KEY EVALUATION CHECKLIST

Intended for use in designing and evaluating programs, plans, and policies; for writing evaluation reports on them; for assessing their evaluability; & for evaluating evaluations of them.

Michael Scriven
Edition of 23may06

GENERAL NOTE A: Throughout this document, “evaluation” is taken to mean the determination of merit, worth, or significance (abbreviated m/w/s); and “evaluand” means whatever is being evaluated. This is a tool for the working evaluator, so knowledge of some terms from evaluation vocabulary is assumed, e.g., formative, goal-free, ranking; their definitions can be found in the Evaluation Thesaurus (4e, Sage, 1991), or in the Evaluation Glossary, which will be online at evaluation.wmich.edu, in June 06. Merely for simplification, we talk throughout about ‘programs’ rather than ‘programs, plans, or policies, or evaluations of them, or designs for their evaluation, or reports on their evaluation or their evaluability.’

GENERAL NOTE B: The KEC can also be used, with care, for: (i) the evaluation of products (for which purpose it was originally designed—although since completely rewritten, and then revised and circulated at least 40 times); and (ii) the evaluation of organizational units such as departments, consultancies, associations, and for that matter, (iii) hotels, restaurants, and hamburger joints; and, with some use of the imagination and a heavy emphasis on the ethical values involved, for (iv) some tasks in the evaluation of personnel.

GENERAL NOTE C: This is an iterative checklist, not a one-shot checklist, i.e., you should expect to go through it several times, even for design purposes, since discoveries or problems that come up under later checkpoints will often require modification of what was entered under earlier ones (and no rearrangement of the order will avoid this). 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 in evaluation by various authors, under the listing for the Checklist Project at evaluation.wmich.edu.

GENERAL NOTE D: It is not always helpful to simply list here what allegedly needs to be done; when the reasons for the recommended coverage (or exclusions) are not obvious, especially when the issues are highly controversial (e.g., Checkpoint 12). In such cases I have also provided brief summaries of the reasons for the position taken.

GENERAL NOTE E: The determination of merit, of worth, and of significance, the triumvirate values of evaluation, rely 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, brings in Cost, i.e., Checkpoint 8.

PART A: PRELIMINARIES

These are essential parts of a report, but may seem to have no relevance to the design and execution phases of an evaluation; however, it turns out to be quite useful to begin all one’s thinking about an evaluation by role-playing the situation when you 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 of starting a log on the project as soon as it becomes a possibility.
I. Executive Summary
The aim is to summarize the results and not just the investigatory process. Typically this should be done without even mentioning the process whereby you got them, unless the methodology is especially notable. (In other words, avoid the pernicious practice of using the exec summary as a ‘teaser.’) Through the whole process of evaluation, keep asking yourself what the overall summary is going to say, based on what you have learned so far, and how it relates to the client’s and stakeholders’ and audiences’ needs for, and prior, information; this helps you to focus on what still needs to be done to learn about what matters most. The exec summary should usually be a selective summary of Checkpoints 11-15, and should not run more than one or at most two pages if you expect it to be read by executives.

II. Preface
Now is the time to identify and define in your notes: (i) the client (the person who officially requests, and, if it’s a paid evaluation, pays for or arranges payment for) the evaluation, and—you hope—the same person to whom you report—if not, try to arrange this, to avoid crossed wires in communications; (ii) the prospective audiences (for the report); (iii) the stakeholders (those with a substantial vested interest in the outcome of the evaluation); and (iv) who else will see, will have the right to see, or should see, (a) the results, and/or (b) the raw data. Get clear in your mind your actual role—internal evaluator, external evaluator, a hybrid (e.g., an outsider brought in for a limited time to help the staff with setting up and running evaluation processes), an evaluation trainer (sometimes called an empowerment evaluator), etc. Each of these roles has different risks, and is viewed with different expectations by your staff and colleagues, the clients, the staff of the program being evaluated, et al.

And now is the time to get down to the nature and details of the job or jobs, as the client sees them—or encourage the client to clarify their position on the details that they have not yet thought out. Can you determine the source and nature of the request, need, or interest, leading to the evaluation: for example, is the request, or the need, for an evaluation of worth—which involves serious cost analysis—rather than of merit; or of significance; or of more than one of these? Exactly what are you supposed to be evaluating (what’s the evaluand): how much of the context is to be included; how many of the details are important (enough to replicate the program elsewhere, or 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 success and failure (one type of analytic evaluation); or the quality or the contribution of each of its components; (another two types of analytic evaluation); or perhaps additionally the accuracy of the client’s theory of how the components work (i.e., an evaluation of, and/or creation of, a program theory—probably in addition to an evaluation of the program)? Is the evaluation to be formative, summative, ascriptive (‘ascriptive’ means simply for the record, or for accountability purposes, rather than to support any decision), or more than one of these? Should it yield grades, ranks, scores, profiles, or apportionments? Are recommendations, faultfinding, or predictions requested, or expected, or feasible (see Checkpoint 12)? Is the client really willing and anxious to learn from faults or is that 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.) Have they thought about post-report help with interpretation and utilization? (If not, offer it without extra charge—see Checkpoint 12 below.)

