

KEY EVALUATION CHECKLIST (KEC)

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

Michael Scriven

October 23, 2005

MAIN 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 some knowledge of some terms from evaluation vocabulary is assumed, e.g., formative, goal-free, ranking; their definitions can be found in the *Evaluation Thesaurus* (4th ed., Sage, 1991). As a 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.”

MAIN NOTE B: The KEC can also be used, with care, for the evaluation of products, and organizational units such as departments; and, with considerable imagination, for some tasks in the evaluation of personnel.

MAIN 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 useful papers about checklists in evaluation by various authors, at the site of the Checklist Project, which is at evaluation.wmich.edu.

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. For one thing, it makes you realize the importance 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 process. Typically this should be done without even mentioning the process. (In other words, avoid the pernicious practice of using the executive summary as a “teaser.”) Through the whole process of evaluation, keep asking yourself how the overall summary is going to look, based on what you have learned so far, and how it relates to the client’s and stakeholders’ and audiences’ needs. This helps you focus on what still needs to be done 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.

II. Preface

Now is the time to identify and set out 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 straighten this out); (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 (an outsider brought in for a limited time to help the staff with setting up 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, etc.

And now is the time to get down to the 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 significance), or of more than one of these? Exactly what are you supposed to be evaluating: How much of the context is to be included? How many of the details are important? Are you supposed to be evaluating the effects of the program as a whole or the contribution of each of its components, or perhaps additionally the client’s theory of how the components work? Are you to consider impact in all relevant respects or just some respects? Is the evaluation to be formative, summative, ascriptive, or more than one of these? Should it yield grades, ranks, scores, profiles, or apportionments? Are recommendations, explanations (i.e., your own theory), fault-finding, or predictions requested, or expected, or feasible? Is the client *really* willing and eager to learn from faults or is that just a pose? (Your contract or, for an internal evaluator, your job, depends 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—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: It’s best to discuss these issues about what’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. Be sure to note later any subsequently negotiated changes in any of the preceding. And here’s where you give acknowledgments/thanks . . .

III. Methodology

Now that you’ve got the questions straight, how are you going to find the answers? Examples of questions that have to be answered under this checkpoint: 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)? Can you use control or comparison groups to determine causation of supposed effects? If there’s to be a control group, can you randomly allocate subjects to it? Can you double- or single-blind the study? If a sample is to be used, how will it be selected, and if stratified, how will it be stratified? If none of these approaches are possible, how will you determine causation? (Of effects by the evaluand; depending on the task, you may also need to determine the contribution to the effects of various components of the evaluand.) Will/should the evaluation be goal-based or goal-free? 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? As soon as possible, identify other investigative procedures (observations, participant observations, logging, journaling, audio/video recording, tests, simulating, role-playing, survey, interview, focus groups, text analysis, library/online searches/search engines, etc.) and data-analytic procedures (stats, cost-analysis, modeling, expert consulting, etc.) to be used in this evaluation, plus reporting techniques (text, stories, plays, graphics, freestyle drawings, stills, movies, etc.) and their justification (may require a literature review on some of these methods). Hence, provide the “logic of the evaluation,” i.e., general justification of its total design.

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, simultaneous, and any projected settings for the program. Identify (i) any “upstream stakeholders”—and their stakes—other than clients (i.e., identify people or groups or organizations that assisted in implementation of the program or its evaluation, e.g., with funding or advice or housing); (ii) enabling (and any more recent relevant) legislation/policies—and any legislative/executive/practice or attitude changes since start-up; (iii) the underlying rationale, a.k.a. official program theory, and political logic (if either exist or can be reliably inferred; though neither are necessary, they are sometimes useful); (iv) general results of lit review on similar interventions (including “fugitive” studies

[those not published in standard media], and the Internet [consider including the “invisible web,” and the latest group and blog search engines]); (v) previous evaluations, if any; (vi) their impact, if any.

2. Descriptions and Definitions

Record any official description of program + components + context/environment, but don't assume it's correct. Develop a correct and complete description, which may be very different, in enough detail to recognize the evaluand, and perhaps—depending on the purpose of the evaluation—to replicate it. Get a detailed description of goals/mileposts (if not operating in goal-free mode). Explain meaning of any “technical terms,” i.e., those that will not be in prospective audiences' vocabulary. 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, 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), sometimes called “clients” (but they are clients of the program not of the evaluation, so it's usually better to restrict the use of this term in the context of talking about evaluation to the sponsor of the evaluation) 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 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'être* 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 and/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 or contract the program when/if you get to checkpoint 12.

