beyond reasonable doubt); crucially important in Checkpoint C1.

(x) Sub-legal but still important legislative preferences (e.g., GAO used to determine these from an analysis of the hearings in front of the sub-committee in Congress from which the legislation emanated.)

(xi) Professional standards (i.e., standards set by the profession) of quality that apply to the

---

[37] The resolution of conflicts between ethical systems, which is essential in order to avoid death in the swamp of relativism, is easier once you accept the fact that the best-supported evaluative conclusions (such as ethical conclusions), are simply a subset of all truths. Then one follows the practice of good scientists and mathematicians, who (typically although not always) show how to accept the need for respecting the views of others while committed to the view that the truth must be respected above all, even when it contradicts the views of others; i.e., they go with the balance of evidence. This means that the evaluator must understand the arguments about the basis for ethics itself, not just the ethical code that governs evaluators.

evaluand,[38] as long as not already covered in (v) on this list.

(xii) Expert refinements of any standards lacking a formal statement, e.g., ones in (ix); but it is important to avoid double counting.

(xiii) Historical/Traditional standards.

(xiv) Scientific merit (or worth or significance).

(xv) Technological m/w/s.

(xvi) Marketability, in commercial program/product evaluation.

(xvii) Political merit, if you can establish it BRD.

(xviii) Risk reduction (risk sometimes just means chance, but more usually and here means the probability of failure (or loss); or, sometimes, of the loss (or gain) that would *result from* failure; or, sometimes, the product of these two). Risk in this context does *not* mean the probability of error about the facts or values we are using as parameters—i.e., the level of confidence we have in our data or conclusions. Risk here is the value or disvalue of the chancy element in the enterprise in itself, as an independent positive or negative element—positive for those who are positively attracted by gambling as such (this is usually taken to be a real attraction, unlike risk-tolerance) and negative for those who are, by contrast, risk-averse (a.k.a. conservative, in one sense). This consideration is particularly important in evaluating *plans* (preformative evaluation) and in formative evaluation, but is also relevant in summative and ascriptive evaluation when either is done prospectively (i.e., before all data is available, as is common in *policy analysis* and *futurism*). There is an option of including this under personal preferences, item (vii) above, but it is often better to consider it separately since: (i) it improves decision-making quality if it is explicit, (ii) can be very important, and (iii) is a matter on which evidence/expertise (in the logic of probability) should be brought to bear, not simply a matter of personal taste.[39]

---

[38] Since one of the steps in the evaluation checklist is the meta-evaluation, in which the evaluation itself is the evaluand, you will also need, when you come to that checkpoint, to apply professional standards for evaluations to the list. Currently the best ones would be those developed by the Joint Committee (*Program Evaluation Standards 2e* (Sage), but there are several others of note, e.g., the GAO Yellow Book), and perhaps the KEC. And see the final checkpoint in the KEC, D5.

[39] Note that, as briefly indicated above, risk is often defined in the technical literature as the *product* of the likelihood of failure and the magnitude of the disaster if the program, or part of it, does fail (the possible loss itself is often called 'the hazard'); but in common parlance, the term 'risk' is often used to mean either the probability of disaster ("very risky") or the disaster itself ("the risk of death"). Now the classical definition of a gambler is someone who will prefer to pay a dollar to get a 1 in 1,000 chance of making $1,000 over paying a dollar to get a 1 in 2 chance of making $2, even though the expectancy is the same in each case; the risk-averse person will reverse those preferences and in extreme cases will prefer to simply keep the dollar; and the rational risk-tolerant person will, supposedly, treat all three options as of equal merit. So, if this is correct, then one might argue that the more precise way to put the value differences here is to say that the gambler is not attracted by the element of chance *in itself* but by the *possibility of making the larger sum* despite the low probability of that outcome, i.e., that s/he is less risk-averse, not more of a risk-lover. (I think this way of putting the matter actually leads to a better analysis, i.e., the view that any of these positions can be rational depending on contextual specification of the cost of Type 1 vs. Type 2 errors.)

(xix) Last but not least—Resource economy (i.e., how low-impact the program is with respect to short-term and long-term limits on resources of money, space, time, labor, contacts, expertise and the eco-system). Note that 'low-impact' is not what we normally mean by 'low-cost' (covered separately in Checkpoint C3) in the normal currencies (money and non-money), but refers to absolute (usually meaning irreversible) loss of available resources in some framework, which might range from a single person to a country or the globe. This could be included under an extended notion of (opportunity) cost or need, but has become so important in its own right that it is probably better to put it under its own heading as a value, as here. It partly overlaps with Checkpoint C5, because a low score on resource economy undermines sustainability, so watch for double counting. Also check for double counting against value (ix), if that is being weighted by client or audiences and is not overridden by ethical or other higher-weighted concerns.

Fortunately, bringing this long list of values and their standards to bear[40] is less onerous than it may appear, since many of these values will be unimportant or only marginally important in the specific case, although each one will be crucially important in other particular cases. And doing all this values-analysis will be easy to do sometimes because all stakeholders agree on the ones involved, although very hard on other occasions—it can often require expert advice and/or impactee/stakeholder advice. Of course, some of these values will conflict with others (e.g., impact size with resource economy), so their relative weights may then have to be determined for the particular case, a non-trivial task by itself. Because of these possibly extreme difficulties, you need to be very careful not to *assume* that you have to generate a *ranking* of evaluands in the evaluation you are asked to do, since if that's not required, you can often avoid settling the issue of relative weights of criteria, or at least avoid any precision in settling it, by simply doing a grading of each evaluand, on a profiling display (i.e., showing the merit on all relevant dimensions of merit in a bar-graph for each evaluand(s)). That profile will exhibit the various strengths and weaknesses of each evaluand, ideal for helping them to improve, and for helping clients to refine their weights for the criteria of merit, which will often make it obvious which one is the best choice.

***Note B5.1*:** You must cover in this checkpoint all values that you will use, *including those used in evaluating the side-effects* (if any), not just the intended effects (if those materialize). Some of these values will probably occur to you only after you *find* the side-effects (Checkpoint C2), but that's not a problem—this is an iterative checklist, and in practice that means you will often have to come back to modify findings on earlier checkpoints. But finding them is a specific task that must be allocated time and skill in the evaluation plan, and the damage done by side-effects must be sought out, not just treated as 'lucky that I noticed this.' For example, few program evaluation designs that I have seen in philanthropic work include looking for increased dependency (i.e., loss of motivation to solve problems without external help) as a negative side-effect of benevolent interventions.

However described, this can be a major value difference between people and organizations e.g., venture capitalist groups vs. city planning groups.

[40] 'Bringing them to bear' involves: (a) identifying the relevant ones, (b) specifying them (i.e., determining the dimensions for each and a method of measuring performance/achievements on all of these scales), (c) validating the relevant standards for the case, and (d) applying the standards to the case.

***Note B5.2*:** Multivalue indices. While it is often important to look, even if briefly, at all the possible values that may be relevant to a particular evaluand, or for a particular client, custom packages arrived at this way have the disadvantage of not yielding good comparability. On the other hand, only looking at one or two indicators is usually misleading. In our times, when the focus is so often on economic variables, sometimes excessively so, sometimes inadequately, there has been a long tradition of using the GDP (Gross Domestic Product), measured in currency terms, as the supreme comparison variable for many comparisons of international welfare, e.g., by the World Bank. A six-dimensional alternative to this was created by Scriven et al. in 2005 for international evaluations of the Heifer philanthropy's work and developed further in the six succeeding years of that work in 23 countries. As previously mentioned, in May, 2011, the OECD released an eleven-dimensional alternative called the BLI (Better Life Index), with details on its application to 34 countries; the two approaches should be integrated at some point. These both now have enough practical experience built into them to be viable options for many purposes, but some tasks will require new task-specific indexes, although it is to be hoped that fewer will try to get by with the GDP. In policy analysis, a branch of evaluation that adjoins program evaluation, although it is at a somewhat earlier stage of development, there's a well-known list of values to be considered in judging policies; it overlaps some but not all of the list here and may be useful just because it's a different shuffle of the values deck. One version of this (Wikipedia 3/13) is: (i) Ecological impacts—such as biodiversity, water quality, air quality, habitat quality, species population, etc. (ii) Economic efficiency—commonly expressed as benefits and costs. (iii) Distributional equity—how policy impacts are distributed amongst different demographics. Factors that can affect the distribution of impacts include location, ethnicity, income, and occupation. (iv) Social/Cultural acceptability—the extent to which the policy action may be opposed by current social norms or cultural values. (v) Operational practicality—the capacity required to actually operationalize the policy. (vi) Legality—the potential for the policy to be implemented under current legislation versus the need to pass new legislation that accommodates the policy. (vii) Uncertainty—the degree to which the level of policy impacts can be known.

PART C: SUBEVALUATIONS

Each of the following five core dimensions of an evaluation requires both: (i) a 'fact-finding'[41] phase, followed by (ii) the process of combining those facts with whatever values from B5 that are relevant to this dimension of merit bear on those facts, which yields (iii) the subevaluations. In other words, Part C requires[42] the completion of five separate kinds of inference from [(i) plus (ii)] to (iii), i.e., from What's So? to So What? For example (in the case of C2, Outcomes), from (i) 'the outcomes were measured as XX,' and (ii) 'outcomes of this size are valuable to the degree YY' to (iii) 'the effects were extremely beneficial,' or 'insignificant in this context,' etc. Making that step requires, in each case, a premise of type (ii)

[41] Here, and commonly, this sense of the term means *non-evaluative* fact-finding. There are plenty of evaluative facts that we often seek or pronounce, e.g., whether the records show that an attorney we are considering has a history of malpractice; whether braided nylon fishline is as good as wire for fish over 30kg.

[42] Although this is generally true, there are evaluations in which one or more of the sub-evaluations are irrelevant, e.g., when cost is of no concern.

that forms a bridge between facts and values; these are usually some kind of 'rubric,' discussed further in the D1 (Synthesis) checkpoint. (iii) Finally, you will often need to do an overall *synthesis of the sub-evaluations* into a holistic final evaluation… The first two of the following core checkpoints will, in one case or another, use rubrics referring to nearly all the values from Checkpoint B5 and bear most of the load in determining *merit*; the next three checkpoints are defined in terms of specific values of great general importance, named in their heading, and particularly relevant to *worth* (Checkpoint C3 and C4) and *significance* (Checkpoints C4 and C5).