NOTE II.1: It’s best to discuss these issues about what it’s feasible to evaluate, and clarify your commitment, only after doing a quick trial run through the KEC, so ask for a little time to do this, overnight if possible (see NOTE 12.3 near the end of the KEC). Be sure to note later any subsequently negotiated, or imposed, changes in any of the preceding. And here’s where you give acknowledgments/thanks . . . so it’s likely to be the last section you revise.
Ill. Methodology

Now that you’ve got the questions straight, how are you going to find the answers? Examples of this kind of question that has to be answered here: Do you have adequate domain expertise? If not, how will you add it to the evaluation team (via consultant(s), advisory panel, full team membership, split contract)? Can you use control or comparison groups to determine causation of supposed effects/outcomes? If there’s to be a control group, can you randomly allocate subjects to it? How will you control differential attrition; cross-group contamination, other threats to internal validity? Can you double- or single-blind the study? If a sample is to be used, how selected, and if stratified, how stratified? If none of these, how will you determine causation (of effects by the evaluand); and, depending on the job, you may also need to determine the contribution to the effects from individual components of the evaluand—how will you do that? How will you find side-effects? Will/should the evaluation be goal-based or goal-free? To what extent participatory or collaborative; what standards for selecting partners/assistants? If judges are to be involved, what reliability and bias controls will you need (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 11)? Identify, as soon as possible, other investigative procedures for which you’ll need expertise, time, and staff in this evaluation: observations, participant observations, logging, journaling, audio/photo/video recording, tests, simulating, role-playing, surveys, interviews, focus groups, text analysis, library/online searches/search engines, etc.; and data-analytic procedures (stats, cost-analysis, modeling, expert consulting, etc.), plus reporting techniques (text, stories, plays, graphics, freestyle drawings, stills, movies, etc.), and their justification (may need to allocate time for a lit review on some of these methods).

Most important of all, how are you going to identify, particularize, validate, and incorporate all significantly relevant values and associated standards, since each of those steps is essential for supporting an evaluative conclusion? In the light of this and all the preceding, set out the ‘logic of the evaluation,’ i.e., a general description and justification of its total design, for the report; the same process will also generate a list of needed resources for your planning and budgeting efforts.

Part B: Foundations

This is the set of investigations that lays out the context and nature of the program, which you’ll need in order to start specific work on the key dimensions of merit in Part C.

1. 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 program; and put matched numbers on every pair that looks as if they may interact. Identify (i) any ‘upstream stakeholders’—and their stakes—other than 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); (ii) enabling legislation—and any other relevant legislation/policies—and add any legislative/executive/practice or attitude changes after start-up; (iii) the underlying rationale, a.k.a. 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 lit review on similar interventions (including ‘fugitive studies’ (those not published in standard media), and the Internet (consider checking the ‘invisible web,’ and the latest group and blog/wiki with the specialized search engines needed to access either)); (v) previous evaluations, if any; (vi) their impact, if any.

2. Descriptions and Definitions

Record any official descriptions of program + components + context/environment, + (client’s) program logic, but don’t assume they are correct; develop a correct and complete description of the first three,
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 a correct 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 find some obvious flaws in the client’s effort here and can point that out at some appropriate time.) Get a detailed description of goals/mileposts for the program (if not operating in goal-free mode). Explain meaning of any ‘technical terms,’ i.e., those that will not be in prospective audiences’ vocabulary, e.g., ‘hands-on’ 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, including some discussion of whether or not they can be justified. 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.

3. **Consumers (Impactees)**
Consumers 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 coworkers, who are impacted via ripple effect). Program staff are also impactees, but we usually keep them separate by calling them the midstream impactees, because the obligations to them are very different and much weaker in most kinds of program evaluation (their welfare is not the raison d’etre of the program). The funding agency, taxpayers, and political supporters, who are also impactees in some sense, 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). 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 &/or modus operandi methodology.) 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 impacted populations, which may suggest some ways to expand, modify, or contract the program when/if you spend time on Checkpoint 11 (Synthesis: sub-issue, what might have been done that was not done), and Checkpoint 12 (recommendations). The possible variations are of course constrained by the resources available—see next checkpoint.

NOTE 3.1 Do not use or allow the use of the term ‘beneficiaries’ for impactees, since it begs the question of whether all the effects of the program are beneficial.

4. **Resources (a.k.a. “strengths assessment”)**
This checkpoint 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. 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, the comparisons, and directions for improvement that an evaluation considers, hence crucial in Checkpoint 9 (Comparisons), Checkpoint 11 (Synthesis, for achievement), Checkpoint 12 (Recommendations), and Checkpoint 13 (Responsibility). Particularly for #9 and #13, 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 involved? As a cross-check, and as a complement, consider all constraints on the program, including legal and fiscal. This checkpoint is the one that covers individual and social capital available to the program; there is also social
capital used by the program (part of its Costs), and sometimes social capital benefits produced by the program (part of the Outcomes).¹

5. Values

The values of primary interest in typical program evaluations are for the most part not mere personal preferences of the impactees, unless those overlap with their needs and the community/societal needs and committed values, e.g., those in the Bill of Rights and the wider body of law. But preferences as such are not irrelevant, and on some issues, e.g., surgery options, they are 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. While there are intercultural and international differences of great importance in evaluating programs, virtually all of the values listed here are highly regarded in all cultures; the differences are generally in their precise interpretation, the exact standards based on each of them, and the relative weight assigned to them, and taking those differences into account is fully allowed for in the approach here.