4. Resources (a.k.a. “strengths assessment”)

The financial, physical, and intellectual-social-relational assets of the program. 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 comparisons that an evaluation considers. It may be helpful to list specific resources that were *not* used/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 check, and complement, consider all *constraints* on the program.

5. Values

Begin by identifying the relevant values for evaluating this evaluand in these circumstances from the following list. Validate them for the present project as current, and as currently and contextually supportable. Add a preliminary indication of “stars, bars, and steps” as appropriate for this evaluand in this (or these) implementation(s). The “stars” are the weights, i.e., the relative or absolute importance of the dimensions of merit or other values that will be used to get from the facts about the evaluand, as you locate or determine them, to the evaluative conclusions. (Their importance might be expressed qualitatively (e.g., letter grades) or quantitatively (e.g., points on a ten point scale, or—usually a better method—by allocation of 100 “weighting points” across the set of values) or relatively in terms of an

ordering of their importance). The “bars” are minimum standards for acceptability, if any. Bars and stars may be set on any relevant properties (a.k.a., dimensions of merit or values) or on dimensions of measured (valued) performance, and may additionally include holistic bars or stars.¹ These are commonly referred to as standards. In serious evaluation, it may also be appropriate to establish “steps,” i.e., the points or intervals on measured dimensions of merit where the weight changes. (Bars and steps may be fuzzy as well as precise.)

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

- (i) 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 act, not values in themselves, but in any context that accepts the most rudimentary ethical considerations, those facts are value-imbued.
- (ii) criteria of merit from the definition of the evaluand and from standard usage (e.g., since a program is usually regarded as definitionally better if it reaches more people and has a larger good effect on them [other things being equal], the criteria of merit typically include the number of people impacted by the program and the depth of desirable impact)
- (iii) logical requirements (e.g., consistency)
- (iv) legal and (v) ethical requirements (they overlap), usually including (reasonable) safety, confidentiality, 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 the ethical backing, but are enough by themselves to drive the inference to an evaluative conclusion about, e.g., which apartment to rent
- (vii) fidelity to alleged specs (a.k.a. “authenticity,” “adherence,” 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)
- (viii) sublegal but still important legislative preferences
- (ix) professional standards of quality that apply to the evaluand²
- (x) expert judgment
- (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
- (xvi) last but definitely not least—resource economy (i.e., how low-impact is the program with respect to money, space, time, labor, contacts, expertise and the ecosystem?)

¹ Example: The candidates for admission to a graduate 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 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.

² Since one of the steps in the evaluation 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 are those developed by the Joint Committee (*The Program Evaluation Standards, 2nd ed.*, Sage).

Of course, identifying/validating/applying some of these is unimportant in some cases, crucially important in others; easy to do sometimes, very hard on other occasions; and it will often require expert advice and/or impactee/stakeholder advice. Also of course, some of these will conflict with others (e.g., impact size with economy), so their relative weights must then be determined for the particular case, a nontrivial task. Hence you need to be very careful not to assume that you need to generate a ranking from the evaluation. 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 of merit on all relevant dimensions of merit, in a bar-graph) of the evaluand(s).

NOTE: 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). 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 list, which means you will often have to come back to modify findings on earlier checkpoints.

PART C: SUBEVALUATIONS

Each of these involves (i) a fact-finding phase, followed by (ii) the process of combining the facts with whatever values (from 5 above) bear on those facts, which yields the subevaluation. In other words, Part C requires the completion of five separate steps from What's So? to So What?, e.g., from "the effects were measured as XXX" to "the effects were extremely beneficial" (or "a bargain" etc.).

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 the design provisions for resilience under environmental or political or fiscal duress (including planning for worst-case outcomes); implementation fidelity (i.e., degree of implementation of the supposed archetype program, a.k.a. "authenticity", "adherence", "compliance"); 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; activities; procedures; the learning process; attitudes/values; morale; perhaps also, if you're covering this in any detail, the quality of the original logic of the program and its current logic (both the current official one and the possibly different one implicit in the operations/staff behavior). Process evaluation may also include the evaluation of what are often called "outputs," (usually taken to be "intermediate outcomes" 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.

