KEY EVALUATION CHECKLIST

(mainly for evaluating programs and policies and evaluations of them)

Michael Scriven

October 2004

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 a few basic terms from evaluation vocabulary is assumed, e.g., "formative," "goal-free," "ranking;" their definitions can be found in the *Evaluation Thesaurus* (Scriven, 1991).

MAIN NOTE B: This is an iterative checklist, not a one-shot checklist, i.e., you should expect to go through it many times, even for design purposes, since discoveries under later checkpoints will often require modifications of what was entered under earlier ones (and no rearrangement of the order will avoid this). For more on the use of checklists 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: www.wmich.edu/evalctr/checklists/.

PART A: PRELIMINARIES	
I. Executive Summary	Usually a selective summary of checkpoints 11-15.
II. Preface	Source and nature of the request or need leading to the evaluation, e.g., is the request or the need for an evaluation of worth or of merit (or significance) or of more than one? Is it to be formative, summative, ascriptive, or both? Should it yield grading, ranking, or profiling? Are recommendations, explanations, faultfinding, or predictions requested or expected? How about postreport help with utilization? (If not, offer it). Now is the time to identify (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; and (iv) who else will see, will have the right to see, or should see the results and/or the raw data. It's best to discuss what's feasible and clarify your commitment only after doing a quick run through the KEC. Be sure to note later any subsequently negotiated changes in any of the preceding. Acknowledgments/thanks.
III. Methodology	Examples of questions that have to be answered here: 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? If a sample, how selected, and if stratified, how stratified? If none of these, how will you determine causation of effects by the evaluand? Will/should the evaluation be goal-based or goal-free? If judges are to be involved, what bias controls (for credibility as well as validity)? How will you search for side effects? Identify, as soon as possible, other investigative and data-analytic procedures used in this evaluation and their justification (may require a lit review). Hence, provide the "logic of the evaluation," i.e., general justification of its design.

PART B: FOUNDATIONS	
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., people or groups or organizations that assisted in implementation, e.g., with funding to housing); (ii) enabling and any more recent relevant legislation and any policy 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); (iv) general results of lit review on similar interventions (including "fugitive" studies (those not published in standard media) and the Internet (including the "invisible web" [e.g., by using Copernic Personal Agent ($29, WinTel only[1]) to access it]); (v) previous evaluations, if any; (vi) their impact, if any.
2. Descriptions and Definitions	Record any official description of program + components + context/environment, but provide 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 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 (or implicit in participants' accounts) or that occur to you; these are potential descriptions and may be more enlightening than literal prose if 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	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 the 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'etre of the program). The funding agency, taxpayers, and political supporters, who are also impactees in some sense, are also treated differently (and called upstream impactees or, sometimes, stakeholders, although that term is often used more loosely to include all impactees), except when they are also direct recipients. 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 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 social-relational assets of the program—include the abilities, knowledge, and goodwill of staff, volunteers, community members, and other supporters. This includes 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 reviews. 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?

[1] WinTel = Windows (operating system) and Intel (processor) computers

5. Values	Identify the relevant values for evaluating this evaluand in these circumstances from the following list, and add "stars and bars" 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. (They might be expressed qualitatively [e.g., letter grades] or quantitatively [e.g., points on a ten-point scale]). 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), values, and standards or on dimensions of valued performance and may additionally include holistic bars or stars.[2] In serious evaluations, it may also be appropriate to establish "steps:" the points or intervals on 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 [consider the Resources checkpoint]) (ii) criteria of merit from the definition of the evaluand and from standard usage (e.g., these typically include numbers impacted by the program and average/range of depth of impact) (iii) legal and (iv) ethical requirements (they overlap), including (reasonable) safety, perhaps privacy, for all impactees (v) fidelity to alleged specs ("authenticity"—this is often usefully expressed via an "index of implementation") (vi) personal and organizational goals/desires, if not goal-free (vii) professional standards of quality that apply to the evaluand (viii) logical requirements (e.g., consistency) (ix) sub-legal legislative preferences (x) scientific feasibility (xi) technological feasibility (xii) marketability, where relevant/important (xiii) expert judgment (xiv) historical/traditional/cultural standards (xv) political feasibility, if you can establish it BRD (beyond reasonable doubt) (xvi) consistency with the supposed program model (if BRD). Of course, identifying/applying some of these is unimportant in some cases, crucially important in others, and it will often require expert advice and/or impactee/stakeholder advice. NOTE: You must include any values that you will use in evaluating the side effects (if any) here, not just the intended effects (if any). Some of these values will of course, occur to you only after you find the side effects, 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.

