KEY EVALUATION CHECKLIST

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

Michael Scriven
February 2007

GENERAL NOTE A: Throughout this document, "evaluation" is taken to mean the determination of merit, worth, or significance (abbreviated m/w/s); "evaluand" means whatever is being evaluated; and "dimensions of merit" a.k.a., "criteria of merit" refers to the characteristics of the evaluand that bear on its m/w/s. 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*** (Scriven, 1991), or in the ***Evaluation Glossary***, online at evaluation.wmich.edu. Merely for simplification, the term, "programs," is used 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 also can be used, with care, for the evaluation of (i) products (for which it was originally designed—although since completely rewritten and then revised and circulated at least 40 times); (ii) organizational units such as departments, consultancies, associations, and for that matter, and for that matter (iii) hotels, restaurants, and hamburger joints; (iv) services, which can be treated as if they were aspects of programs; (v) practices, which are either implicit policies ("Our practice at this school is to provide guards for children walking home after dark"), hence evaluable using the KEC, or habitual patterns of behavior, i.e., performances (as in "In my practice as a consulting engineer, I often assist designers, not just manufacturers"), which is a slightly different subdivision of evaluation; 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. 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 (Scriven, 2005) and a number of other papers about, and examples of, checklists in evaluation by various authors, under "Checklists" 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), 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 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, 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. Among other benefits, it makes you realize the importance of describing context, settling on a level of technical terminology, and 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 executive summary as a "teaser" that only describes what you looked at or how you looked at it, instead of what you found.) Through the whole evaluation process, 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' prior knowledge and needs for information. This helps you focus on what still needs to be done in order to learn about what matters most. The executive 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. It should convey some sense of the strength of the conclusions—which includes both the weight of the evidence for the premises and the robustness of the inference(s) to the conclusion(s).

II. Preface

Now is the time to identify and define in your notes, for assertion in the final report the (i) client, if there is one: this is 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) prospective audiences (for the report); (iii) stakeholders in the program (who will have a substantial vested interest in the outcome of the evaluation and may have important information about the program and its situation/history); and (iv) who else will see, have the right to see, or should see the (a) results and/or (b) 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 described as an empowerment evaluator), 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.

And now is the time to get down to the nature and details of the job or jobs, as the client sees them—and to encourage the client-stakeholders 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 usually involves more serious attention to cost analysis—rather than of *merit,* or of *significance,* or of more than one of these? Is the evaluation to be formative, summative, ascriptive ("ascriptive" means simply for the record, or for interest, rather than to support any decision), or more than one of these? Exactly what are you supposed to be evaluating (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 merely 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), the dimensions of success and failure (one type of analytic evaluation), the quality on each of those dimensions, or the contribution of each of its components (another two types of analytic evaluation)?

To what extent is a conclusion that involves generalization from this context being requested/required?

Are you also being asked (or expected) either to evaluate the client's theory of how the program components work or to create such a "program theory"—keeping in mind that this is something over and above the literal evaluation of the program and sometimes impossible in the present state of subject-matter knowledge? Is the evaluation to yield grades, ranks, scores, profiles, or (a different level of difficulty altogether) apportionments? Are recommendations, faultfinding, or predictions requested or expected or feasible (see Checkpoint 12)? Is the client *really* willing and eager 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 postreport 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 is 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 probably should be the last section you revise in the final report.

III. Methodology

Now that you've got the questions straight, how are you going to find the answers? Examples of the kind of question that has to be answered here include these: 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, subcontract)? 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, and other threats to internal validity? If you can't control these, what's the decision-rule for aborting the study? Can you double- or single-blind the study? If a sample is to be used, how will it be selected; and if stratified, how stratified? If none of these apply, how will you determine causation (of effects by the evaluand)? 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? Will/should the evaluation be goal-based or goal-free? To what extent will it be participatory or collaborative, and what standards will be used 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.; data-analytic procedures (statistics, 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 literature 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 and for the report. The same process also will 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 the success/failure of the program and put matched numbers on any that look 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 creating or implementing or supporting 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 literature 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 see or find later 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—and discuss 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 they usually are kept 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

often is 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 (called 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 and/or modus operandi methodology.) While looking for the impacted population, you also may 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 subissue, 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 and 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), Checkpoint11 (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 constraints. This checkpoint is the one that covers individual and social capital *available* to the program; there also is social capital *used* by the program (part of its Costs) and sometimes social capital benefits *produced* by the program (part of the Outcomes).[1]