7. Outcomes

Evaluation of (good and bad) effects on consumers: direct/indirect, intended/unintended, immediate/short-term/long-term. 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 legendary "Book of Causes"³). Immediate outcomes 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 are 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 can adjust usage of these terms to client/audience/-stakeholder preferences.) (Sometimes, not always, it's useful and feasible to provide explanations of success/failure in terms of components/context/decisions. To do this may or may not require the

³ The Book of Causes (BofC) 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.

identification of the true operating logic/theory of program operation [by contrast with (i) the original, (ii) the current, (iii) the official, and (iv) the implicit, logics or theories). 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 *side impacts* (i.e., impacts 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., look for (ii) social, and (iii) environmental, outcomes as well as (i) conventional outcomes. Finally, don't forget (i) the effects on the program staff, especially lessons and skills learned 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 immigration or emigration to the community, and are often more serious in at least the economic dimension than the directly caused results of the program's implementation.

8. Costs

Cover money *and* nonmoney costs; direct *and* indirect; actual *and* opportunity costs; itemize by developmental stage, i.e., start-up/maintenance/upgrade/shutdown costs; and/or by calendar time period; and by components (rent, equipment, personnel, etc.), if relevant and possible. Include use of expended but never realized value, if any, e.g., social capital. The most common nonmoney 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); and stress, political and personal capital, and environmental impact, which are rarely *fully* coverable by money.

9. Comparisons

The key comparisons tend to be with other means for getting the same or similar benefits from about the same or lesser resources; anything that does this is known as a "critical competitor." It is often worth looking for, and reporting on, one alternative—if you can find one—that is much cheaper but nearly as effective ("el cheapo") and one much stronger although costlier alternative, i.e., one that produces many more payoffs or process advantages ("el magnifico"); and it's sometimes worth comparing the evaluated with a widely adopted/admired approach that is perceived as an alternative, though not really in the race, e.g., a local icon. Remember that "having the same effects" covers side effects as well as intended effects. Treading on potentially thin ice, there are also sometimes strong reasons to compare with a demonstrably possible alternative (the next checkpoint is one 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 if you have an ongoing role as formative evaluator—and you need to be sure that your invitation to do this was genuine and authoritative (see also Checkpoint 12).

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

Can the program be used with similar results if we use it with other content, at other sites, with other staff, with other recipients, in other climates (social, political, physical), etc. Generalization to later times is longevity (under adverse conditions, it's durability), and it is almost always crucial to consider this. Making sure this checkpoint covers the financial, social, spatial, temporal, environmental, political, and other nonmoney costs, capacity, and conditions for survival, yields the sustainability (a.k.a. "resilience to risk") rating, which is even more important than longevity, for example when evaluating international or crosscultural developmental programs.

NOTE: What you're generalizing about then is "the program in context," and the context should be specified and include infrastructure. Here, as in the last sentence in Checkpoint 9, we are talking about possibilities, and these, although risky, sometimes generate the greatest contribution of the evaluation to improvement of the world; see also the "possible scenarios" of Checkpoint 14. (All three show how much great evaluation is a creative and not just a reactive enterprise.)

PART D: CONCLUSIONS

11. Overall Significance

Combine the subevaluations of Part C into an overall evaluation, i.e., at least into a profile (one means of representing a multidimensional conclusion) or into a unidimensional conclusion—a grade or a rank, if that is required (usually much harder). The focus (point of view) should usually be the present and future impact on consumers' needs, subject to the constraints of ethics and the law (and feasibility, etc.—i.e., including all the other relevant values), but usually there should also be some conclusion(s) that refers to the client's (and other stakeholders') needs for information (and wants or hopes, if feasible), e.g., goals met; unrealized value, if calculable.

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 a few caveats. 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 relevant field such as information technology or clinical psychology, and they are often very useful. (Getting them is one of the good side effects of using an external evaluator.) But macro-recommendations—which are about the disposition of the whole program (fund, cut, modify, export, etc.—what we might call *external* management recommendations)—are usually another matter. They will often require: (i) extensive knowledge of the context-of-decision for the about-program decision makers (who are not always the clients for the evaluation *and* who may be unwilling or psychologically unable to provide full details about the context of decision about the program); (ii) considerable extra effort; *and* (iii) knowledge of all the within-program management options available. Key elements in this list are often not available to anyone, including the most expert of experts, in the present state of the art on that particular topic (improving that program in that context), or only available to a board of directors, or to select legislators, or perhaps only to their psychotherapists, hence inaccessible to the evaluator. These extra requirements, 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 often requires more than—and rarely requires less than—state-of-the-art subject-matter expertise. The combination of these requirements often means that the attempt to provide recommendations or explanations 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 a 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. Plain *predictions* are also often requested (e.g., Will the program work reliably with the recommended changes, without staff changes?) and are often very hazardous.