Begin by identifying the relevant general values for evaluating this evaluand in these circumstances. Some of these follow simply from understanding the nature of the evaluand (these are often called criteria of merit or dimensions of merit). For example, if it’s a health program, then the extent or reach or breadth of its impact (i.e., the size of the impactee population), and the impact’s depth (magnitude or extent or duration of average effect) are leading values. Other primary criteria in such a case include low adverse eco-impact, physical ease of access/entry, broader age/gender range for acceptance. Then turn to the following list to find other relevant values. Validate them as current for the present project, and as contextually supportable.² Now you need to establish the particular standards (the specific values, a.k.a., the ‘cut scores’ if the dimension is measurable) that will identify the levels of quality relevant for this evaluation. What

¹ Individual human capital is the sum of the physical and intellectual abilities, skills, powers, experience, health, and attitudes a person has acquired; these blur into their—and their community’s—social capital, which includes their relationships (‘social networks’) and their share of any latent attributes that their group acquires over and above the sum of their individual human capital (those that depend on interactions with others). For example, the extent of the trust or altruism that pervades a group, be it family, army platoon, corporation, or other organization, is part of the value the group has acquired, a survival-related value that they (and perhaps others) benefit from having in reserve. (Example of non-additive social capital: the skills of football or other team members that only pay off for anyone when they are part of 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 has been taken over to mean “employees,” and that is not what we are 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 include simplified and erroneous as well as diverse uses—few dictionaries have yet caught up with these terms.

² 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, contrary to the myth of value-free social science. The ones in the list here are usually fully supportable to the degree needed by the evaluator, by appeal to publicly available evidence and careful reasoning. Bringing them into consideration is what distinguishes evaluation from 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 cost to implement? In other words, the most important practical questions, for most people looking at programs (the same approach applies to product evaluation, personnel evaluation, etc.).
Given is a fairly detailed account of the logic of integrating facts and values, the core methodological skill in the evaluator’s repertoire.

This means, first, attaching a grading scale to each dimension, e.g., in evaluating students for admission to a graduate school program, the standards for admission set by a department might include a minimum score on the GRE quantitative test of 550. That would be represented by a C grade on that dimension of applicant merit, meaning Acceptable. In evaluating that graduate program, however, the evaluator might have reasons for treating that definition of admissibility as too low a standard for the purposes of choosing programs that are to be rated as ‘amongst the best in the country’, and give the program a D for using 550 as a C (D meaning Less than satisfactory). So s/he is setting standards at different points on this performance dimension from those used by the program itself; a common occurrence, since the two purposes for grading are completely different—one is grading students for admission and the other is grading programs for ‘outstanding in the US.’

All of this is the first stage in doing what we can call the process of identifying “stars, bars, and steps” for our listed values. 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 other values 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 importance might be expressed qualitatively (e.g., major/medium/minor, or by letter grades); or quantitatively (e.g., points on a five or ten point scale, or—usually a better method—by the allocation of 100 ‘weighting points’ across the set of values); or relatively, in terms of an ordering of their importance. The “bars” are absolute minimum standards for acceptability, if any: that is, they are minima on a particular scale, that 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 the bar,’ e.g., an F on the GRE quantitative scale may be acceptable if overshadowed by other virtues for entry into a creative writing program. 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. The grades, bars, and stars (weights), are often loosely included under what is called ‘standards.’ In serious evaluation, it may also be appropriate to establish “steps” i.e., points or intervals on measured dimensions of merit where the weight changes, if the mere starts don’t provide enough effect. (Bars and steps may be fuzzy as well as precise.)

Three values are of such general importance that they receive independent checkpoint status and are listed in the next section: cost (reduction), relative superiority, and generalizability. Their presence in the KEC brings the number of types of values considered up to a round 20.

At least check the following values for relevance and look for others:

3 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 a credit, on one of which the candidate must score high, but no particular one is required. In other words the applicant has to have talent, but a wide range of talents are acceptable. This might be described as a case where there is a group bar, i.e., a ‘floating’ bar on a group of dimensions, which must be cleared on one of them. )It can be exhibited in the list of dimensions of merit by bracketing the group of dimensions in the margin, and stating the height of the floating bar in an attached note.

4 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 respect for which these were specified (i.e., they ‘clear these bars’), but may be so close to 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 (or overall) 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.)
(i) criteria of merit from standard usage (e.g., breadth and depth for a health program) and from the contextual meaning of an ideal or excellent evaluand (e.g., ‘good shuttle bus service for clinic patients in Harlem’).

(ii) needs of the impacted population via a needs assessment (distinguish performance needs from treatment needs, met needs from unmet needs, and meetable needs from ideal but impractical or impossible-with-present-resources needs (consider the Resources checkpoint)). The needs are matters of fact, not values in themselves, but in any context that accepts the most rudimentary ethical considerations, those facts are value-imbued.