[2] Example: The candidates for admission to a graduate program may meet minimum standards in each respect for which these were specified, 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 thinks they are not good enough for the program. We can describe this as a case where they failed to clear the holistic bar (which was implicit in this example, but can often be made explicit through dialog).

PART C: SUBEVALUATIONS	
Each of these involves (i) a fact-finding phase, followed by (ii) the work on combining the facts with whatever values (from 5 above) bear on those facts, which yields the subevaluation. In other words, Part C makes the step from What's So? to So What?	
6. Process	Assessment of the m/w/s of everything significant that happens or applies before true outcomes emerge, especially: goals (if you're not operating in goal-free mode) that may have changed or be changing; design (including consideration of design for resilience under environmental or political or fiscal duress); implementation fidelity (i.e., degree of implementation of supposed program, a.k.a. "authenticity"); accuracy of official name, subtitle, or description of program (e.g., "an inquiry-based science education program for middle school"); management; activities; procedures; learning; attitudes/values; morale; perhaps also, if you're covering this, 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. results or impact) such as knowledge 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 role is that of an intermediate 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 or impacts; adjust use 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 identification of the true operating logic/theory of program operation, by contrast with (i) the original and (ii) the current, (iii) the official, and (iv) the implicit logics. Remember that the most important outcomes may have been unintended, even unanticipated; these may be side effects (which affect the target population) or side impacts (i.e., impacts on nontargeted populations) or both. Remember that success cases may require their own treatment, regardless of average improvement (since the benefits to them 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 ones.

[3] The Book of Causes shows, when opened at the name of a factor or intervention, (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. Since the BofC is a virtual book, you have to create these pages, using all your resources for the required literature/Internet search.