C1. Process

We start with this core checkpoint because it forces us to confront immediately the merit of the *means* this intervention employs, so that we are able, as soon as possible, to answer the question whether the (intended or unintentionally produced) *ends*—many of which we'll cover in the next checkpoint—justify the *means*, in this specific case or set of cases. The Process checkpoint involves the assessment of the m/w/s of everything that happens or applies before true outcomes emerge, especially: (i) the vision, design, planning and operation of the program, from the justification of its goals (if you're not operating in goal-free mode)—and note that goals may have changed or be changing since the program began—through design provisions for reshaping the program under environmental or political or fiscal duress (including planning for worst-case outcomes); to the development and justification of the program's supposed 'logic' a.k.a. design (but see Checkpoint D2), along with (ii) the program's 'implementation fidelity' (i.e., the degree of implementation of the supposed archetype or exemplar program, if any). This index is also called "authenticity," "adherence," "alignment," "fidelity," "internal sustainability," or "compliance".[43] You must also, under Process, check the accuracy of the official name or subtitle (whether descriptive or evaluative), or the official description of the program e.g., "an inquiry-based science education program for middle school"—one, two, three, or even four of the components of this compound descriptive claim (it may also be contextually evaluative) can be false. (Other examples: "raises beginners to proficiency level", "advanced critical thinking training program"). Also check (iii) the quality of its management (especially (a) the arrangements for getting and appropriately reporting evaluative feedback (that package is often much of what is called accountability or transparency), along with support for learning from that feedback, and from any mistakes/solutions discovered in other ways, along with meeting more obviously appropriate standards of accountability and transparency[44]; (b) the quality

[43] Several recent drug studies have shown huge outcome differences between subjects filling 80% or more of their prescriptions and those filling less than 80%, in both the placebo and treatment groups, *even when it's unknown* how many of those getting the drug from the pharmacy are actually taking it, and *even though there is no overall difference* in average outcomes between the two groups. In other words, mere aspects of the process of treatment can be more important than the nature of the treatment or the fact of treatment status. So be sure you know what the process actually comprises, and whether any comparison group is closely similar on each aspect.

[44] It's important to check whether evaluation began when it should begin i.e., well before the program begins, so that formative advice on the design, the evaluability, and precursor effects could be obtained when it was most useful. (This is sometimes called 'preformative' evaluation.) It is usually a mistake to dismiss this as 'evaluation of the design, not the program' (cf. volcanic eruption).

of the risk-management,[45] including the presence of a full suite of 'Plan B' options; (c) the extent to which the program planning covers issues of sustainability and not just short-term returns (this point can also be covered in C5). You need to examine all activities and procedures, especially the program's general learning/training process (e.g., regular 'updating training' to cope with changes in the operational and bio-environment, staff aging, essential skill pool, new technology[46]); attitudes/values e.g., honesty, class or gender bias; and morale. Of course, management quality is something that continues well beyond the beginning of the program, so in looking at it, you need to be clear when it had which form or you won't be able to ascribe results—good or bad—to management features, if you are hoping to be able to do that. Organization records often lack this kind of detail, so try to improve their practice, at least for the duration of your evaluation.

As mentioned before, under this heading you may or may not need to examine the quality of the original 'logic of the program' (the rationale for its design) and its current logic (both the current official version and the possibly different one implicit in the operations or in staff behaviour rather than rhetoric). It is not generally appropriate to try to determine and affirm whether the model is correct in detail and in scientific fact unless you have specifically undertaken that kind of (usually ambitious and sometimes unrealistically ambitious) analytic evaluation of the program design/plan/theory. You need to judge with great care whether comments on the plausibility of the program theory are likely to be helpful, and, if so, whether you are sufficiently expert to make them. Just keep in mind that it's never been hard to evaluate aspirin for e.g., its analgesic and side-effects, although it is only very recently that we had any idea how/why it works. It would have been a logical error—and unhelpful to society—to make the earlier evaluations depend on solving the causal mystery. It helps to keep in mind that there's no mystery until you've done the evaluation, since you can't explain outcomes if there aren't any (or explain why there aren't any until you've shown that that's the situation). So if you can be helpful by evaluating the program theory, and you have the resources to spare, do it; but doing this is not an essential part of doing a good evaluation, will often be a diversion, is always an extra cost, and is sometimes a cause for disruptive antagonism.

Process evaluation may also include (iv) the evaluation of what are often called "outputs," (usually taken to be 'intermediate outcomes' that are developed en route to 'true outcomes,' the longer-term results that are sometimes called 'impact'). Typical outputs are knowledge, skill, or attitude changes in staff (or clients), when these changes are not major outcomes in their own right. Remember that in any program that involves learning, whether incidental or intended, the process of learning is gradual and at any point in time, long before you can talk about outcomes/impact, there will have been substantial learning that produces a gain in individual or social capital, which must be regarded as a tangible gain for the program and for the intervention. It's not terribly important whether you call it

---

[45] Risk-management has emerged fairly recently as a job classification in large organizations, growing out of the specialized task of analyzing the adequacy and justifiability of the organization's insurance coverage, but now including matters such as the adequacy and coordination of protocols and training for emergency response to natural and human-caused disasters, the identification of responsibility for each risk, and the sharing of risk and insurance with other parties.

[46] See also my paper on "Evaluation of Training" at ***michaelscriven.info*** for a checklist that massively extends Kirkpatrick's groundbreaking effort at this task.

process or output or short-term outcome, as long as you find it, estimate it, and record it—once. (Recording it under more than one heading—other than for merely annotative reasons—leads to double counting when you are aiming for an overall judgement.)

Note C1.1: Six other reasons or aspects of reasons why process is an essential element in program evaluation, despite the common tendency in much evaluation to place almost the entire emphasis on outcomes: (v) gender or racial (etc.) prejudice in selection/promotion/-treatment of staff is an unethical practice that must be checked for, and comes under process; (vi) in evaluating health or other training programs that involve medication or exercise, 'adherence' or 'implementation fidelity' means following the prescribed regimen including drug dosage, and it is often vitally important to determine the degree to which this is occurring—which is also a process consideration. We now know, because researchers finally got down to triangulation (e.g., via randomly timed counts by a nurse-observer, of the number of pills remaining in the patient's medicine containers), that adherence can be very low in many needy populations, e.g., Alzheimer's patients, a fact that completely altered some evaluative conclusions about treatment efficacy; (vii) the process may be where the value lies—writing poetry in the creative writing class may be a good thing to do in itself, not because of some later outcomes (same for having fun at school, in kindergarten at least; painting as an art; and marching to protest war or exploitation, even if it doesn't succeed); (viii) the treatment of human subjects must meet federal, state, and other ethical standards, and an evaluator can rarely avoid the responsibility for checking that they are met; (ix) as the recent scandal in anaesthesiology underscores, many widely accepted evaluation procedures, e.g., peer review, rest on assumptions that are sometimes completely wrong (e.g., that the researcher actually got the data he reported from real patients), and the evaluator should try to do better than rely on such assumptions. (x) We've listed above some examples of ethical considerations that need checking, but it's as well to keep in mind that many others can emerge: the basic slogan has to be that process must always be evaluated in terms of ethics, which will turn up in new places just when you think you've covered all the specific possibilities.

Note C1.2: It is still frequently said that formative evaluation is the same as process evaluation, but the two are completely different: the latter refers to a component of all evaluations and the former to one of several possible roles/purposes for an evaluation.

C2. Outcomes (a.k.a. Effects)

This checkpoint (a.k.a. 'impact evaluation') is the poster-boy of many evaluations, and the one that many people mistakenly think of as covering 'the results' of an intervention. (In fact, the results are everything covered in Part C plus Part D.) This checkpoint does cover the 'ends' at which the 'means' discussed in C1 (Process) are aimed (i.e., the goals of the program, but remember that not all evaluands have goals, e.g., rough diamonds and neonates, and also that many programs with huge impacts, good or bad, score nothing at all with respect to reaching their intended goals). Your task requires the identification of *all* effects (good *and* bad; intended *and* unintended; immediate[47], short term *and* long term

[47] The 'immediate' effects of a program are not only the first effects that occur *after* the program starts up, but should also include major effects that occur *before* the program starts. These (preformative) effects are the ones that impact 'anticipators' who react to the announcement of—or have secret intelligence about—the future start of the program. For example, the award of the 2012

(i.e., occurring long after the evaluation concludes—check 'sustainability' in D5 below)) on: (i) program recipients (both targeted and untargeted—an example of the latter are thieves of aid or drug supplies); (ii) other impactees, e.g., families and friends—and enemies—of recipients; and (iii) the environment (biological, physical, local and remote social environments). (These are all, roughly speaking, the focus of Campbell's 'internal validity.') Finding outcomes cannot be done by hypothesis-testing methodology, because: (i) some of the most important effects are unanticipated ones (the four main ways to find these are: (a) goal-free evaluation, (b) trained observation, (c) interviewing (of impactees and critics of the program[48]) that is explicitly focused on finding side-effects, and (d) using previous experience (as provided in the research literature and the mythical "Book of Causes"[49]). And (ii) because determining the m/w/s of the effects—that's the bottom line result you have to get out of this sub-evaluation—is often the hard part, not just determining whether there are any effects, or even what they are intrinsically, and who they affect (some of which you *can* get by hypothesis testing)… Immediate outcomes (e.g., the publication of instructional leaflets for AIDS caregivers) are often called 'outputs,' especially if their role is that of an intermediate cause or intended cause of main outcomes, and they are normally covered under Checkpoint C1. But note that some true outcomes (including results that are of major significance, whether or not intended) can occur during the process but may be best considered here, especially if they are highly durable… (Long-term results are sometimes called 'effects' (or 'true effects' or 'results') and the totality of these is often referred to as 'impact'; but you should adjust to the highly variable local usage of these terms by clients/-audiences/stakeholders.)… Note that you must pick up effects on individual and social capital here (see the earlier footnote): much of that ensemble of effects is normally not counted as outcomes, because they are gains in latent ability (capacity, potentiality), not necessarily in observable achievements or goods. Particularly in educational evaluations aimed at improving test scores, a common mistake is to forget to include the (possibly life-long) gain in ability as an effect.

Sometimes, not always, it's useful and feasible to provide explanations of success/failure in terms of components/context/decisions. For example, when evaluating a statewide consortium of training programs for firemen dealing with toxic fumes, it's probably fairly easy to identify the more and less successful programs, maybe even to identify the key to success as particular features—e.g., realistic simulations—that are to be found in and only in the

---

Olympic Games to Rio de Janeiro, made several years in advance of any implementation of the planned constructions etc. for the games, had a huge immediate effect on real estate prices, and later on favela policing for drug and violence control.

[48] Thanks to Jonny Morell for mentioning this.

[49] The Book of Causes shows, when opened at the name of a condition, factor, or event: (i) on the left (verso) side of the opening, all the things which are known to be able *to cause it*, in some context or other (which is specified); and (ii) on the right (recto) side, all the things which *it can cause*: that's the side you need in order to guide the search for side-effects. Since the BofC is only a virtual book, you usually have to create the relevant pages, using all your resources such as accessible experts and a literature/internet search. Good forensic pathologists and good field epidemiologists, amongst other scientists, have very comprehensive 'local editions' of the BofC in their heads and as part of the informal social capital of their guild. Modern computer technology makes real BofCs feasible, perhaps imminent (a Google project?).

successful programs. To do this usually does not require the identification of the whole operating logic/theory of program operation. (Remember that the operating logic is not necessarily the same as: (i) the original official program logic, (ii) the current official logic, (iii) the implicit, logics or theories of field staff). Also see Checkpoint D2 below.

Given that the most important outcomes may have been unintended (a broader class than unexpected), it's worth distinguishing between side-effects (unintended effects on the target population and possibly others) and side-impacts (unintended impacts of any kind on non-targeted populations).