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

(iv) legal requirements

(v) ethical requirements (overlaps with legal), usually including (reasonable) safety, confidentiality (and perhaps anonymity), for all impactees;

(vi) personal and organizational goals/desires, if not in conflict with ethical/legal/practical considerations (unless you’re doing a goal-free evaluation). These are usually much 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 an evaluative conclusion about e.g., which college to attend;

(vii) fidelity to alleged specs (a.k.a. “authenticity,” “adherence,” “implementation,” 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 6.

(viii) sub-legal but still important legislative preferences;

(ix) professional standards (i.e., standards set by the profession) of quality that apply to the evaluand;

(x) expert refinements of any standards lacking a formal statement like those in (ix);

(xi) historical/traditional/cultural standards;

(xii) scientific merit (or worth or significance);

(xiii) technological m/w/s;

(xiv) marketability;

(xv) political merit, if you can establish it BRD, which means cross-party agreement;

(xvi) risk, meaning the probability of failure/loss (particularly important in evaluating plans (pre-formative evaluation) and in formative evaluation, but also relevant in summative and ascriptive evaluation when the latter two are done prospectively). There is an option of including this under personal preferences (as a positive or negative preference, for gamblers and risk-aversive individuals), but it is better to consider it separately since it is often a matter on which a community must decide together. Note that risk is normally taken to be 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’);

(xvii) last but not least—resource economy (i.e., how low-impact is the program with respect to available resources of money, space, time, labor, contacts, expertise and the eco-system. Note that this is not the same as low-cost (covered separately in Checkpoint 8) in the normal currencies (money and non-money), but of absolute diminution of available resources (in some framework, which might range from a single family to a country). This could be included un-

---

5 Since one of the steps in the evaluation 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 the KEC.
der an extended notion of 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. It partly overlaps with Checkpoints 10, so watch for double-counting.

Fortunately, bringing these standards to bear\(^6\) is less onerous than it may appear, since many of these values will be unimportant or only marginally important in many particular cases, although each one will be crucially important in other cases. And doing all this values-analysis will be easy to do sometimes, very hard on other occasions; it will often require expert advice and/or impactee/stakeholder advice. And, 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. Hence you need to be very careful not to assume that you have to generate a ranking from the evaluation, since if that’s not required or useful, you can often avoid settling the issue of relative weights, or at least avoid any precision in settling it, by simply doing a grading or profiling display (showing the merit on all relevant dimensions of merit, in a bar-graph of the evaluand(s)).

NOTE 6.1: 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 any materialize). Some of these values will occur to you only after you find the side-effects (Checkpoint 7), 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.

**PART C: SUBEVALUATIONS**

Each of the following five core dimensions requires both: (i) a fact-finding phase, followed by (ii) the process of combining the facts with whatever values bear on those facts, which yields the subevaluation. In other words, Part C requires the completion of five separate inferences from (i) to (ii), i.e., from What’s So? to So What?—e.g., from ‘the effects were measured as XXX,’ to ‘the effects were extremely beneficial’ (or ‘insignificant in this context,’ etc. The first two of these checkpoints can use all relevant values from Checkpoint 5 and bear most of the load in determining merit; the next three are defined in terms of specific values of great general importance, named in their heading, and particularly relevant to worth (Checkpoint 8 and 9) and significance (Checkpoints 9 and 10).

**6. Process**

This is the assessment of the m/w/s of everything that happens or applies before true outcomes emerge, especially: the vision, design, planning and operation of the program, from the justification of its goals (if you’re not operating in goal-free mode) which may have changed or be changing; through design provisions for reshaping under environmental or political or fiscal duress (including planning for worst-case outcomes), to its justification of its program logic, and its implementation fidelity (i.e., degree of implementation of the supposed archetype program, if any; this is also called “authenticity”, “adherence”, “compliance”); also check accuracy of official name (if it’s descriptive), subtitle, or description of program (e.g., “an inquiry-based science education program for middle school”, “raising to proficiency level”, “critical thinking training program”); management (especially the arrangements for evaluative feedback, and support for learning from it, and from any mistakes discovered in other ways); activities; procedures; the general program learning process (e.g., constant updating training to cope with changes in the operational environment, staff aging, skill pool, new technology); attitudes/values; morale.

\(^6\) ‘Bringing them to bear’ involves: (a) identifying the relevant ones, (b) specifying (meaning, 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.
You may perhaps also, if you’ve undertaken to cover this in any detail, check under this heading the quality of the original logic of the program and its current logic (both the current official version and the possibly different one implicit in the operations/staff behavior). It is not appropriate to try to determine whether the model is correct in scientific fact unless you have specifically undertaken that kind of (usually ambitious and sometimes unrealistically ambitious) analytic evaluation of the program theory. You only need to decide whether it is correct as a description of what appears to be actually going on, i.e., don’t, if possible, undertake to decide whether the alleged ‘phosphate reduction unit’ does reduce phosphate content, only whether the stuff that needs to be cleaned up actually goes to this unit (and of course whether the entire plant produces the needed results for the ecosystem, measured by examination of the ecosystem, under Outcomes).