Policy analysis, when the policy is an alternative being considered for future adoption, is essentially program evaluation of future (possible) programs and hence necessarily involves predictions. Extensive knowledge of the fate of similar programs in the past is then the key resource. If the policy has already been implemented, policy analysis then boils down to something very close to program evaluation.

The fact that clients expect/request explanations, macro-recommendations, and 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 are not an excuse for doing, trying to do, and especially 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 an evaluative conclusion about the evaluand. The merit, worth, or significance of a program are often hard to determine; determining how to improve it, why it works or fails to work, and what one should do with it, are simply other tasks, often of great scientific and/or managerial/social interest, but 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, frequently the only feasible one.)

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 they may have. 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; 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 investigations of aircraft crashes by the National Transportation Safety Board in the U.S. are 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 due to natural causes. (Operating room deaths pose similar problems and are sometimes as well investigated.)

NOTE: 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 relief efforts, (iv) an evaluation of lessons learned, and (v) an evaluation of corrective action. All five involve some evaluation of responsibility and the allocation of praise/blame. Recent efforts appear not to have distinguished all of these and not to have covered all of them.

14. Report and Support

Now we come to the task of conveying the conclusions in an appropriate way and at appropriate times. 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 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 or barebones. 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 that are or are not consistent with the findings, i.e., doing some problem-solving for the client. In this process, a wide range of communication skills are often useful, e.g., audience “reading,” use and reading of body language, understanding the cultural iconography and connotative implications of the presentations and responses (the “connotative implications” are the subexplicit but suprasymbolic realm of communication, manifested in, e.g., the use of gendered or genderless language). There should be an explicit effort to identify “lessons learned.” Checkpoint 14 should also cover getting the results (and incidental knowledge findings) into the relevant databases, if any, possibly but not necessarily via journal publication; 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, discussions at scholarly meetings, or in hardcopy posters, graffiti, and movies.

15. Metaevaluation

This is evaluation of an evaluation, including those to which the earlier checkpoints are addressed, in order to identify its strengths/limitations/other uses. This should always be done, as a separate quality control step, as follows: (i) to the extent possible, by the evaluator, certainly—but not just—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, along with (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 five ways to go about this: you and then the metaevaluator can: (a) apply the KEC list to the evaluation itself; and/or (b) use a special metaevaluation checklist (there are several available); 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.

NOTES: (a) Utility is usability and not actual use, the latter—or its absence—being at best an indicator of the former. (b) Implementation does not prove high usability. (c) 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” is a better generic term for the ideal response. (d) Evaluation impact often occurs years after submission and often occurs even if the evaluation was rejected completely when submitted. (e) Help with utilization beyond submitting the report should at least have been offered. (f) Look for contributions to the client organization's knowledge management system; if none exists, recommend creating one. (g) Remember that effects of the evaluation are not effects of the program: an empowerment evaluation produces 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 external credibility, since it's not a primary criterion of merit, and these are common problems with empowerment evaluation. (h) Similarly, the usual nonmoney cost of an evaluation—disruption of work by program staff—is *not* a bad effect of the program; and, of course, it's a minimal effect in goal-free evaluation, since the evaluators do not talk to program staff. (It *is* one of the items that should be picked up in a metaevaluation.) Careful design can sometimes bring these evaluation costs either near to zero, or ensure that there are benefits which more than offset the cost.

MAIN NOTE D: The two checkpoints marked “possible”—12 and 13—are not always relevant or feasible and very often require some extra time/costs. They are mentioned because they are often supposed to be obligatory or “obviously” part of any professional evaluation, although this is not true since black box evaluation is often useful, often all that's possible, and often the best way to answer the client's question (certainly the fastest and cheapest, which is a long start towards best). The descriptions under these checkpoints will provide some guidance, if doing them is feasible and desired by the client.

MAIN NOTE E: 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 can be found in the *Evaluation Thesaurus* (Michael Scriven, 4th ed., Sage, 1991), under the checkpoint's name, or in the references cited there, or, the best 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, Jane Davidson, Rob Brinkerhoff, Lori Wingate, and Andrea Wulf; and a thought or two from Michael Quinn Patton's work. More suggestions and criticisms are very welcome—please send to: michael.scriven@wmich.edu.

This checklist is being provided as a free service to the user. The provider of the checklist has not modified or adapted the checklist to fit the specific needs of the user, and the user is executing his or her own discretion and judgment in using the checklist. The provider of the checklist makes no representations or warranties that this checklist is fit for the particular purpose contemplated by users and specifically disclaims any such warranties or representations.