The three biggest methodological problems with this checkpoint are: (i) establishing the causal connection, especially when there are many possible or actual causes, and—a separate point—(ii) the attribution of portions of the effect to each of them, if this is requested,[50] and (iii) deciding how to describe the outcomes. We've already stressed that the description normally needs to be evaluative, not just descriptive, and of course should include a comment on the level of statistical significance if appropriate measures are available. But that leaves open many choices. To give just one example from qualitative work: the usual way is to describe what changes have occurred—but it may be more important, depending on the context, to describe what changes did *not* occur (e.g., in a situation where the context changed considerably for the worse). In quantitative methodology it took half a century to move the norm of reporting from measured change + statistical significance, to add the effect size; it should not take that long to add cases where zero effect size is a big gain, and robustness estimates and prevented change when appropriate. Or, to add a different twist, it may be more important to provide the outcomes as seen by the subjects not the client; or as seen by the donor, not the manager of the program. Or, at least, include more than one of these approaches.

Note C2.1: As Robert Brinkerhoff argues, success cases may be worth their own analysis as a separate group, regardless of the average improvement (if any) due to the program (since the benefits in those cases alone may justify the cost of the program)[51]; the failure cases should also be examined, for differences and toxic factors.

Note C2.2: Keep the "triple bottom-line" approach in mind. This means that, *as well as* (i) conventional outcomes (e.g., learning gains by impactees), you should always be looking for (ii) community (include social capital) changes, and (iii) environmental impact… And always comment on (iv) the risk aspect of outcomes, which is likely to be valued very differently by different stakeholders… Especially, do not overlook (v) the effects on the program staff, good and bad, e.g., lessons and skills learned, and the usual effects of stress; and (vi) the pre-program effects mentioned earlier: that is, the (often major) effects of the announcement or discovery that a program will be implemented, or even *may* be implemented. These effects include booms in real estate and migration of various groups to/from the community, and are sometimes more serious in at least the economic dimension than the directly caused results of the program's implementation on this impact group,

---

[50] On this, consult recent literature by, or cited by, Cook or Scriven, e.g., in the 6th and the 8th issues of the *Journal of Multidisciplinary Evaluation (2008)*, at jmde.com, and *American Journal of Evaluation (3, 2010)*), and in "Demythologizing Causation and Evidence" in "What Counts as Credible Evidence in Applied Research and Evaluation?" eds. S. Donaldson and C. Christie (Sage, 2011, 2014).

[51] Robert Brinkerhoff in *The Success Case Method* (Berrett-Koehler, 2003).

the 'anticipators.' Looking at these effects carefully is sometimes included under what is called 'preformative evaluation' (which also covers looking at other dimensions of the planned program, such as evaluability).

Note C2.3: It is usually true that evaluations have to be completed long before some of the main outcomes have, or indeed could have, occurred—let alone have been inspected carefully. This leads to a common practice of depending heavily on predictions of outcomes based on indications or small samples of what they will be. This is a risky activity, and needs to be carefully highlighted, along with the assumptions on which the prediction is based, and the checks that have been made on them, as far as is possible. Some very expensive evaluations of giant international aid programs have been based almost entirely on outcomes estimated by the same agency that did the evaluation *and* the installation of the program—estimates that, not too surprisingly, turned out to be absurdly optimistic. Pessimism can equally well be ill-based, for example predicting the survival chances of Stage IV cancer patients is often done using the existing data on five-year survival—but that ignores the impact of research on treatment in (at least) the last five years, which has often been considerable. On the other hand, waiting for the next Force 8 earthquake to test disaster plans is stupid; simulations, if designed by a competent external agency, can do a very good job in estimating long-term effects of a new plan.

Note C2.4: Identifying the impactees is not only a matter of identifying each individual—or at least small group—that is impacted (targeted or not), hard though that is; it is also a matter of finding patterns in them, e.g., a tendency for the intervention to be more successful (or unsuccessful) with women than men. Finding patterns in the data is of course a traditional scientific task, so here is one case amongst several where the task of the evaluator includes one of the core tasks of the traditional scientist.

Note C2.5: Furthermore, if you have discovered any unanticipated side-effects at all, consider that they are likely to require evaluation against some values that were not considered under the Values checkpoint, since you were not expecting them; you need to go back and expand your list of relevant values, and develop scales and rubrics for these, too. (For example, reduced autonomy is a common 'sleeper side-effect' of benevolent interventions.)

Note C2.6: Almost without exception, the social science literature on *effects* identifies them as 'what happened after an intervention that would not have happened without the presence of the intervention'—this is the so-called 'counterfactual property.' This identification is a serious fallacy, and shows culpable ignorance of about a century's literature on causation in the logic of science (see references given above on causation, e.g., in footnote 8). Many effects would have happened anyway, due to the presence of other factors with causal potential; this is the phenomenon of 'overdetermination,' which is common in the social sciences. For example, the good that Catholic Charities does in a disaster might well have occurred if they were not operating, since there are other sources of help with identical target populations; this does not show they were not in fact the causal agency nor does it show that they are redundant.

Note C2.7: A professional evaluator needs to be sophisticated about a wide range of investigatory designs aimed at impact estimation, notably the quantitative/experimental ones. A good discussion of these is the *Handbook on impact evaluation: quantitative methods and*

practices, a World Bank publication available online—Google the title to get the current location. But don't make the mistake of thinking that such approaches are in general either superior to, or necessary for, identification of impacts. Sometimes they will save the day, sometimes they will be impossible, an extravagance, or unethical.

Note C2.8: Evaluators are very commonly called in much too late in the day, notably much too late to get baseline ('pre-test') data for the evaluand. They must be adept at 'reconstructing the baseline' and at recognizing when this can't be done to the extent required. Generally speaking, a good deal can be learnt about doing this by studying the methodology of historians, not a common part of evaluation training programs. It often requires skilled interviews of elders, skilled reading of middens and the aging of artifacts, and skilled digging in archives of biographies, newspapers, weather and market records.

C3. Costs

This checkpoint brings in what might be called 'the other quantitative element in evaluation' besides statistics, i.e., (most of) cost analysis. It is certainly the neglected quantitative component. But don't forget there is also such a thing as *qualitative* cost analysis, which is also very important—and, done properly, it's not a feeble surrogate for quantitative cost analysis but an essentially independent effort. Note that both quantitative and qualitative cost-analysis are included in the economist's definition of cost-effectiveness. Both are usually very important in determining worth (or, in one sense, value) by contrast with plain merit (a.k.a. quality). Both were almost totally ignored for many years after program evaluation became a matter of professional practice; and a recent survey of journal articles by Nadini Persaud shows both are still seriously underused in evaluation. An impediment to progress that she points out is that today, CA (cost analysis) is done in a different way by economists and accountants,[52] and you will need to make clear which approach you are using, or that you are using both—and, if you do use both, indicate when and where you use each. There are also a number of different types of quantitative CA—cost-benefit analysis, cost-effectiveness analysis, cost-utility analysis, cost-feasibility analysis, etc., and each has a particular purpose; be sure you know which one you need and explain why in the report (the definitions in Wikipedia are better than many in Google). The first two require calculation of *benefits as well as costs*, which usually means you have to find, and monetize if important (and possible), the benefits and damages from Checkpoint C2 as well as the more conventional (input) costs.

At a superficial level, cost analysis requires attention to and distinguishing between: (i) money vs. non-money vs. non-monetizable costs; (ii) direct and indirect costs; (iii) both actual and opportunity costs;[53] and (iv) sunk (already spent) vs. prospective costs. It is also

[52] Accountants do 'financial analysis' which is oriented towards an individual's monetary situation, economists do 'economic analysis' which is takes a societal point of view.

[53] Economists often define the costs of P as the value of the most valuable forsaken alternative (MVFA), i.e., as the same as opportunity costs. This risks circularity, since it's arguable that you can't determine the value of the MVFA without knowing what it required you to give up, i.e., identifying *its* MVFA. In general, it may be better to define ordinary costs as the tangible valued resources that were used (not the same as 'required') to cause the evaluand to come into existence (money, time, expertise, effort, etc.), and opportunity costs as another dimension of cost, namely the desiderata you spurned by choosing to create the evaluand rather than the best alternative path to your goals, using about the same resources. The deeper problem is this: the 'opportunity cost of the evaluand'

often helpful, for the evaluator and/or audiences, to itemize these by: developmental stage, i.e., in terms of the costs of: (a) start-up (purchase, recruiting, training, site preparation, etc.); (b) maintenance (including ongoing training and evaluating); (c) upgrades; (d) shut-down; (e) residual (e.g., environmental damage); and/or by calendar time period; and/or by cost elements (rent, equipment, personnel, etc.); and/or by payee. Include use of expended but never utilized value, if any, e.g., social capital (such as decline in workforce morale).

The most common significant non-money costs that are often monetizable are space, time, expertise, and common labor, to the extent that these are not available for purchase in the open market—when they are so available, they can be monetized. The less measurable but often more significant ones include: lives, health, pain, stress (and other positive or negative affects), social/political/personal capital or debts (e.g., reputation, goodwill, interpersonal and other trade and professional skills), morale, energy reserves, content and currency of technical knowledge, and immediate/long-term environmental costs… Of course, in all this, you should be analyzing the costs and benefits of unintended as well as intended outcomes; and, although unintended heavily overlaps unanticipated, both must be covered… The non-money costs are almost never trivial in large program evaluations (and in technology assessment or senior staff evaluation), and very often decisive… The fact that in rare contexts (e.g., insurance suits) some money equivalent of e.g., a life, is treated seriously is not a sign that life is a monetizable value in general i.e., across more than that very limited context,[54] let alone a sign that if we only persevere, cost analysis can be treated as really a quantitative task *or even* as a task for which a quantitative version will give us a useful approximation to the real truth. Both views are categorically wrong, as is apparent if you think about the difference between the value of a particular person's life to their family, vs. to their employer/employees/coworkers, vs. to their profession, and vs. to their friends; *and* the difference between those values as between different people whose lost lives we are evaluating. And don't think that the way out is to allocate different money values to each specific case, i.e., to each person's life-loss for each impacted group: not only will this destroy generalizability but the cost to some of these impactees is clearly still not covered by money, e.g., when a great biochemist or musician dies.

As an evaluator you aren't doing a literal audit, since you're (usually) not an accountant, but you can surely benefit if an audit is available, or being done in parallel. Otherwise, consider hiring a good accountant as a consultant to the evaluation; or an economist, if you're going that way. But even without the accounting expertise, your cost analysis and certainly your evaluation, if you follow the lists here, will include key factors—for decision-making or simple appraisal—usually omitted from standard (financial) auditing practice. In general, professional auditing uses quantitative methods and hence usually involves monetizing everything considered: this is a mistake, as recent professional practice has made clear, since audits can and often must include attention to many qualitative considerations if they

---

is ambiguous; it may mean the value of something else to do the same job, or it may mean the value of the resources if you didn't attempt this job at all. (See my "Cost in Evaluation: Concept and Practice", in *The Costs of Evaluation*, edited by Alkin and Solomon, (Sage, 1983) and "The Economist's Fallacy" in jmde.com, 2007).