Process evaluation may also include the evaluation of what are often called “outputs,” (usually taken to be ‘intermediate outcomes’ that are developed en route to ‘true outcomes,’ a.k.a. longer-term results or impact) such as knowledge, skill, and 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, there will have been substantial learning that produces a gain in individual or social capital, that must be regarded as a tangible gain for the program and for the intervention. It’s not terribly important whether you call it process or output or short-term outcome, as long as you find it, estimate it, and record it—only once, though.

Two important examples of why process is an essential element in program evaluation, despite the common emphasis on outcomes-based evaluation: (i) gender or racial prejudice in selection/promotion/treatment of staff is a common unethical practice that must be checked for, and comes under process; (ii) in evaluating health programs that involve medication, ‘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., through 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 evaluative conclusions about treatment efficacy.

7. Outcomes

Evaluation of (good and bad) effects on consumers: these must include direct and indirect, intended and unintended, immediate and short-term and long-term. (These are, roughly speaking, the focus of Campbell’s ‘internal validity.’) Finding outcomes cannot be done by hypothesis-testing methodology, because often the most important effects are unanticipated ones (the two main ways to find such side-effects are goal-free evaluation and using the mythical “Book of Causes”\textsuperscript{8}), and because determining the m/w/s of the effects—this is the result you have to get out of this sub-evaluation—is often the hard part, not determining whether there are any and what they are. Immediate outcomes (e.g., 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; they are normally covered under Checkpoint 6. But note that some true out-

\textsuperscript{7} The ‘immediate’ effects of a program are not just the first effects that occur after the program starts up, but also the effects on anticipators who react to the announcement of, or secret intelligence about, the forthcoming start of the program.

\textsuperscript{8} The Book of Causes shows, when opened at the name of a 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 circumstances; and (ii) on the right (recto) side, all the things which it can cause: that’s the side you need to guide the search for side-effects. Since the BofC is only a virtual book, you have to create these pages, using all your resources such as accessible expertise 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 social capital of their guild.
comes (i.e., results that are of major significance, whether or not intended) can occur during the process and should be considered here. (Long-term results are sometimes called ‘effects’ (or ‘true effects’ or ‘results’) and the totality of these is often referred to as the ‘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, much of which is normally not counted as outcomes, partly because they are gains in latent ability (capacity, potentiality), not necessarily in observable achievements or goods.

Sometimes, not always, it’s useful and feasible to provide explanations of success/failure in terms of components/context/decisions, e.g., when evaluating a statewide consortium of training programs for firemen dealing with toxic fumes, it’s often easy enough to identify the successful and failing ones, maybe even to identify the key to success as particular features that are to be found in and only in the successful programs. To do this usually does not require the identification of the true operating logic/theory of program operation (by contrast with (i) the original official, (ii) the current official, (iii) the implicit, logics or theories). Also see Checkpoint 12 below.

Given that the most important outcomes may have been unintended, even unanticipated, it’s worth distinguishing between side-effects (which affect the target population and possibly others) and side-impacts (i.e., impacts, of the kind that were intended or of other kinds, on non-targeted populations).

Remember that success cases may require their own treatment, regardless of average improvement (since the benefits in those cases alone may justify the cost of the program); if so, so too will the failure cases. Keep the “triple bottom-line” approach in mind, i.e., as well as (i) conventional outcomes, also look for (ii) social (include social capital), and (iii) environmental, outcomes. And always comment on the risk aspect of outcomes, which is likely to be valued very differently by different stakeholders. Finally, don’t forget (i) the effects on the program staff, good and bad; e.g., lessons and skills learned, and the effects of stress; and (ii) the pre-program effects: 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. Looking at these carefully is sometimes called preformative evaluation; and the impact group, as previously mentioned, is sometimes called the ‘anticipators.’

NOTE 7.1 It is usually true that evaluations have to be completed long before some of the main outcomes have occurred, let alone been inspected carefully. This leads to a common practice of depending heavily on predictions of total 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.

8. Costs

This checkpoint is crucial in the determining worth (or, in one sense, value) by contrast with plain merit (or quality). It requires attention to both: (i) money and non-money costs; (ii) direct and indirect costs; and (iii) both actual and opportunity costs. They should be itemized by: developmental stage, i.e., (a) start-up, (b) maintenance, (c) upgrade, (d) shutdown costs; and/or by calendar time period; and by components (rent, equipment, personnel, etc.); and by payee, all of these whenever relevant and possible. Include use of expended but never realized value, if any, e.g., social capital, (e.g., decline in workforce morale). The most common non-money costs are space, time, expertise, and common labor (when these are not available in the market—if they are so available, they become money costs); PLUS the less measurable ones—stress, political and personal capital (e.g., reputation and goodwill), and immediate environmental impact, which are rarely fully coverable by money. You aren’t doing an audit, since you’re (usually) not an accountant, but you can surely benefit if one is available or being done in parallel. But even

---

without the accounting expertise, your cost analysis and certainly your evaluation, if you follow the KEC, is reaching key factors omitted from standard auditing practice.

NOTE 8.1 This sub-evaluation is the key element in the determination of worth.