[1] 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 relationships ("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 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 nonadditive social capital: The skills of football or other team members that will not provide (direct) benefits for people who are not 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 toward 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 definitions of "human capital" and the 10 for "social capital" at 6/06 included simplified and erroneous as well

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 in evaluation, and on some issues, e.g., surgery options, they are usually 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 intercultural and international differences are 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. 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 criteria of merit, by the meaning of the terms, include the extent (a.k.a., reach or breadth) of its impact (i.e., the size and range of categories of the impactee population), and the impact's depth (a.k.a., magnitude or extent or duration) of average effect. Other primary criteria in such a case, from a general understanding of the nature of a health program, include low adverse eco-impact, physical ease of access/entry, a broader age/gender/ethnic range for acceptance. Then turn to the list below 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 follows is a fairly detailed account of the core logic of integrating facts and values, the distinctive methodological skill in the evaluator's repertoire.

This means, first, attaching a grading scale to each dimension for each evaluative purpose. For example, 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 for a national review, however, an 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 "among the best in the country" and give the program a D for using 550 as a C (D meaning less than satisfactory). So evaluators often set standards at different points on the same performance dimension when the purposes for grading are completely different—one is grading students for admission, and the other is grading programs for "outstanding in the U.S."

---

as diverse uses—few dictionaries have yet caught up with these terms (although the term "human capital" dates from 1916).

[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. These are often the most important practical questions, for most people—and their representatives—looking at programs (the same applies to product evaluation, personnel evaluation, etc.). Examples of such questions include these: 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?

All of this is one part of 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 5- or 10-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 she or 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 offset by other virtues *for entry into a creative writing program.*[3] 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.[4] In serious evaluation, it is often appropriate to establish "steps," i.e., points or intervals on measured dimensions of merit where the weight changes, if the mere stars don't provide enough effect. An example of this is the setting of several cutting scores on the GRE for different grades in the use of that scale for two types of evaluation given above. 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 (-reduction), superiority to feasible alternatives, and generalizability. Their presence in the KEC brings the number of types of values considered up to 21.

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

(i) values 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 an ideal or excellent evaluand of this type (e.g., a good shuttle bus service for night shift workers would have 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 (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

[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 weighted as 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 thinks 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.)

that accepts the most rudimentary ethical considerations (i.e., the non-zero value of the welfare of other human beings), 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 and overrides when in conflict), usually including (reasonable) safety, confidentiality (and perhaps anonymity) of all records, for all impactees. (Problems like conflict of interest and protection of human rights have federal legal status in the U.S., plus scientific good procedural standards, and also universal ethical status.)

(vi) cultural values held with a high degree of respect (and thus distinguished from matters of manners, style, taste, etc.), of which an extremely important one is honor; another group, not always distinct from that one, concerns respect for ancestors or tribal or totem spirits or local deities. These, like legal requirements, are subject to override, in principle at least, by ethical values, although often taken to have the same status.

(vii) personal, group, and organizational goals/desires (unless you're doing a goal-free evaluation) if not in conflict with ethical/legal/practical considerations. 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 many evaluative conclusions about e.g., what recreational facilities to provide in community-owned parks.

(viii) 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

(ix) sublegal but still important legislative preferences (GAO used to determine these from an analysis of the hearings in front to the subcommittee in Congress from which the legislation emanated

(x) professional standards (i.e., standards set by the profession) of quality that apply to the evaluand[5]

(xi) expert refinements of any standards lacking a formal statement, e.g., those in (ix)

(xii) historical/traditional standards

(xiii) scientific merit (or worth or significance)

[5] Since one of the steps in the evaluation checklist is the metaevaluation, 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 on Standards for Educational Evaluation (1994); but there are several others of note, e.g., the *GAO Yellow Book* (GAO, 2007) and this KEC.