[54] The World Bank since 1966 has recommended reporting mortality data in terms of lives saved or lost, not dollars.

are to avoid a parody of comprehensive accountability analysis. Racism or religious bias, for example, are faults even if legal (e.g., because they constrict hiring/promotion of best talent) and should show up as costs or liabilities in a good audit of an organization's value. They are often fatal faults, for legal or ethical reasons, right up there with underfunded pension funds and poor insurance coverage, and should be potential dealbreakers in amalgamation or takeover or funding decisions or any summative evaluation… Also keep in mind that there are evaluations where it is appropriate to analyze benefits (a subset of outcomes) and assets in just the same way, i.e., by type, time of appearance, durability, etc. This is especially useful when you are doing an evaluation with an emphasis on cost-benefit tradeoffs.

Note C3.1: This sub-evaluation (especially item (iii) in the first list) is the key element in the determination of worth.

Note C3.2: If you have not already evaluated the program's risk-management efforts under Process, consider doing—or having it done—as part of this checkpoint.

Note C3.3: Sensitivity analysis is the cost-analysis analog of robustness analysis in statistics and testing methodology, and equally important. It is essential to do it for any quantitative results.

Note C3.4: The discussion of C/A in this checkpoint so far uses the concept of cost-effectiveness in the usual economic sense, but there is another sense of this concept that is of considerable importance in evaluation, in some but not all contexts, and this sense does not seem to be discussed in the economic or accounting literature. (It is the 'extended sense' mentioned in the Resources checkpoint discussion above.) In this sense, efficiency or cost-effectiveness means the ratio of benefits to resources *available* not resources *used.* In this sense—remember, it's only appropriate in certain contexts—one would say that a program, e.g., an aid program funded to provide potable water to refugees in the Haitian tent cities in 2010, was (at least in this respect) inefficient/cost-ineffective if it did not do as much as was *possible* with the resources provided. There may be exigent circumstances that deflect any imputation of irresponsibility here, but the fact remains that the program needs to be categorized as unsatisfactory with respect to getting the job done, if it had unused access to adequate resources to do it. Moral: when you're doing C/A in an evaluation, don't just analyze what was spent but also what was available i.e., could have been spent.

C4. Comparisons

Comparative or relative m/w/s, which requires comparisons, is often extremely illuminating, and sometimes absolutely essential—as when a government has to decide on whether to refund a health program, go with a different one, or abandon the sector to private enterprise. Here you must look for programs or other entities that are alternative ways for getting the same or similar benefits from about the same resources, especially those that use fewer resources. Anything that comes close to this is known as a "critical competitor". Identifying the most important critical competitors is a test of high intelligence, since they are often very unlike the standard competitors, e.g., a key critical competitor for telephone and email communication in extreme disaster planning is carrier pigeons, even today. It is also often worth looking for, and reporting on, at least one other alternative—if you can find one—that is much cheaper but not much less effective ('el cheapo'); and one much stronger

although costlier alternative, i.e., one that produces far more payoffs or process advantages ('el magnifico'), although still within the outer limits of the available Resources identified in Checkpoint B4; the extra cost may still be the best bet. (But be sure that you check carefully, e.g., don't assume the more expensive option is higher—or even of equal—quality because it's higher priced.) It's also sometimes worth comparing the evaluand with a widely adopted/admired approach that is perceived by important stakeholders as an alternative, though not really in the race, e.g., a local icon. Keep in mind that looking for programs 'having the same effects' means looking at the side-effects as well as intended effects, to the extent they are known, though of course the best available critical competitor might not match on side-effects… Treading on potentially thin ice, there are also sometimes strong reasons to compare the evaluand with a demonstrably *possible* alternative, a 'virtual critical competitor'—one that could be assembled from existing or easily constructed components (the next checkpoint is another place where ideas for this can emerge). The ice is thin because you're now moving into the partial role of a program designer rather than an evaluator, which creates a risk of conflict of interest (you may be ego-involved as author (or author-wannabe) of a possible competitor and hence not objective about evaluating it or, therefore, the original evaluand). Also, if your ongoing role is that of formative evaluator, you need to be sure that your client can digest suggestions of virtual competitors (see also Checkpoint D2). The key comparisons should be constantly updated as you find out more from the evaluation of the primary evaluand, especially new side-effects, and should always be in the background of your thinking about the evaluand.

Note C4.1: It sometimes looks as if looking for critical competitors is a completely wrong approach, e.g., when we are doing formative evaluation of a program i.e., with the interest of improvement: but in fact, it's important even then to be sure that the changes made or recommended really do add up, taken all together, to an improvement; so you need to compare version 2 with version 1, *and also* with available alternatives since the set of critical competitors may change as you modify the evaluand.

Note C4.2: It's tempting to collapse the Cost and Comparison checkpoints into 'Comparative Cost-Effectiveness' (as Davidson does, for example) but it's better to keep them separate because for certain important purposes, e.g., fund-raising, you will need the separate results. Other examples: you often need to look at simple cost-feasibility, which does not involve comparisons (but give the critical competitors a quick look in case one of them *is* cost-feasible); or at relative merit when 'cost is no object' (which means 'all available alternatives are cost-feasible, *and* the merit gains from choosing correctly are much more important than cost savings').

Note C4.3: One often hears the question: "But won't the Comparisons checkpoint double or triple our costs for the evaluation—after all, the comparisons needed have to be quite detailed in order to match one based on the KEC?" Some responses: (i) "But the savings on purchase costs may be much more than that;" (ii) "There may already be a decent evaluation of some or several or all critical competitors in the literature;" (iii) "Other funding sources may be interested in the broader evaluation, and able to help with the extra costs;" (iv) "Good design of the evaluations of alternatives will often eliminate potential competitors at trifling cost, by starting with the checkpoints on which they are most obviously vulnerable;" (v) " Estimates, if that's all you can afford, are much cheaper than evaluations, and better than not doing a comparison at all."

Note C4.4: A common mistake in looking for alternatives is to assume the point of view of an impactee, perhaps a typical impactee. Of course, you must cover that point of view, but not only that one. Go back to the 'cost cube' of the Costs checkpoint (C3) and look at the list of 'costs to whom' you constructed there; each of those players has a point of view from which you should think up the relevant alternatives. So, if you're evaluating a college, the first thing you think of is other colleges: then, we hope you also think of online instruction, computer-assisted self-education, etc. But those are just the alternatives for a potential student; think about the point of view of a donor (including taxpayers and their elected representatives if it's a tax-supported institution); of an employer who hires heavily from the graduates of this college; and so on.

C5. Generalizability

Other names for this checkpoint (or something close to, or part of it) are: exportability, transferability, transportability—which would put it close to Campbell's "external validity"— but it also covers sustainability, longevity, durability, and resilience, since these tell you about generalizing the program's merit (and problems) to other *times* rather than (or as well as) other *places* or *circumstances* besides the one you're at (in either direction in time, so the historian is involved.) Note that this checkpoint concerns the sustainability of the *program*, with the effects you have discovered, but not the sustainability of the effects themselves, which is also important but covered under impact (Checkpoint C1).

Although other checkpoints bear on <u>significance</u> (because they are needed to establish that the program has non-trivial benefits), this checkpoint is frequently the most important one of the core five when attempting to determine significance, a.k.a. importance. (The other highly relevant checkpoint for that is C4, where we look at how much better it is compared to whatever else is available; and the final word on that comes in Checkpoint D1, especially Note D1.1.) Under Checkpoint C5, you must find the answers to questions like these: Can the program be used, with similar results, if we use it: (i) with other content; (ii) at other sites; (iii) with other staff; (iv) on a larger (or smaller) scale; (v) with other recipients; (vi) in other climates (social, political, physical); (vii) in other times, etc. An affirmative answer on any of these 'dimensions of generalization' is a merit, since it adds another universe to the domains in which the evaluand can yield benefits (or adverse effects). Looking at generalizability thus makes it possible (sometimes) to benefit greatly from, instead of dismissing, programs and policies whose use at the time of the study was for a very small group of impactees—such programs may be extremely important because of their generalizability.

Generalization to (vii) later times, a.k.a. longevity, is nearly always important (under common adverse conditions, it's durability). Even more important is (viii) sustainability (this is external sustainability, not the same as the internal variety mentioned under Process). It is sometimes inadequately treated as meaning, or as equivalent to, 'resilience to risk.' Sustainability usually requires making sure the evaluand can survive at least the termination of the original funding (which is usually not a risk but a known certainty), and also some range of hazards under the headings of warfare or disasters of the natural as well as financial, social, ecological, and political varieties. Sustainability isn't the same as resilience to risk especially because is must cover future *certainties*, such as seasonal changes in temperature, humidity, water supply—and the end of the reign of the present CEO, political

party in power, or of present funding. But the 'resilience to risk' definition has the merit of reminding us that this checkpoint will require some effort at identifying and then estimating the likelihood of the occurrence of the more serious risks, and costing the attendant losses. Sustainability is sometimes even more important than longevity, for example when evaluating international or cross-cultural developmental programs; longevity and durability refer primarily to the reliability of the 'machinery' of the program and its maintenance, including availability of the required labor/expertise and tech supplies; but are less connotative of external threats such as the '100-year drought,' fire, flood, earthquake, or civil war, and less concerned with 'continuing to produce the same results' which is what you primarily care about. Note that what you're generalizing—i.e., predicting—about these programs is the future (effects) of 'the program in context,' not the mere existence of the program, and so any context required for the effects should be specified, and include any required infrastructure. Here, as in the previous checkpoint, we are making predictions about outcomes in certain scenarios, and, although risky, this sometimes generates the greatest contribution of the evaluation to improvement of the world (see also the 'possible scenarios' of Checkpoint D4). All three show the extent to which good evaluation is a creative and not just a reactive enterprise. That's the good news way of putting the point; the bad news way is that much good evaluation involves raising questions that can only be answered definitively by doing work that you are probably not funded to do.

Note C5.1: Above all, keep in mind that the absence of generalizability has absolutely no deleterious effect on establishing that a program is meritorious, unlike the absence of a positive rating on any of the four other sub-evaluation dimensions. It only affects establishing the extent of its benefits. This can be put by saying that generalizability is a plus, but its absence is not a minus—unless you're scoring for the Ideal Program Oscars. Putting it another way, generalizability is highly *desirable*, but that doesn't mean that it's a *requirement* for m/w/s. A program may do the job of meeting needs just where it was designed to do that, and not be generalizable—and still rate an A+.

Note C5.2: Although generalizability is 'only' a plus, it needs to be explicitly defined and defended. It is still the case that good researchers make careless mistakes of inappropriate implicit generalization. For example, there is still much discussion, with good researchers on both sides, of whether the use of student ratings of college instructors and courses improves instruction, or has any useful level of validity. But any conclusion on this topic involves an illicit generalization, since the evaluand 'student ratings' is about as useful in such evaluations as 'herbal medicine' is in arguments about whether herbal medicine is beneficial or not. Since any close study shows that herbal medicines with the same label often contain completely different substances (and almost always substantially different amounts of the main element), and since most though not all student rating forms are invalid or uninterpretable for more than one reason, the essential foundation for the generalization—a common referent—is non-existent. Similarly, investigations of whether online teaching is superior to onsite instruction, or vice versa, are about absurdly variable evaluands, and generalizing about their relative merits is like generalizing about the ethical standards of 'white folk' compared to 'Asians.' Conversely, and just as importantly, evaluative studies of a nationally distributed reading program must begin by checking the fidelity of your sample (Description and Process checkpoints). This is checking instantiation (sometimes this is part of what is called 'checking dosage' in the medical/pharmaceutical context), the complementary problem to checking generalization.