9. Comparisons

The key comparisons should be constantly updated as you find out more about the evaluation and the evaluand, and constantly in the background of all your thinking about the evaluand. You look for entities that are alternative ways for getting the same or similar benefits from about the same resources; anything that does this is known as a “critical competitor” (‘criticomp’ here). It is often worth looking for, and reporting on, at least one other alternative—if you can find one—that is much cheaper but not so much less effective (‘el cheapo’); and one much stronger although costlier alternative, i.e., one that produces many more payoffs or process advantages (‘el magnifico’), although still within the outer limits of the available Resources identified in Checkpoint 4. It’s also sometimes worth comparing the evaluand with a widely adopted/admired approach that is perceived as an alternative, though not really in the race, e.g., a local icon. Keep in mind that ‘having the same effects’ covers side-effects as well as intended effects, though the best available criticomp 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 criticomp’—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 role of a program designer rather than an evaluator, which creates a risk of conflict of interest (you may be ego-involved as author of a competitor and hence not objective about evaluating the original evaluand). So, if you have an ongoing role as formative evaluator, you need to be sure that your client can digest suggestions of virtual competitors (see also Checkpoint 12).

NOTE 9.1: Comparisons are often extremely illuminating, and sometimes absolutely essential—as when a government has to decide on a health program. It looks as if they are sometimes a completely wrong approach, as when we are doing formative evaluation for 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 1 with version 2.

NOTE 9.2: It’s tempting to collapse the Cost and Comparison checkpoints into ‘Comparative Cost-Effectiveness’ but it’s better to keep them separate because for certain important purposes you will need the separate results. For example, you often need to know cost-feasibility, which does not involve comparisons; or relative merit when ‘cost is no object’ (which means ‘all relevant alternatives are cost-feasible, and merit gains from choosing correctly are much more important than cost savings).

10. Generalizability (a.k.a. exportability, transferability, transportability; roughly the same as Campbell’s ‘external validity;” but also covers sustainability, longevity, durability, resilience.)

Although other checkpoints bear on it, this checkpoint is frequently the most important one of the core five when attempting to determine significance. (The final word on that comes in Checkpoint 11.) Here you must find the answers to questions like these: Can the program be used with similar results if we use it: with other content, at other sites, with other staff, on a larger (or smaller) scale, with other recipients, in other climates (social, political, physical), 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. Looking at generalizability thus makes it possible (sometimes) to benefit from, instead of dismissing, programs and policies whose first use was in a very small region of payoff. Generalization to later times, a.k.a. longevity, is nearly always important (under adverse conditions, it’s durability). Even more important is sustainability (a.k.a. ‘resilience to risk’), which requires making sure the evaluand can survive at least the termination of direct funding, and also some magnitude of hazards under the headings of warfare and disasters of the natural as well as financial, social, ecological, and political varieties. This is even more important than longevity, for example when evaluating international or cross-cultural developmental
programs. Note that what you’re generalizing---i.e., predicting---about them is the future of ‘the program in context,’ and so any required context should be specified, and include any required infrastructure. Here, as in the last sentence in the previous checkpoint (9), we are making predictions about outcomes in these 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 14. All three show the extent to which good evaluation is a creative and not just a reactive enterprise.

NOTE 10.1: Generalizability is highly desirable, but that doesn’t mean that generalization is highly desirable. It is still the case that good researchers make careless mistakes of inappropriate generalization; for example, there is still much discussion of whether the use of student ratings improves instruction. But 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 since most student rating forms are invalid or uninterpretable for more than one reason, the essential foundation for the generalization is shifting sand. Similarly, investigations of whether online teaching is superior to onsite instruction 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, the complementary problem to checking generalization.

GENERAL NOTE F: Comparisons, Costs, and Generalizability are related to values from the list in Checkpoint 5; why do they get special billing with their own checkpoint in the list of sub-evaluations? Basically, because of (i) their virtually universal importance, (ii) the frequency with which they 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 5, which will come up under Process, Outcomes, and Comparisons.

PART D: CONCLUSIONS & IMPLICATIONS

11. Synthesis
Here we combine the basics of Part B with the subevaluations of Part C and any other synthesis of empirical results with values, into an overall evaluation, i.e., at least into a profile (a.k.a., bar graph or histogram, the simplest graphical means of representing a multidimensional conclusion, and incomparably superior to a table for most clients and audiences); or into a unidimensional conclusion—a grade or a rank, if that is required (usually much harder). The focus (point of view) of the synthesis should usually be the present and future impact on consumer and community needs, subject to the constraints of ethics and the law (and feasibility, etc.—i.e., showing the clearance over any bars on other relevant values), but usually there should also be some conclusion(s) that are aimed at the client’s and other stakeholders’ needs for concise evaluative information. This means referencing the results against their goals, wants, or hopes (if feasible), e.g., goals met; ideals realized; created but unrealized value, when determinable. But the overarching obligation of the evaluator is to reference the results to the needs of the impacted population.

NOTE 11.1 One special conclusion to go for, often a major part of determining significance: look at what was done against what could have been done with the Resources available, including social and individual capital.