(xiv) technological m/w/s

(xv) marketability

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

(xvii) risk (or chance), meaning the probability of failure (or success) or of the loss (or gain) that would result from failure (or success). This is *not* 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. This is the value or disvalue of the chance element in the enterprise in itself, as a 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. This consideration is particularly important in evaluating plans (preformative evaluation) and in formative evaluation, but *also relevant in summative and ascriptive evaluation when either is done prospectively (i.e., before all data are available)*. There is an option of including this under personal preferences, item (vii) above, but it is often better to consider it separately since it can be very important and it is a matter on which evidence/expertise (in the logic of probability) can be brought to bear, not simply a matter of taste.[6]

(xviii) last but not least—resource economy (i.e., how low impact is the program with respect to limited 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 8) in the normal currencies (money and nonmoney), but refers to absolute loss of available resources (in some framework, which might range from a single person to a country). 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. It partly overlaps with Checkpoint 10, because a low score on resource economy undermines sustainability, so watch for double counting.

Fortunately, bringing these standards to bear[7] 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 will be crucially important in some other particular cases. Doing all this values-analysis will be easy to do sometimes, although very hard on other occasions; it often can require expert advice and/or impactee/stakeholder advice. Of course, some of these values will conflict with others (e.g., impact

[6] Note that 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 is often used to mean *either* the probability of disaster ("very risky") *or* the disaster itself ("the risk of death"). Now the classical concept 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 risk-tolerant person will 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 difference 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 risk factor, i.e., that he or she is less risk-averse, not more of a risk-lover. However described, this can be a major value difference between people and organizations e.g., venture capitalist groups or city planning groups.

[7] "Bringing them to bear" involves (a) identifying the relevant ones, (b) specifying (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.

size with resource economy), so their *relative* weights may then have to be determined for the particular case, a nontrivial task by itself. Hence, you need to be very careful not to *assume* that you have to generate a ranking of evaluands from an evaluation you are asked to do of several of them, 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 a profiling display (showing the merit on all relevant dimensions of merit in a bar graph for each evaluand[s]).

NOTE 5.1: 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 may well 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 (iii) the subevaluation. In other words, Part C requires the completion of five separate inferences from (i) to (iii), 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 the following checkpoints will 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 since the program began, through design provisions for reshaping under environmental or political or fiscal duress (including planning for worst-case outcomes); to the development and justification of the program logic (but see Checkpoint 12), along with the program's "implementation fidelity" (i.e., degree of implementation of the supposed archetype program, if any; this is also called "authenticity," "adherence," "alignment," or "compliance"). You must also check the accuracy of the official name or subtitle (if either is descriptive or evaluative) or the official description of the program (e.g., "an inquiry-based science education program for middle school," "raising beginners to proficiency level," "advanced critical thinking training program") and its management (especially the arrangements for getting and appropriately reporting evaluative feedback [that package is most 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 appropriate standards of accountability and transparency). You need to examine all activities and procedures and the program's general learning process (e.g., regular "updating training" to cope with changes in the operational environment, staff aging, essential skill pool, new technology), attitudes/values, and morale.

Under this heading you may or may not 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/staff behavior). It is not always 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 its analgesic effects; only recently have we had any idea how/why it works. It would have been a logical error—and unhelpful to society—to make the evaluation depend on solving the causal mystery; and of course 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's the situation). So if you can be helpful by evaluating the program theory, do it; but it's not an essential part of doing a good evaluation and will often be a diversion and sometimes a cause for antagonism.

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, 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, 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—once.

Here are four other examples of why process is an *essential element* in program evaluation, despite the common tendency in much evaluation to place almost the entire emphasis on outcomes: (i) gender or racial prejudice in selection/promotion/treatment of staff is an 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; (iii) 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, in kindergarten at least; painting; and marching to protest war, even if it doesn't succeed); (iv) the treatment of human subjects must meet federal, state, and other ethical standards, and an evaluator cannot avoid the responsibility for checking that they are met.

7. Outcomes

Evaluation of (good and bad) effects on program recipients, on others, and on the environment: these must include direct *and* indirect, intended *and* unintended, immediate[8] and short-term and long-term effects. (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"[9]) and because determining the m/w/s of the effects—this is the result you

[8] 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.