Note C5.3: Checkpoint C5 is, perhaps more than any others, the residence of prediction, with all its special problems. Will the program continue to work in its present form? Will it work in some modified form? In some different context? With different personnel/clients/-recipients? These, and the others listed above, are each formidable prediction tasks that will, in important cases, require separate research into their special problems. When special advice cannot be found, it is tempting to fall back on the assumption that, absent ad hoc considerations, the best prediction is extrapolation of current trends. That's the best simple choice, but it's not the best you can do; you can at least identify the most common interfering conditions and check to see if they are/will be present and require a modification or rejection of the simple extrapolation. Example: will the program continue to do as well as it has been doing? Possibly not if the talented CEO dies/retires/leaves/burns out? So check on the evidence for each of these possibilities, thereby increasing the validity of the bet on steady-state results, or forcing a switch to another bet. See also Note D2.2.

General Note 7: Comparisons, Costs, and Generalizability are in the same category as values from the list in Checkpoint B5; they are all considerations of certain dimensions of value—comparative value, economic value, general value. Why do they get special billing with their own checkpoint in the list of sub-evaluations? Basically, because of (i) their virtually universal critical importance[55], (ii) the frequency with which one or more are omitted from evaluations when they should have been included, and (iii) because they each involve some techniques of a relatively special kind. Despite their idiosyncrasies, it's also possible to see them as potential exemplars, by analogy at least, of how to deal with some of the other relevant values from Checkpoint B5, which will come up as relevant under Process, Outcomes, and Comparisons.

PART D: CONCLUSIONS & IMPLICATIONS

Now we're beginning to develop the last few payoff components to go into the final report and the executive summary.

D1. Synthesis

You have already done a great deal of the required synthesis of facts with values using the scales and rubrics developed in Checkpoint B5, Values, in order to get the sub-evaluations of Part C. This means you already have an *evaluative profile* of the evaluand, i.e., a bar graph, the simplest graphical means of representing a multidimensional evaluative conclusion, and greatly superior to a table for most clients and audiences. You may even have profiles at two levels of detail. But for some evaluative purposes you need a further synthesis, this time of the bars, the sub-evaluations, because you need to get a one-dimensional evaluative conclusion, i.e., an overall grade—or, if you can justify a more precise quantitative scale, an overall score. For example, you may need to assist the client in choosing the best of several evaluands, which means ranking them, and the fast way to do this is to have each

[55] Of course, ethics (and the law) is critically important, but only as a framework constraint that must not be violated. Outcomes are the material benefits or damage within the ethical/legal framework and their size and direction are the most variable and antecedently uncertain, and hence highly critical findangs from the evaluation. Ethics is the boundary fence; outcomes are what grow inside it.

of them evaluated on a single overall summative dimension. That's easy to say, but it's not easy to justify most efforts to do that, because in order to combine those multiple dimensions into a single one, you have to have a legitimate common metric for them, which is rarely supportable. For example, perhaps the most commonly used example of this synthesis, the amalgamation of student grades into an overall index of academic achievement—the GPA (Grade Point Average)—is certainly invalid, since grades on different courses (or given by different instructors) often represent very different levels of academic achievement and the GPA treats them as equal… At the least, you'll need a supportable estimate of the relative importance of each dimension of merit, and not even that is easy to get. Details of how and when it can be done are provided elsewhere and would take too much space to fit in here.[56] The measurement scale (point of view) of the synthesis, on which the common metric should be based, should usually be the present and future total impact on consumer (e.g., employer, employee, patient, student) or community needs, subject to the constraints of ethics, the law, and resource-feasibility, etc… Apart from the need for a ranking there is very often also a practical need for a *concise presentation of the most crucial evaluative information*. A profile showing merit on the five core dimensions of Part C can often meet that need, without going to a uni-dimensional compression into a single grade… Another possible profile for such a summary would be based on the SWOT checklist widely used in business: Strengths, Weaknesses, Opportunities, and Threats.[57] Sometimes it makes sense to provide both profiles… This part of the synthesis/summary could also include referencing the results against the clients' and perhaps other stakeholders' goals, wants, and (feasible) hopes, e.g., goals met, ideals realized, created but unrealized value, when these are determinable, which can also be done with a profile… But the primary obligation of the public service or even just any professional evaluator is to reference the results to the *needs* of the impacted population, within the constraints of overarching values such as ethics, the law, the culture, debt, etc. Programs are not made into good programs by matching someone's goals, but by doing something worthwhile, on balance. Of course, for public or philanthropic funding, the two should coincide, but you can't just assume they do; in fact, they are all-too-often provably incompatible.

Another popular focus for the overall report is the ROI (return on investment), which is superbly concise, but it's too limited (no ethics scale, no side-effects, no goal critique, usually requires implausible monetizing of outcomes, etc.) The often-suggested 3D expansion of ROI gives us the 3P dimensions—benefits to People, Planet, and Profit—often called the 'triple bottom line.' It's still a bit narrow and we can do better with the 5 dimensions listed here as the sub-evaluations listed in Part C: Process, Outcomes; Costs; Comparisons; Generalizability. A bar graph showing the merit of the achievements on each of these can provide a succinct and insightful profile of a program's value. To achieve it, you will need defensible definitions of the standards you are using on each column (i.e., rubrics), e.g., "An A grade for Outcomes will require…" and there will be 'bars' (i.e., absolute minimum stand-

[56] A 2012 article (Scriven, M. (2012). The logic of valuing. *New Directions for Evaluation, 133*, 17-28) does a better job on this than my previous efforts, which do not now seem adequate as references. E. Jane Davidson's book *Evaluation Methodology Basics* (Sage, 2005) does a very good job of explaining the problem and the best available solutions.

[57] Google provides 6.2 million references for SWOT (@1/23/07), but the top two or three are good introductions.

ards) on several of these, e.g., ethical acceptability on the Process scale, cost-feasibility on the Costs scale. Since it's highly desirable that you get these for any serious program evaluation, getting this 5D summary should not be a dealbreaker requirement. (Another version of a 5D approach is given in the paper "Evaluation of Training" that is online at michael-scriven.info.)… Apart from the rubrics for each relevant value, if you have to come up with an overall grade of some kind, you will need to do an *overall* synthesis to reduce the two-dimensional profile to a 'score' on a single dimension. (Since it may be qualitative, we'll use the term 'grade' for this property.) Getting to an overall grade requires what we might call a meta-rubric—a set of rules for converting profiles that are typically themselves a set of grades on several dimensions—to a grade on a single scale. What we call 'weighting' the dimensions is a basic kind of meta-rubric since it's an instruction to take some of the constituent grades more seriously than others *for some further, 'higher-level' evaluative purpose.* (A neat way to display this graphically is by using the width of a column in the profile to indicate importance.)… If you are lucky enough to have developed an evaluative profile for a particular evaluand, in which each dimension of merit is of equal importance (or of some given *numerical* importance compared to the others), *and* if each sub-grade can be expressed numerically, then you can just calculate the weighted average of the grades. But legitimate examples of such cases are almost unknown, although we often oversimplify and act as if we have them when we don't. For example, we average college grades to get the GPA, and use this in many overall evaluative contexts such as selection for admission to graduate programs. Of course, this oversimplification can be, and frequently is, 'gamed' by students e.g., by taking courses where grade inflation means that the A's do not represent excellent work by any reasonable standard. A better meta-rubric results from including the score on a comprehensive exam, graded by a departmental committee instead of a single instructor, and then giving the grade on this test double weight, or even 80% of the total weight. (A slight problem here is finding a valid comprehensive test—the CLA has been suggested but is fatally flawed.)… Another common meta-rubric in graduate schools is setting a meta-bar, i.e., an overall absolute requirement for graduation, e.g., that no single dimension (e.g., the grade on a particular course or a named subset of crucially important courses) be graded below B-.[58] And of course grad schools almost always build in an overall weight for the quality/difficulty of the undergraduate college from which the grades come, usually a rather arbitrary process (but Yale does it right by getting the weights from an analysis of the Yale grades for every student from the undergraduate college that has so far been admitted).

Note D1.1: One special conclusion to go for, often a major part of determining significance**,** comes from looking at *what was done* against *what could have been done* with the Resources available, including social and individual capital. This is one of several cases where imagination is needed to determine a grade on the Opportunities part of the SWOT analysis; and hence one of the cases that shows evaluation requires creative thinking not just critical thinking (and local knowledge). But remember, establishing possibilities is thin ice territory (see Note C4.1).

Note D1.2: Be sure to convey some sense of the *strength of your overall conclusions* (often

[58] I have recently discovered a little discussion of a related issue in the standard-setting literature, under the topic of 'performance profile' and 'dominant profile' methods; see *Cutscores,* Zieky et al., ETS, 2008, pp. 170-174.

called robustness), which means the combination of: (i) the *net weight of the evidence for* the premises, with (ii) the *probability of the inferences from* them to the conclusion(s), and (iii) the *probability that there is no other relevant evidence*. (These probabilities will often be qualitative.) For example, indicate whether the performance on the various dimensions of merit was a tricky inference or directly observed; did the evaluand clear any bars or lead any competitors 'by a mile' or just scrape over (i.e., use gap-ranking not just ranking[59]); were the predictions involved double-checked for invalidating indicators (see Note C5.2); was the conclusion established 'beyond any reasonable doubt,' or merely 'supported by the balance of the evidence'? This complex property of the evaluation is referred to as 'robustness.' Some specific aspects of the limitations also need statement here e.g., those due to limited time-frame (which often rules out some mid- or long-term follow-ups that are badly needed).

D2. Recommendations, Explanations, Predictions, and Redesigns.

All of these possibilities are examples of the 'something more' approach to evaluation, by contrast with the more conservative 'nothing but' approach, which advocates rather careful restriction of the evaluator's activities to evaluation, 'pure and simple.' These alternatives have analogies in every profession—judges are tempted to accept directorships in companies who may come before them as defendants, counselors consider adopting counselees, etc. The 'nothing more approach' can be expressed, with thanks to a friend of Gloria Steinem, as: 'An evaluation without recommendations (or explanations, etc.) is like a fish without a bicycle.' Still, there are more caveats about pressing for evaluation-separation than with the fish. In other words, 'lessons learned'—of whatever type—should be sought diligently, expressed cautiously, and applied even more cautiously.

Let's start with recommendations. Micro-recommendations—those concerning the internal workings of program management and the equipment or personnel choices/use—often become obvious to the evaluator during the investigation, and are demonstrable at little or no extra cost/effort (we sometimes say they "fall out" from the evaluation; as an example of how easy this can sometimes be, think of copy-editors, who often do both evaluation and recommendation to an author in one pass), or they may occur to a knowledgeable evaluator who is motivated to help the program, because of his/her expert knowledge of this or an indirectly or partially relevant field such as information or business technology, organization theory, systems concepts, or clinical psychology. These 'operational recommendations' can be very useful—it's not unusual for a client to say that these suggestions alone were worth more than the cost of the evaluation. (Naturally, these suggestions have to be within the limitations of the (program developer's) Resources checkpoint, except when doing the Generalizability checkpoint.) Generating these 'within-program' recommendations as part of formative evaluation (though they're one step away from the primary task of formative evaluation which is straight evaluation of the present quality of the evaluand),

[59] Gap-ranking is a refinement of ranking in which a qualitative or quantitative estimate of the size of intervals between evaluands is provided (modeled after the system in horse-racing—'by a head,' 'by a nose,' 'by three lengths,' etc. This is often enormously more useful than mere ranking e.g., because it tells a buyer that s/he can get very nearly as good a product for much less money, and it does not require a ratio scale, or even a strict interval scale—only a partially interval scale.

is one of the good side-effects that may come from using an external evaluator, who often has a new view of things that everyone on the scene may have seen too often to see critically.