12. (possible) Recommendations and Explanations
The general principle governing these can be expressed, with thanks to Gloria Steinem, as: An evaluation without recommendations (or explanations) is like a fish without a bicycle. Still, there are more caveats than for the fish. In other words, ‘lessons learned’ should be sought diligently, but very selectively. Micro-recommendations—those concerning the internal workings of program management and equipment choices/use—often become obvious to the evaluator during the investigation, and are demonstrable at little or no extra cost/effort (we say they “fall out” from the evaluation), or they may occur to the smart 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 technology, organization theory, 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 made within the limitations of the Resources checkpoint.) Generating these ‘within-program’ recommendations as part of formative evaluation (though they’re not the primary task of formative evaluation which is straight evaluation of the present state of the evaluand) is one of the good side-effects of using an external evaluator, who often has a new view of things everyone on the scene may have seen too often to see critically. But macro-recommendations—which are about the disposition of the whole program (refund, cut, modify, export, etc.—what we might call external management recommendations)—are usually another matter. These are the decisions serviced by summative evaluations, but making recommendations about them is not part of the task of evaluation as such.

For the evaluator to make recommendations about these decisions will typically require: (i) extensive knowledge of the context-of-decision for the external (‘about-program’) decision-makers. Remember that those people are often not the clients for the evaluation—they’re often further up the organization chart—and they may be unwilling or psychologically unable to provide full details about the context-of-decision concerning the program—implicit values are not always recognized by those who operate using them; (ii) considerable extra effort; and (iii) knowledge of all the internal (‘within-program’) management options available, and the probable outcomes if they are implemented, since the real value of the program includes its value if the optimal choices are made at all these ‘inner’ choice-points. Key elements in this list—notably the effects tree for the program manager’s choices—are often not available to anyone, including the most expert of experts, in the present state of the art of scientific knowledge about that particular evaluand. Or they exist but are trade secrets or national security matters not available to the evaluator. This is also often the case with the true values, with respect to this decision, of the ‘disposition decision-makers,’ 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. It is 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, 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. 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 which you can sometimes include an evaluation of the internal choices available to the program manager, i.e., recommendations for improvements.

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 practical expertise about its inner workings (i.e., about what optional changes would lead to what results). A good automobile mechanic has the latter knowledge about cars that s/he works on regularly; but only the automobile engineer 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 impose considerable, and sometimes enormous, extra time and research costs which has often meant that the attempt to provide recommendations or explanations (correct pro-
gram 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. Note that macro-recommendations typically also require the ability to predict the results of recommended changes in the program, in this specific context, something that the program logic or program theory (like many social science theories) is often not able to do with any reliability. However, procedural recommendations in the future tense, e.g., about needed further research or data-gathering or evaluation procedures, are often possible—although much less useful.

NOTE: 12.1 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 recommended changes, without staff changes?) and are often very hazardous.¹⁰ 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 (i.e., Note 12.2), and then call in the subject matter experts. In most cases, the best thing you can do, 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)¹¹. 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 (minimax, etc.) based on their risk-aversiveness.

NOTE 12.2 Policy analysis, in the common situation when the policy is a proposal being considered for future adoption, is close to being program evaluation of future (possible) programs (a.k.a., ex ante, or prospective program evaluation) and hence necessarily involves all the checkpoints in the KEC including a large dose of prediction. (A policy is a 'course or principle of action' for a certain domain of action and implementing it produces a 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, the probability of their departure for better offers, epidemics, natural disasters, legislative ‘counter-revolutions’ by groups of opponents, general economic decline 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 simpler case.

NOTE 12.3 Evaluability assessment is a useful part of good program planning, whenever it is required or hoped that evaluation will be used to help improve as well as determine the m/w/s of the program. 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. And that is, of course, exactly the process 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 evaluality is a matter of degree and circumstance, not of absolute possibility. So, while everything is evaluable, not everything is evaluable to a reasonable degree of confidence, with the available resources, in every context. For example, the atomic power plant

¹⁰ Evaluators sometimes say, in response to such questions, Well, why wouldn’t it work,—the reasons for doing it 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)

¹¹ In PERT charting, 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, a pragmatic solution to consider. (PERT = Program Evaluation and Review Technique.)
program for Iran in 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 4/2006—since getting the data is not feasible, and predicting sustainability is highly speculative—historians will no doubt be able to evaluate it retrospectively, since we will eventually know whether that program paid off or brought on an attack.

NOTE 12.4 **Inappropriate expectations** The fact that clients often expect/request explanations of shortcomings, or macro-recommendations, and inappropriate 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 to do, and especially not for promising to do, these extra things if you lack the very stiff 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 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, and what all that means for those involved. To add on the tasks of determining how to improve it, explaining why it works (or fails to work), and what one should do with supporting and exporting it, are simply 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 and affordable approach, and frequently the only feasible one. And in fact, most evaluations are of partially blacked-out boxes. This is perhaps most obviously true in pharmacological evaluation, but it is also true in every branch of the discipline and every one of its applied fields. 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 12.5 Finally, there are extreme situations in which the evaluator does have an ethical responsibility to move beyond the role of the evaluator, e.g., because it becomes clear early in a formative evaluation either that some gross improprieties are involved, or that 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, and *proformative* evaluation in the second, a cross between formative evaluation and proactivity. If macro-recommendations by evaluators require great care, proactivity requires even greater care.