[9] 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 that are known to be able to cause it in some circumstances or other and (ii) on the right (recto) side, all the things that it can cause: that's the side you need to guide the search for side effects.

have to get out of this subevaluation—is often the hard part, not determining whether there are any or even 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 outcomes (i.e., results that are of major significance, whether or not intended) can occur during the process and should be 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 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 (see the earlier footnote), 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 programs, maybe even to identify the key to success as particular features that are to be found only in the successful programs. To do this usually does not require the identification of the true operating logic/theory of program operation (remember to contrast this with the (i) original official, (ii) current official, and (iii) the implicit logics or theories). Also see Checkpoint 12 below.

Given that the most important outcomes may have been unintended (a broader class than unanticipated), it's worth distinguishing between *side effects* (which affect the target population and possibly others) and *side impacts* (i.e., impacts of any kind on nontargeted populations).

The biggest methodological problem with this checkpoint is establishing the causal connection, especially when there are many possible or actual causes and attribution of portions of the effect to each of them must be attempted. On this, consult recent articles by Cook (2006) and Scriven (2006).

NOTE 7.1: Remember that success cases may require their own treatment as a group, regardless of average improvement due to the program (since the benefits in those cases alone may justify the cost of the program)[10]; the failure cases should also be examined. 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 preprogram 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. This impact group, as previously mentioned, is sometimes called the "anticipators." Looking at these effects carefully is sometimes included under preformative evaluation (which also covers looking at other dimensions of the planned program, such as evaluability).

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, among other scientists, have comprehensive "local editions" of the BofC in their heads and as part of the social capital of their guild.

[10] As argued very well by Robert Brinkerhoff (2003) in *The Success Case Method*.

NOTE 7.2: It is usually true that evaluations have to be completed long before some of the main outcomes have occurred, let alone have 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 determining *worth* (or, in one sense, *value)* by contrast with plain *merit* (or *quality*). It requires attention to both (i) money *and* nonmoney costs, (ii) direct *and* indirect costs, and (iii) both actual *and* opportunity costs.[11] 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; by cost elements (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 nonmoney costs are space, time, expertise, and common labor (when these are not available for purchase in the open market—if they are so available, they just represent 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 surely can benefit if one is available or being done in parallel. Even without the accounting expertise, your cost analysis and certainly your evaluation, if you follow the KEC, reaches key factors omitted from standard auditing practice.

NOTE 8.1: This subevaluation 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 or fewer resources; anything that does this is known as a "critical competitor." 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; the extra cost may still yield a bargain. Sometimes it's also 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 critical competitor *might* not match on side effects. Treading on potentially thin ice, sometimes there also are 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

[11] 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 you might argue that you can't determine the value of the MVFA without knowing what *it* required you to give up. In general, it's better to define ordinary costs as the *tangible* valued resources that were used to cause the evaluand to come into existence (money, time, expertise, effort, etc.) and opportunity costs as another dimension of cost, namely, the MVFA you spurned by choosing to create the evaluand rather than that alternative path to your goals, using about the same resources. The deeper problem is this: the "opportunity cost of the evaluand" 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 Scriven, 1983, "Costs in Evaluation: Concept and Practice.")

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). Also, 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 often are extremely illuminating and sometimes absolutely essential—as when a government has to decide on a health program. It sometimes looks as if they are 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, taken all together, really do add up to an improvement; so you need to compare version 2 with version 1, *and also* with available alternatives.