On the other hand, macro-recommendations—which are about the disposition or classification of the whole program (refund, cut, modify, export, etc.—which we might also call external management recommendations, or dispositional recommendations)—are usually another matter. These are important decisions serviced by and properly dependent on, summative evaluations, but making recommendations about the evaluand is not intrinsically part of the task of evaluation as such, since it depends on other matters besides the m/w/s of the program, which is all the evaluator normally can undertake to determine.

For the evaluator to make dispositional recommendations about a program's disposition will typically require two extras over and above what it takes to evaluate the program: (i) extensive knowledge of the other factors in the context-of-decision for the top-level ('about-program') decision-makers. Remember that those people are often not the clients for the evaluation—they are often further up the organization chart—and they may be unwilling or psychologically or legally unable to provide full details about the context-of-decision concerning the program (e.g., unable because implicit values are not always recognized by those who operate using them). The correct dispositional decisions often rightly depend on legal or donor constraints on the use of funds, and sometimes on legitimate political constraints not explained to the evaluator, not just m/w/s; and any of these can arise after the evaluation begins and the evaluator is briefed about then-known environmental constraints, if s/he is briefed at all.

Such recommendations will also often require (ii) considerable extra effort e.g., to evaluate each of the other macro-options. Key elements in this may be trade secrets or national security matters not available to the evaluator, e.g., the true sales figures, the best estimate of competitors' success, the extent of political vulnerability for work on family planning, the effect on share prices of withdrawing from this slice of the market. This elusiveness also often applies to the macro-decision makers' true values, with respect to this decision, which are quite often trade or management or government secrets of the board of directors, or select legislators, or perhaps personal values only known to their psychotherapists.

So it is really a quaint conceit of evaluators to suppose that the m/w/s of the evaluand are the only relevant grounds for deciding how to dispose of it; there are often entirely legitimate political, legal, public-perception, market, and ethical considerations that are at least as important, especially in toto. So it's simply presumptuous to propose macro-recommendations as if they follow directly from the evaluation: they almost never do, even when the client may suppose that they do, and encourage the evaluator to produce them. (It's a mistake I've made more than once.) If you do have the required knowledge to infer to them, then at least be very clear that you are doing a different evaluation in order to reach them, namely an evaluation of the alternative options open to the disposition decision-makers, by contrast with an evaluation of the evaluand itself. In the standard program evaluation, but not in the evaluation of various dispositions of it, you can sometimes include an evaluation of the internal choices available to the program manager, i.e., recommendations for improvements.

There are a couple of ways to 'soften' recommendations in order to take account of these

hazards. The simplest way is to preface them by saying, "Assuming that the program's disposition is dependent only on its m/w/s, it is recommended that…" A more creative and often more productive approach, advocated by Jane Davidson, is to convert recommendations into options, e.g., as follows: "It would seem that program management/staff faces a choice between: (i) continuing with the status quo; (ii) abandoning this component of the program; (iii) implementing the following variant [here you insert your recommendation] or some variation on this." The program management/staff is thus invited to adopt and become a co-author of an option, a strategy that is often more likely to result in implementation than a mere recommendation from an outsider.

Many of these extra requirements for making macro-recommendations—and sometimes one other—also apply to providing explanations of success or failure. The extra requirement is possession of the correct (not just the believed) logic or theory of the program, which typically requires more than—and rarely requires less than—state-of-the-art subject-matter expertise, both practical and 'theoretical' (i.e., the scientific or engineering account), about the evaluand's inner workings (i.e., about what optional changes would lead to what results). A good automobile mechanic has the practical kind of knowledge about cars that s/he works on regularly, which includes knowing how to identify malfunction and its possible causes; but it's often only the automobile engineer who can give you the reasons why these causal connections work, which is what the demand for explanations will usually require. The combination of these requirements imposes considerable, and sometimes enormous, extra time and research costs which has too often meant that the attempt to provide recommendations or explanations (by using the correct program logic) is done at the expense of doing the basic evaluation task well (or even getting to it at all), a poor trade-off in most cases. Moreover, getting the explanation right will sometimes be absolutely impossible within the 'state of the art' of science and engineering at the moment—and this is not a rare event, since in many cases where we're looking for a useful social intervention, no-one has yet found a plausible account of the underlying phenomenon: for example, in the cases of delinquency, addiction, autism, serial killing, ADHD. In such cases, what we need to know is whether we have found a cure—complete or partial—since we can use that knowledge to save people immediately, and also, thereafter, to start work on finding the explanation. That's the 'aspirin case'—the situation where we can easily, and with great benefit to many sufferers, evaluate a claimed medication although we don't know why it works, and don't need to know that in order to evaluate its efficacy. In fact, until the evaluation is done, there's no success or failure for the scientist to investigate, which vastly reduces the significance of the causal inquiry, and hence the probability/value of its occurrence.

It's also extremely important to realize that macro-recommendations will typically require the ability to predict the results of the recommended changes in the program, at the very least in this specific context, which is something that the program logic or program theory (like many social science theories) is often not able to do with any reliability. Of course, *procedural* recommendations in the future tense, e.g., about needed further research or data-gathering or evaluation procedures, are often possible—although typically much less useful.

'Plain' predictions are also often requested by clients or thought to be included in any good evaluation. (e.g., Will the program work reliably in our schools? Will it work with the rec-

ommended changes, without staff changes?) and are often very hazardous.[60] Now, since these are reasonable questions to answer in deciding on the value of the program for many clients, you have to try to provide the best response. So read *Clinical vs. Statistical Prediction* by Paul Meehl and the follow-up literature, and the following Note D2.1, and then call in the subject matter experts. In most cases, the best thing you can do, even with all that help, is not just to pick what appears to be the most likely result, but to give a range from the probability of the worst possible outcome (which you describe carefully) to that of the best possible outcome (also described), plus the probability of the most likely outcome in the middle (described even more carefully).[61] On rare occasions, you may be able to estimate a confidence interval for these estimates. Then the decision-makers can apply their choice of strategy (e.g., minimax—minimizing maximum possible loss) based on their risk-aversiveness, assuming they are trained in it.

Although it's true that almost every evaluation is in a sense predictive, since the data it's based on is yesterday's data but its conclusions are put forward as true today and probably tomorrow, there's no need to be intimidated by the need to predict; one just has to be very clear what assumptions one is making and how much evidence there is to support them.

Finally, a new twist on 'something more' that I first heard proposed by John Gargani and Stewart Donaldson at the 2010 AEA convention, is for the evaluator to do a redesign of a program rather than giving a highly negative evaluation. This is a kind of limit case of recommendation, and of course requires an extra skill set, namely design skills. The main problem here is role conflict and the consequent improper pressure: the evaluator is offering the client loaded alternatives, a variation on 'your money or your life.' That is, they suggest that the world will be a better place if the program is redesigned rather than just condemned by them, which is probably true; but these are not the only alternatives. The evaluator might instead recommend the redesign, and suggest calling for bids on that, recusing his or her candidacy. Or they might just recommend changes that a new designer should incorporate or consider.

Note D2.1: Policy analysis, in the common situation when the policy is being considered for

---

[60] Evaluators sometimes say, in response to such questions, Well, why wouldn't it work—the reasons for it doing so are really good? The answer was put rather well some years ago: "...it ought to be remembered that there is nothing more difficult to take in hand, more perilous to conduct, or more uncertain of success, than to take the lead in the introduction of a new order of things. Because the innovator has for enemies all those who have done well under the old conditions, and lukewarm defenders in those who may do well under the new." (Niccolo Machiavelli (1513), with thanks to John Belcher and Richard Hake for bringing it up recently (*PhysLrnR*, 16 Apr 2006)).

[61] In PERT charting (PERT = Program Evaluation and Review Technique), a long-established approach to program planning that emerged from the complexities of planning the first submarine nuclear missile, the Polaris, the formula for calculating what you should expect from some decision is: {Best possible outcome + Worst Possible outcome + 4 x (Most likely outcome)}/6. It's a pragmatic solution to consider seriously. My take on this approach is that it only makes sense when there are good grounds for saying the most likely outcome (MLO) is very likely; there are many cases where we can identify the best and worst cases, but have no grounds for thinking the intermediate case is more likely other than the fact it's intermediate. Now that fact does justify some weighting (given the usual distribution of probabilities), but the coefficient for the MLO might then be better as 2 or 3.

future adoption, is close to being program evaluation of future (possible) programs (a.k.a., ex ante, or prospective (virtual) program evaluation) and hence necessarily involves all the checkpoints in the KEC including, in most cases, an especially large dose of prediction. (A policy is a 'course or principle of action' for a certain domain of action, and implementing it typically is or involves at least one program.) Extensive knowledge of the fate of similar programs in the past is then the key resource, but not the only one. It is also essential to look specifically for the presence of indicators of future change in the record, e.g., downturns in the performance of the policy in the most recent time periods, intellectual or motivational burn-out of principal players/managers, media attention, the probability of personnel departure for better offers, the probability of epidemics, natural disasters, legislative 'counter-revolutions' by groups of opponents, general economic decline, technological breakthroughs, or large changes in taxes or house or market values, etc. If, on the other hand, the policy has already been implemented, then we're doing historical (a.k.a. ex post, or retrospective program evaluation) and policy analysis amounts to program evaluation without prediction, a much easier case.

Note D2.3: Evaluability assessment is a useful part of good program planning whenever it is required, hoped, or likely that evaluation could later be used to help improve as well as determine the m/w/s of the program to assist decision-makers and fixers. It can be done well by using the KEC to identify the questions that will have to be answered eventually, and thus to identify the data that will need to be obtained; and the difficulty of doing that will determine the evaluability of the program as designed. Those preliminary steps are, of course, exactly the ones that you have to go through to design an evaluation, so the two processes are two sides of the same coin. Since everything is evaluable, to some extent in some contexts, the issue of evaluability is a matter of degree, resources, and circumstance, not of absolute possibility. In other words, while everything is evaluable, by no means is everything evaluable to a reasonable degree of confidence, with the available resources, in every context. (For example, the atomic power plant program for Iran after 4/2006, when access was denied to the U.N. inspectors) As this example illustrates, 'context' includes the date and type of evaluation, since, while this evaluand is not evaluable prospectively with any confidence, in 4/06—since getting the data is not feasible, and predicting sustainability is highly speculative—historians will no doubt be able to evaluate it retrospectively, because we will eventually know whether that program paid off, and/or brought on an attack.