13. **(possible) Responsibility and Justification**

If any can be determined, and if appropriate (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, as well as (iii) an ethical analysis of their options, and the excuses or justifications they 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 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 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 13.1 The evaluation of disasters, recently an area of considerable activity, usually involves some or all of the following five components: (i) an evaluation of the extent of preparedness, (ii) an evaluation
of the immediate response, (iii) an evaluation of the totality of the relief efforts until termination, (iv) an evaluation of the lessons learned; these lessons learned should be a part of each of the evaluations done of the response, and (v) an evaluation of corrective/preventative action. All five involve some evaluation of responsibility and sometimes the allocation of praise/blame. Recent efforts (2005) referred to as general approaches to the ‘evaluation of disasters’ appear not to have distinguished all of these and hence not to have covered all of them, although it seems plausible that all should have been covered.

14. Report and Support

Now we come to the task of conveying the conclusions in an appropriate way, and at appropriate times and locations. This is a very different task from—although frequently confused with—handing over a semi-technical report at the end of the study, the paradigm for typical research studies. Evaluation reporting for a single evaluation may require 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 step in 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, other impactees. This in turn may involve creation and depiction of various possible scenarios of interpretations and associated actions that are and—the contrast is extremely helpful—are not consistent with the findings. Essentially, this means doing some problem-solving for the client, 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.12 There should usually be an explicit effort to identify ‘lessons learned,’ failures and limitations, costs if requested, and ‘who evaluates the evaluators’? Checkpoint 14 should also cover getting the results (and incidental knowledge findings) into the relevant databases, if any; possibly but not necessarily into the information ocean via journal publication (with careful consideration of the cost of purchasing these 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, and movies.

15. Metaevaluation

This is evaluation of an evaluation, including those based on the use of this checklist, in order to identify its strengths/limitations/other uses. This should always be done, as a separate quality control step(s), as follows: (i) to the extent possible, by the evaluator, certainly—but not only—after completion of the final draft of any report; and (ii) whenever possible also by an external evaluator of the evaluation (a meta-evaluator). The primary criteria of merit for evaluations are: (i) validity; (ii) utility (usually to clients, audiences, and stakeholders); (iii) credibility (to select stakeholders, especially funders, regulatory agencies, and usually also to program staff); (iv) cost-effectiveness; and (v) ethicality. There are at least five ways to go about this. You and then the metaevaluator can: (a) apply the KEC list to the evaluation itself (the Cost checkpoint then addresses the cost of the evaluation, not the program, and so on for all checkpoints); and/or (b) use a special meta-evaluation checklist (there are several available, including the one earlier in this paragraph); and/or (c) replicate the evaluation, doing it in the same way, and compare the results; and/or (d) do the evaluation using a different methodology and compare the results; and/or (e) apply the Program Evaluation Standards to it. It’s highly desirable to employ more than one of these approaches.

12 The ‘connotative implications’ are in the sub-explicit but supra-symbolic realm of communication, manifested in, e.g., the use of gendered or genderless language.
NOTE 15.1 Utility is usability and not actual use, the latter—or its absence—being at best an indicator of the former, since it may have been very hard to use the results of the evaluation, and utility/usability means (reasonable) ease of use.

NOTE 15.2 Literal or direct use is not a term clearly applicable to evaluations without recommendations, a category that includes many important, complete, and influential evaluations: ‘due consideration’ or ‘utilization’ is a better generic term for the ideal response to a good evaluation.

NOTE 15.3 Evaluation impacts often occur years after completion and often occur even if the evaluation was rejected completely when submitted.

NOTE 15.4 Help with utilization beyond submitting the report should at least have been offered—see Checkpoint 14.

NOTE 15.5 Look for contributions from the evaluation to the client organization’s knowledge management system; if they lack one, recommend creating one.

NOTE 15.7 Remember that effects of the evaluation are not effects of the program: an empowerment evaluation should produce substantial gains in the staff’s knowledge about and tendency to use evaluations, but 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—and these are common threats to empowerment evaluation—since training the program staff is not a primary criterion of merit for evaluations.

NOTE 15.8 Similarly, the usual non-money cost of an evaluation—disruption of work by program staff—is not a bad effect of the program. It is one of the items that should always be picked up in a meta-evaluation. Of course, it’s minimized in goal-free evaluation, since the evaluators do not talk to program staff. Careful design (of program plus evaluation) can sometimes bring these evaluation costs either near to zero, or ensure that there are benefits that more than offset the cost.

GENERAL NOTE G: The explanatory remarks here should not be regarded as more than approximations to the content of each checkpoint. More detail on them and on items mentioned in them can be found in the Evaluation Thesaurus, Michael Scriven, (4th edition, Sage, 1991), under the checkpoint’s name, or in the references cited there, or in the online Evaluation Glossary (2006) at evaluation.wmich.edu, or in the best expository source now, E. Jane Davidson’s Evaluation Methodology Basics (Sage, 2004). The above version of the KEC itself is, however, much better than the ET one, with help from many students and colleagues, most recently Emil Posavac, Chris Coryn, Jane Davidson, Rob Brinkerhoff, Lori Wingate, Daniela Schroeter, and Andrea Wulf; with a thought or two from Michael Quinn Patton’s work. More suggestions and criticisms are very welcome—please send to: Scriven@aol.com. [10,469 words]