NOTE 9.2: It's tempting to collapse the Cost and Comparison checkpoints into "Comparative Cost-Effectiveness" (as Davidson [2004] does, for example), 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 available alternatives are cost feasible, and the 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 frequently is 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 for a very small group of impactees—they may be extremely important because of their generalizability. Generalization to later *times*, a.k.a. longevity, is nearly always important (under adverse conditions, it's durability). Even more important is sustainability (sometimes referred to as "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 or disasters of the natural as well as financial, social, ecological, and political varieties. This is sometimes 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 these programs is the future of "the program in context"; and so any required context should be specified, including 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 always desirable. It is still the case that good researchers make careless mistakes of inappropriate 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 since most but not all student rating forms are invalid or uninterpretable for more than one reason, the essential foundation for the generalization—a common reference—is nonexistent. 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 in the same category as values from the list in Checkpoint 5. Why do they get special billing with their own checkpoints in the list of subevaluations? 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 as relevant 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 and values into an overall evaluation, i.e., at least into a profile (a.k.a., bar graph, the simplest graphical means of representing a multidimensional conclusion and greatly superior to a table for most clients and audiences) or into a unidimensional conclusion—a grade or a rank, if that is required (doing this is 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 dimensions), but usually there should also be some conclusion(s) aimed at the client's and other stakeholders' need for concise evaluative information. A possible framework for that slice of the summary is the SWOT checklist used in business: Strengths, Weaknesses, Opportunities, and Threats.[12] This part of the summary should include referencing the results against the client's and perhaps other stakeholders' goals, wants, or hopes (if feasible)—e.g., goals met, ideals realized, created but unrealized value—when these are determinable. But the primary obligation of the evaluator is usually to reference the results to the needs of the impacted population, within the constraints of overarching values such as ethics, the law, the culture, etc. Programs are not made into good programs by matching someone's goals, but by doing someone some good. Of course, the two should coincide, but you can't assume they do; they are often incompatible.

NOTE 11.1: One special conclusion to go for, often a major part of determining <u>significance,</u> comes from looking at what was done against what could have been done with the resources available, including social and individual capital. This is where imagination is needed to get the opportunities part of the SWOT analysis.

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

NOTE 11.2: Be sure to convey some sense of the *strength* of the conclusions, which means the combination of the net weight of the evidence for the premises with the robustness of the inferences from them to the conclusion(s). (For example, was the performance on the various dimensions of merit sometimes a tricky inference or directly observed? Did it clear any bars or lead any competitors "by a mile" or just scrape over? Was the conclusion established "beyond any reasonable doubt" or merely "supported by the balance of the evidence?")

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. Let's start with recommendations. *Micro-recommendations*—those concerning the *internal* workings of program management and the equipment or personnel choices/uses—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), *or* they may occur to the smart evaluator who is motivated to help the program, because of his or her expert knowledge of this or an indirectly or partially relevant field such as information 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 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 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.—what we might call *external* management recommendations)—are usually another matter. These are important decisions *serviced by* and usually *dependent on* summative evaluations; but making recommendations about them is not 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 recommendations about these decisions will typically require (i) extensive knowledge of the other factors in the context-of-decision for the external ("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 unable to provide full details about the context-of-decision concerning the program (unable because implicit values are not always recognized by those who operate using them); (ii) considerable extra effort e.g., to evaluate each of the macro-options. Key elements in this list 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. 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, 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. 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, in contrast with an evaluation of the evaluand itself, typically against alternative available evaluands rather than against the alternative of not having it at all. In the standard program evaluation, but not 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.

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 as well as "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; 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 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. Worse, 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, serial killing, ADHD. In such cases, what we need to know is whether we have found a cure—complete or partial—and the explanation can wait. That's the "aspirin case"—the situation where we can easily, and with great benefit to all sufferers, evaluate a claimed medication but don't know why it works and don't need to know that in order to evaluate its efficacy. The bottom line is this: 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. Of course, ***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*** also are 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.[13] 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 (1954) 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, even with all that help, is not 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).[14] On rare

[13] 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 on PhysLrnR, the Physical Learning Research List, April 16, 2006.)

[14] In PERT (Program Evaluation and Review Technique) charting—a long-established approach to program planning that emerged from the complexities of planning the first submarine nuclear missile, the Polaris—the

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.

NOTE 12.2: ***Policy analysis***, in the common situation when the policy is 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 typically 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 also is 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 period; 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 "counterrevolutions" 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 12.3: ***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. 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; 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 evaluability is a matter of degree, resources, and circumstance, not of absolute possibility. Hence, 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 program for Iran in April 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 April 2006—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 or brought on an attack.