Note D2.3: Inappropriate expectations The fact that clients often expect/request explanations of success or shortcomings, or macro-recommendations, or impossible predictions, is grounds for educating them about what we can definitely do vs. what we can hope will turn out to be possible. Although tempting, these expectations on the client's part are not an excuse for doing, or trying for long to do, and especially not for promising to do, these extra things if you lack the very substantial extra requirements for doing them, especially if that effort jeopardizes the primary task of the evaluator, viz. drawing the needed type of evaluative conclusion about the evaluand. The merit, worth, or significance of a program is often hard to determine; it (typically) requires that you determine whether and to what degree and in what respects and for whom and under what conditions and at what cost it does (or does not) work better or worse than the available alternatives, and what all that means for all those involved. To add on the tasks of determining how to improve it, explaining why it works (or fails to work), now and in the future, and/or what one should do about supporting or exporting it, is simply to add other tasks, often of great scientific and/or

managerial/social interest, but quite often beyond current scientific ability, let alone the ability of an evaluator who is perfectly competent to evaluate the program. In other words, 'black box evaluation' should not be used as a term of contempt since it is often the name for a vitally useful, feasible, and affordable approach, and frequently the only feasible one. And in fact, most evaluations are of partially blacked-out boxes ('grey boxes') where one can only see a little of the inner workings. This is perhaps most obviously true in pharmacological evaluation, but it is also true in every branch of the discipline of evaluation and every one of its application fields (health, education, social services, etc.). A program evaluator with some knowledge of parapsychology can easily evaluate the success of an alleged faith-healer whose program theory is that God is answering his prayers, without the slightest commitment to the truth or falsehood of that program theory.

Note D2.4: Win-win recommendations It is almost always worth thinking about the possibility of a recommendation, consistent with your evaluation findings, which will provide benefits to all stakeholders; or as next best, beneficial to some of them with no harm to the others (i.e., one that is Pareto-optimal). Creating such an option is a truly creative activity, not something that is entailed by your evaluation findings; but it is sometimes a great service to the client, who of course has not previously had that task, since they have not previously had the results you have turned up in the evaluation. Note, however, that even a win-win solution will not automatically be the best option from the client's point of view: there are expenses involved in getting it accepted by the other stakeholders, and there may be sub-optimal solutions that do better for the client. This places some ethical responsibility on the evaluator to convince the client of the social superiority of the win-win solution.

Note D2.5: Proformative evaluation Finally, there are extreme situations in which the evaluator does have a responsibility—an ethical responsibility—to move beyond the role of the evaluator, e.g., because it becomes clear, early in a formative evaluation, either that (i) some gross improprieties are involved, or that (ii) certain actions, if taken immediately, will lead to very large increases in benefits, and it is clear that no-one besides the evaluator is going to take the necessary steps. The evaluator is then obliged to be proactive, and we can call the resulting action whistle-blowing in the first case (see Note D4.1 below), and proformative evaluation in both cases, this being a cross between formative evaluation and proactivity. While macro-recommendations by evaluators require great care, proactivity requires even greater care.

D3. Responsibility and Justification

If either can be determined, and if it is appropriate to determine them. Some versions of accountability that stress the accountability of people do require this—see examples below. Allocating blame or praise requires extensive knowledge of: (i) the main players' knowledge-state at the time of key decision making; (ii) their resources and responsibilities for their knowledge-state as well as their actions; as well as (iii) an ethical analysis of their options, and of the excuses or justifications they (or others, on their behalf) may propose. Not many evaluators have the qualifications to do this kind of analysis. The "blame game" is very different from evaluation in most cases and should not be undertaken lightly. Still, sometimes mistakes are made, are demonstrable, have major consequences, and should be pointed out as part of an evaluation; and sometimes justified choices, with good or bad effects, are made and attacked, and should be praised or defended as part of an evaluation. The evaluation of accidents is an example: the investigations of aircraft crashes by the Na-

National Transportation Safety Board in the US are in fact a model example of how this can be done; they are evaluations of an event with the added requirement of identifying responsibility, whether it's human or natural causes. (Operating room deaths pose similar problems but are often not as well investigated.)

Note D3.1: The evaluation of disasters, (a misleadingly narrow title) recently an area of considerable activity, typically involves one or more of the following six elements: (i) an evaluation of the magnitude, nature, physical details, and social/ecological significance of the event; (ii) an evaluation of the extent of preparedness; (iii) an evaluation of the immediate response; (iv) an evaluation of the totality of the relief efforts until termination; (v) an evaluation of the lessons learned (lessons learned should be a part of each of the evaluations done of the response); and (vi) an evaluation of subsequent corrective/preventative action. All six should involve some evaluation of human responsibility and any appropriate allocation of praise/blame. Early efforts (c. 2005) referred to as general approaches to the 'evaluation of disasters' appear not to have distinguished all of these and not to have covered all of them, although it seems plausible that all should have been covered even if one's interest is only to minimize the impact of later disasters.

D4. Report & Support

Now we come to the task of conveying the conclusions in an appropriate way, to appropriatae groups/people, at appropriate times and locations. This is a very different task from—although frequently confused with—handing over a semi-technical report to the client at the end of the study, the paradigm for typical research studies of the same phenomena. Evaluation reporting for a single evaluation may require, or benefit from, radically different presentations to different audiences, at different times in the evaluation: these may be oral or written, long or short, public or private, technical or non-technical, graphical or textual, scientific or story-telling, anecdotal and personal or barebones. And this phase of the evaluation process should include post-report help, e.g., handling questions when they turn up later as well as immediately, explaining the report's significance to different groups including users, staff, funders, and other impactees, and even reacting to later program or management or media documents or actions allegedly reporting or based on the results or implications of the evaluation.

This extension of the research-report paradigm may, in practice or in prospect, involve proactive creation and depiction in the primary report of various possible scenarios of interpretations and associated actions that are, and—this contrast is extremely helpful—are not, consistent with the findings. Essentially, this means doing some problem-solving for the clients, that is, advance handling of difficulties they are likely to encounter with various audiences. In this process, a wide range of communication skills is often useful and sometimes vital, e.g., audience 'reading', use and reading of body language, understanding the multicultural aspects of the situation and the cultural iconography and connotative implications of types of presentations and response.[62] There should usually be an explicit effort to identify 'lessons learned,' failures and limitations, cost details if permitted, and explaining 'who evaluates the evaluators.' Checkpoint D4 should also cover getting the results (and incidental knowledge findings) into the relevant databases, if any; possibly but not neces-

[62] The 'connotative implications' are in the sub-explicit but supra-symbolic realm of communication, manifested in—to give a still-relevant example—the use of gendered or genderless language.

sarily into the information sky (which includes not only clouds) via journal publication (with careful consideration of the cost of subsidizing these if necessary for potential readers of the publication chosen); recommending creation of a new database or information channel (e.g., a newsletter) where beneficial; and dissemination into wider channels if appropriate, e.g., through presentations, online posting, discussions at scholarly meetings, or in hardcopy posters, graffiti, book, blogs, wikis, tweets, and in movies (yes, fans, remember—UTube is free). There is now a much-needed and valuable sub-specialty in evaluation devoted to the graphical aspects of report design, but the previous paragraph refers to a dozen aspects of reporting that go far beyond that, just as the KEC refers to a dozen aspects of content that go far beyond reporting the 'facts of the case.'

Note D4.1: Whistle-blowing It should be understood by every evaluator that there are circumstances in which evaluators are ethically and even legally required go public with some or all of even preliminary findings, sometimes even without first giving the client a chance to take action themselves. In some jurisdictions, for example, it appears to be a legal requirement that if the evaluation turns up some indication—far short of strong evidence—of illegal activity such as sexual abuse, it must be immediately reported.[63] The ethical requirement may be somewhat less stringent, but if time is of the essence (i.e., repetitions of the improprieties are possible), mere reporting one's grounds for concern to the client may not be enough. To avoid complicity, i.e., shared moral responsibility, there must at least be a warning attached to the notification, in some physically recorded form, that the evaluator will have to be convinced that adequate action—not merely investigation, plus cessation, prevention, and punishment as appropriate, *but also external reporting* to appropriate authorities or individuals—has been taken within a specified time, or the evaluator will be obliged to take direct action (normally this would be notification of the police and/or other appropriate civil and/or military authorities).

Delivery of this warning should be certified and the receipt for its delivery filed; and similar proof must be required that the client has in fact reported the violation before the evaluator will be convinced that reporting has occurred. If the violation is continuing, or even if it is merely possible that it may recur within a specified short time, then that interval must be reduced to eliminate that possibility; and if it is still possible within any significant interval, the evaluator probably should *immediately report* any actual, or suspected serious, violation with only a simultaneous report of so doing to the client. It is clear that this lesson was not learnt in or from the recent Wall Street cases, or the Pennsylvania State University case, or the Roman Catholic Church cases. In each of these cases, either managers or auditors (a species of evaluator) failed to report gross and recurrent impropriety. In some areas where evaluators work, it is a requirement for continued funding that there be training of all staff in the responsibility to blow the whistle, and to have a named staff member who is responsible for the training and compliance; this may be a good example for all evaluation consultancies to follow. It should be stressed that this recommendation is not a legal opinion and is sometimes *less than the legal requirement* on the evaluator—which may be immediate reporting of suspicions to the police (etc.)—and may therefore be a risky extension of the evaluator's obligation to the client, defensible, if at all, only if the transgression is historic

[63] Sometimes this obligation is defined only for 'mandated reporters' e.g., child service providers, but sometimes that restriction is not stated; in any case, an evaluator might be construed as included in those addressed by the statute.

or trivial. While it is not appropriate to put any statement like this in one's standard contract, since ethicality is always presumed, it is highly desirable if not essential to avoid signing a contract that always requires notification of the client before any reporting elsewhere.

D5. Meta-evaluation

This is the evaluation of an evaluation or evaluations—including evaluations based on the use of this checklist—in order to identify their strengths/limitations/other uses: they may be done in a formative or summative or ascriptive mode. Meta-evaluation should always be done, as a separate quality control step(s), for three main reasons, one psychological, one (closely linked) methodological, and one ethical. The psychological reason is that a mass of convergent multidisciplinary research has recently clarified the extraordinary extent to which reasoning and problem-solving by groups is typically a *very* large jump better than solo work, something we've always believed in general, so the work of a solo evaluator or an evaluation team simply has to treat getting someone else's take on their work with the same respect as a basic evidence-gathering procedure. The methodological principle that incorporates our long commitment to the qualitative view that more thinkers make better thinking is built into the universal scientific quality control process, peer review. In most professional evaluation that process is bypassed since the evaluator's work goes directly to the client, whereas in most scientific research—which is aimed at publication—the professional report first goes to reviewers. We need to close that gap. The ethical reason is that evaluators sell their work as quality-enhancing or monitoring, and are implicitly rejecting the legitimacy of that claim if they do not apply it to their own work. (Caveat emptor is good advice, albeit cynical, but not an excuse.) Meta-evaluation should also be attempted by the evaluator, both during and after the completion of the work, despite the limitations of self-evaluation, because it forces the evaluator to simulate looking at his or her own work through the eyes and checkpoints of a skilled audience; typically, this requires running through the evaluation using a different approach from the one originally adopted.