8. Costs	Cover both money and nonmoney costs, both direct and indirect, both actual and opportunity costs; itemize by developmental stage, i.e., start-up/ maintenance/upgrade/shutdown and/or by 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, common labor (when these are not accessible via the market) and stress, political and personal capital, and environmental impact, which are rarely fully coverable by money.
9. Comparisons	Other means for getting the same or similar benefits from about the same or lesser resources. These are known as the "critical competitors." It is usually worth looking for one much weaker but nearly as effective alternative (el cheapo) and one much stronger although costlier alternative (el magnifico) that produces many more payoffs or process advantages; and it's sometimes worth comparing the evaluand with a widely adopted/admired approach that is perceived as an alternative, though not really in the race, e.g., a local icon.
10. Generalizability	(A.k.a. exportability; roughly the same as Campbell's "external validity;" but also covers sustainability, durability, resilience.) Can the program be used with similar results with other content, at other sites, with other staff, with other recipients, in other climates (social, political, physical), etc.? Generalization to later times is durability, and it is almost always crucial to consider this. If this checkpoint covers the financial, social, spatial, temporal, environmental, political, and other nonmoney costs, capacity, and conditions for survival, this yields the sustainability (a.k.a. "resilience to risk") rating, even more important than durability; for example, when evaluating international or cross-cultural developmental programs. NOTE: What you're generalizing about, then, is "the program in context," preferably specified, which in turn includes infrastructure.
PART D: CONCLUSIONS	
11. Overall Significance	Combine the subevaluations of Part C into an overall evaluation, i.e., at least into a profile (this is a 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.—the other relevant values). 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, Explanations, and Predictions	Microrecommendations—e.g., those concerning minor management and equipment choices/use—are often possible and supportable at no extra cost/effort (we say they "fall out" from the evaluation), and they are often very useful. But macrorecommendations, which are about the disposition of the whole program (fund, cut, modify, export, etc.)—usually require (i) extensive knowledge of the context of decision for the program decision makers (who are not always the clients for the evaluation and who may be unwilling or psychologically unable to provide this); (ii) considerable extra effort; and (iii) possession of correct (not just believed) logic or theory of the program and of the decision space, key parts of which often are not available to anyone, including the most expert of experts, in the present state of the art on that particular topic, or only available to a board of directors or to select legislators or perhaps only to their psychotherapists; hence, inaccessible to the evaluator. Because of these extra requirements, providing explanations is often done at the expense of doing the basic evaluation task well, a poor trade-off in most cases. Note that macrorecommendations typically also require the ability to predict the results of recommended changes, something that a program theory (like many social science theories) often is 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 and useful. Plain *predictions* are also often requested (e.g., Will the program work with the recommended changes?) and are usually 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. The fact that clients expect/request explanations—macrorecommendations—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 or trying to do these extra things if you lack the strong extra requirements for that OR if that effort jeopardizes the primary task of the evaluator, viz. drawing an evaluative conclusion about the evaluands.
13. [possible] Responsibilities and Justifications	If any can be determined and if appropriate (some versions of accountability that stress the accountability of people do require this). Allocating blame or praise requires extensive knowledge of (i) the main players' knowledge-state at the time of 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 evident, have major consequences, and should be pointed out.
14. Report and Support	Conveying the conclusions in an appropriate way. Not to be identified with handing over a semitechnical report, the paradigm for typical research studies. May require radically different presentations to different audiences; these may be oral or written, long or short, public or private, technical or nontechnical, graphical or textual, anecdotal or bare bones. Should include postreport help, e.g., handling questions; explaining the report's significance for different groups including users, staff, funders, other impactees. This may involve creation and depiction of various possible scenarios that do or do not accommodate the findings in the given context, i.e., doing some problemsolving for the client. In this process, a wide range of communication skills are often useful, e.g., use and reading of body language, understanding the cultural iconography. This checkpoint should also cover getting the results (and incidental knowledge findings) into the relevant databases, if any; recommending creation of one where beneficial; and dissemination into wider publication channels if appropriate.
15. Metaevaluation	I.e., evaluation of this evaluation to identify its strengths/limitations/other uses. This should always be done, as a separate quality control step (i) to the extent possible, by the evaluator, certainly—but not just—after completion of the final report and (ii) whenever possible *also* by an external evaluator of the evaluation (a metaevaluator). The primary criteria of merit for evaluations are validity, along with utility (usually to clients, audiences, and stakeholders) and credibility (to select stakeholders, e.g., funders and usually also to program staff). Can obtain relevant input to the metaevaluation by having yourself (first) and then the metaevaluator (i) apply the KEC list to the evaluation itself and/or (ii) use a special metaevaluation checklist (there are several available) and/or (iii) replicate the evaluation as done and compare the results and/or (iv) do the evaluation using a different methodology and compare the results and/or (v) 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 a weak indicator of the former. (b) Implementation does not prove high usability. (c) Only usability, not use, is applicable to evaluations without recommendations, a category that includes many important, complete, and influential evaluations. (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, recommend creating one. Remember that effects of the evaluation are not effects of the program; an empowerment evaluation produces substantial gains in the staff's knowledge about evaluation, but that's not an effect of the program. 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. 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. Careful design can bring it either near to zero or ensure that there are benefits that more than offset this cost.

MAIN NOTE C: The two checkpoints marked "possible"—12 and 13—are not always relevant or feasible and always require extra time/costs. They are mentioned because they are often supposed to be obligatory or obviously part of any professional evaluation—which is not true, since black box evaluation is often useful and often all that's possible—and supposed to provide some guidance, if they are feasible and desired by the client. And that is only true, roughly speaking, if the evaluator can better see implications of the evaluation results for program improvement than the program designer or manager—a condition that often is not true.

MAIN NOTE D: 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 edition, Sage, 1991), under Key Evaluation Checklist or the item 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 is significantly revised and improved over 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.