NOTE 12.4: ***Inappropriate expectations*** The fact that clients often expect/request explanations of success or shortcomings, or macrorecommendations, or 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 often is 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,*

---

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. My take on this approach is that it makes sense only 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 an 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.

better or worse than the available alternatives, 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/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, affordable approach, and frequently the only feasible one. In fact, most evaluations are of partially blacked-out boxes. This is perhaps most obviously true in pharmacological evaluation, but it also is 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 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 (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 and ***proformative*** evaluation in the second—a cross between formative evaluation and proactivity. If macrorecommendations by evaluators require great care, proactivity requires even greater care.

13. (possible) 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, as well as (iii) an ethical analysis of their options and of 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 U.S. 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, typically involves one or more 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 should be a part of each evaluation done of the response—and (v) an evaluation of subsequent corrective/preventative action. All five involve some evaluation of responsibility and sometimes the allocation of praise/blame. Recent efforts 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.

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 semitechnical 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 nontechnical, graphical or textual, scientific or story-telling, anecdotal and personal, or bare bones. This step in the evaluation process should include postreport 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.[15] There usually should be an explicit effort to identify "lessons learned," failures and limitations, costs if requested, and "who evaluates the evaluators." Checkpoint 14 also should 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 hard-copy posters, graffiti, books, 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 metaevaluator). 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/legality, which includes such matters as conflict of interest[16] and protection of the rights of human subjects. There are several ways to go about metaevaluation. You and later the metaevaluator can (a) apply the KEC, *Program Evaluation Standards* (Joint Committee, 1994), ES or GAO (2007) auditing standards to the evaluation itself (the Cost checkpoint in the KEC then addresses the cost of the evaluation, not the program, and so on for all checkpoints); and/or (b) use a special metaevaluation checklist (there are several available, including the one in the previous sentence); 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. It's highly

[15] The "connotative implications" are in the subexplicit but suprasymbolic realm of communication, manifested in, e.g., the use of gendered or genderless language.

[16] 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 formative evaluation on the same program, of which they are now coauthors or rejected contributor wannabes.

desirable to employ more than one of these approaches, and all are likely to require supplementation with some attention to conflict of interest/rights of subjects.

NOTE 15.1: 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 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 15.6: Since effects of the evaluation usually are not properly included 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 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.

NOTE 15.7: Similarly, the usual nonmoney 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 always should be picked up in a metaevaluation. Of course, it's minimal in goal-free evaluation, since the (field) evaluators do not talk to program staff. 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.

GENERAL NOTE G: The explanatory remarks here should not be regarded as more than approximations to the content of each checkpoint. More detail on many of them and on items mentioned in them can be found in the *Evaluation Thesaurus* (Scriven, 1991), under the checkpoint's name, or in the references cited there, or in the online *Evaluation Glossary* at evaluation.wmich.edu, or in the best expository source now, E. Jane Davidson's (2004) *Evaluation Methodology Basics*. 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, Christian Gugiu, Liliana Rodriguez-Campos, 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.*

References

Brinkerhoff, R. (2003). *The success case method: Find out quickly what's working and what's not.* San Francisco: Berrett-Koehler.

Cook, T. D. (2006). Describing what is special about the role of experiments in contemporary educational research? Putting the "gold standard" rhetoric into perspective. *Journal of Multidisciplinary Evaluation* (6). Available from http://evaluation.wmich.edu/jmde/JMDE_Num006.html.

Davidson, E. J. (2004). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs.* Thousand Oaks, CA: Sage.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Scriven, M. (1983). Costs in evaluation: Concept and practice. In M. C. Alkin & L. C. Solomon (Eds.), *The costs of evaluation*. Beverly Hills: Sage Publications.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed). Newbury Park, CA: Sage.

Scriven, M. (2005). *The logic and methodology of checklists.* Available from www.wmich.edu/evalctr/checklists/

Scriven, M. (2006). Converting perspective to practice. *Journal of Multidisciplinary Evaluation* (6). Available from http://evaluation.wmich.edu/jmde/JMDE_Num006.html.

U.S. Government Accountability Office (GAO). (2007). *Government auditing standards*. Washington, DC: Author. Available from http://gao.gov/govaud/ybk01.htm