The primary criteria of merit for meta-evaluations are: (i) validity, at a contextually adequate level[64]; (ii) credibility (for select stakeholders, especially funders, regulatory agencies, and usually also program staff and community representatives); (iii) utility[65], including cost-feasibility and comprehensibility (usually to clients, audiences, *and* stakeholders) of both the main conclusions about the m/w/s of the evaluand, and the recommendations, if any; and also any utility arising from (iv) generalizability e.g., of novel methodological or interventional approaches; (v) comparative cost-effectiveness, which goes beyond utility to require consideration of alternative possible evaluation approaches, especially

[64] This means, for example, that when balance of evidence is all that's called for (e.g., because a decision has to be made fast) it's an irrelevant complaint that proof of the conclusion beyond any reasonable doubt was not supplied.

[65] Utility is usability and not actual use, the latter—or its absence—being at best a probabilistically sufficient but not necessary condition for the former, since it may have been very hard to use the results of the evaluation, and utility/usability requires (reasonable) ease of use. Failure to use the evaluation may be due to base motives or stupidity or an act of God and hence is not a valid criterion for lack of evaluation merit.

cheaper/faster/simpler ones; (vi) robustness, i.e., the extent to which the evaluation is immune to variations in context, measures used, point of view of the evaluator, small measurement errors, etc, as far as is possible without major losses on other merit criteria; and (vii) ethicality/legality/propriety, which includes such matters as avoidance of conflict of interest,[66] unnecessary intrusiveness, maximization of benefits to, and protection of the rights of, human subjects—of course, these affect credibility, but is not exactly the same since the ethicality may be deeply flawed even though superficially and legally squeaky-clean.

Note that there are secondary criteria of merit for evaluations that should be addressed in meta-evaluations: one of the most important of these is competency-building of evaluation skills of the client, and perhaps also of the program staff and other audiences/stakeholders. These become primary criteria of merit if you're doing an evaluation of an evaluation for significance rather than merit, since utilization is a primary measure of outcome importance: a good example is Berliner's famous meta-evaluation of most educational evaluation, for failing to see that economic status (in particular, poverty) is a more important controlling variable of learning gains than virtually all interventions, in particular No Child Left Behind.[67] (It should have been picked up under Generalizability.)

There are several ways to go about meta-evaluation. You and later another meta-evaluator can: (a) apply the KEC and/or PES and/or GAO list—preferably one or more of these that was not used to do the evaluation—to the evaluation itself. (Then, for example, the Cost checkpoint in the KEC addresses the cost of the evaluation rather than the cost of the program); and/or (b) use a special meta-evaluation checklist (there are several available, including the one sketched in the previous paragraph, which is sometimes called the Meta-Evaluation Checklist or MEC[68]); and/or (c) if funds are available, replicate the evaluation, doing it in the same way, and compare the results; and/or (d) do the same evaluation using a different methodology for data-gathering/analysis and compare the results. It's highly desirable to employ more than one of these approaches.

Note D5.1: As mentioned, failure to use an evaluation's results is often due to bad, perhaps venal, management, and so can never be regarded as a criterion of low utility without further evidence.[69] In fact, literal or direct use are not concepts clearly applicable to evaluations without recommendations (e.g., most done by historians), a category that includes many important, complete, and influential evaluations since evaluations are not in themselves recommendations. 'Due consideration *or* utilization' is a better generic term for the

---

[66] There are a number of cases of conflict of interest of particular relevance to evaluators, e.g., formative evaluators who make suggestions for improvement and then do a subsequent evaluation (formative or summative) of the same program, of which they are now co-authors—or rejected contributor-wannabes—and hence in conflict of interest. See also Note 5.9 below for another extremely important case.

[67] "Our Impoverished View of Educational Reform" *Teachers College Record* 2.2005. He was viciously attacked for this analysis, to the point that citing/discussing the article by an applicant for funding was sometimes used as a criterion for dismissal of funding applications.

[68] An earlier version is online at michaelscriven.info

[69] This view is rejected by the maestro of utilization, Michael Quinn Patton, who insists that failure to use is always a sign of flawed evaluation.

ideal response to a good evaluation; it is an essential part of 'due diligence,' i.e., professionalism in management.

Note D5.2: Evaluation impacts often occur years after completion and often occur even if the evaluation was rejected completely when submitted. Evaluators too often give up their hopes of impact too soon, and meta-evaluators must be cautious about this.

Note D5.3: Help with utilization beyond submitting a report should at least have been offered—see Checkpoint D4.

Note D5.4: Look for contributions from the evaluation to the knowledge management system of the organizations to which the client belongs or which apply to the client (e.g., Lessons Learned databases), as well as the client's own KMS; if they lack one, the evaluator should have encouraged the client to recommend creating one; if this was unsuccessful, the m/evaluator should consider making the recommendation herself/himself.

Note D5.5: Since effects of the evaluation are not usually regarded as effects of the program, it follows that although an empowerment evaluation should produce substantial gains in the staff's knowledge about and tendency to use or improve evaluations, that's not an effect of the *program* in the relevant sense for an evaluator. Also, although that valuable outcome is an effect of the *evaluation*, it can't compensate for low validity or low external credibility—two of the most common threats to empowerment evaluation—since training the program staff is not a primary criterion of merit for evaluations, although it's a desirable outcome (a side-effect, sub-species human capital, sub-sub-species competency enhancement).

Note D5.6: Similarly, one common non-money cost of an evaluation—disruption of the work of program staff—is not a bad effect of the program. But it is one of the items that should always be picked up in a meta-evaluation. (Of course, it's minimal in goal-free evaluation, since the (field) evaluators do not talk to program staff; this is one advantage of GFE.) Careful design (of program plus evaluation) can sometimes bring these evaluation costs near to zero or ensure that there are benefits that more than offset the cost.

Note D5.7:[70] It's obvious that self-evaluation is a weak type of meta-evaluation for reasons that undercut both credibility and validity. But self-selection of a meta-evaluator is only one stage better than self-evaluation—although it's a big step up, it's still 'internal meta-evaluation.' It's better to nudge the client to get another evaluator, 'external' to your own effort, to do the meta-evaluation. As I've suggested elsewhere, for a big evaluation, you could suggest that the client ask the current president of the AEA to suggest candidates; in smaller projects, provide the client with the online address of the AEA's list of professional evaluators to choose from by using their own cvs for indicators of merit/probity or the location of other people for inquiries.

Note D5.8:[71] Perhaps the worst problem in checking credibility is the hardest one to check: the matter of true independence. Apart from the usual cases of financial or social bias, it is seriously compromised when: (i) the evaluation design is specified in detail by the client, in

[70] Thanks to Daniel Stufflebeam for a reminder that led me to add this note to the preceding comments.

[71] Thanks to Robert Picciotto for a comment that inspired this note.

or subsequent to the RFP; (ii) the client requires frequent reports on progress to the client or an agent of the client (i.e., opportunities to influence design and formulation of findings); (iii) the evaluator does very lengthy and/or repeated and/or collaborative formative evaluations for the same client (this appears to include Michael Quinn Patton's 'developmental evaluation'), thus risking co-option or co-authorship (or, if their suggestions are not implemented, 'rejected suitor' negative bias). Meta-evaluation is the best antidote to this loss of independence. And, for internal evaluators, it's even more essential, although considerable diplomacy is required to get approval for it. The big difficulty with this advice is that most of this interaction also serves the admirable purpose of keeping the client informed about what you're doing, and making *appropriate* modifications to the design when they point out gaps in your design as a way of providing them with facts and conclusions they rightly need. So you have to walk a tightrope. Or you can bypass the problem by just doing what the client wants and making clear that this involves research but is not an evaluation.

Note D5.9: Certain evaluation designs should be regarded as raising a red flag, calling for intensive scrutiny by meta-evaluators. For example, any design involving ongoing comparison/control groups instantly creates a conflict of interest for the evaluator, with respect to their treatment of control group subjects, since the evaluator using such an approach is highly motivated to prevent these subjects from (i) getting alternative forms of treatment, or even (ii) just dropping out, either of which may be much better for the subject than continuing to receive no treatment or a placebo. Even if an allowance for attrition has led to an offsetting-oriented increased size for the control group, if there is—or was—substantial attrition the remaining subjects may no longer be a valid match to the experimental group. Again, some interviewing of control group subjects is almost unavoidable, and an ombudsperson may need to be provided for them and indeed all subjects involved in the evaluation. A control group design is often severely handicapped by this consideration, offsetting some or all of its comparative advantages (in some situations) over other designs such as interrupted time series.

General Note 8: The KEC How-to Guide. 1. Describe the evaluand and its context (the infrastructure on which it depends) in detail, including the process (steps and sequence) involved in its delivery (but not its goals or the underlying theory of the intervention); determine its effects including side-effects. 2. Identify the dimensions of merit for this evaluand and its effects, in this context. 3. Classify the dimensions for importance (i.e., weight them), usually into three groups for Very Important, Important, and Slightly Important (and possibly a different set for each of your different audiences). 4. Specify an appropriate grading scale (i.e., a set of rubrics) for performance on each dimension e.g., by setting the 'cut scores' including bars and steps. 5. For each dimension, using relevant prescriptions, preferences, relevant comparisons (i.e., performance of competitive evaluands), needs, and ideals; repeat iff[72] appropriate for your various audiences. 6. Identify the empirical indicators that will provide evidence for an estimate of degree of success or failure on the relevant dimension. 7. Measure or estimate achievement on each indicator. 8. Amalgamate the indicator scores to get an achievement level and hence a grade on each dimension. 9. Iff required, amalgamate the dimensional grades to get an overall grade, using the weights of each dimension. 10. Apply any overall bars. 11. Report results.

[72] Iff = if and only if

General Note 9: References and Acknowledgments The explanatory remarks here should be regarded as first approximations to the content of each checkpoint. More detail on some of them and on items mentioned in them can be found in one of the following: (i) the *Evaluation Thesaurus*, Michael Scriven, (4th edition, Sage, 1991), under the checkpoint's name; (ii) in the references cited there; (iii) in the online *Evaluation Glossary* (2006) at evaluation.wmich.edu, partly written by this author; (iv) in the best expository source now, E. Jane Davidson's *Evaluation Methodology Basics* (Sage, 2005 and 2e, 2013 (projected)); (v) in later editions of this document, at michaelscriven.info… The present version of the KEC itself is, however, in most respects very much better than the ET one, having been substantially refined and expanded in more than 85 'editions' (i.e., widely circulated or online posted revisions), since its birth as a two-pager around 1971—30+ since early 2009—with much appreciated help from many students and colleagues, including: Chris Coryn, Jane Davidson, Rob Brinkerhoff, Christian Gugiu, Nadini Persaud,[73] Emil Posavac, Liliana Rodriguez-Campos, Daniela Schroeter, Natasha Wilder, Lori Wingate, and Andrea Wulf; with a thought or two from Michael Quinn Patton's work and from contributors to Evaltalk, including Stan Capela… More suggestions and criticisms are very welcome—please send to me at mjscriv1@gmail.com, with KEC as the first word in the title line. (Suggestions after 3.28.11 that require significant changes are rewarded, not only with an acknowledgment but a little prize: usually your choice from my list of duplicate books.)

[30,810 words @ July 24th, 2013]

[73] Dr. Persaud's detailed comments have been especially valuable: she was a CPA before she took a doctorate in evaluation. But there are not as many changes in the cost section as she thinks are called for, so she is not to blame for any remaining faults.