

THE LOGIC AND METHODOLOGY OF CHECKLISTS

Michael Scriven
Western Michigan University

June 2000, revised October 2005

Procedures for the use of the humble checklist, while no one would deny their utility, in evaluation and elsewhere, are usually thought to fall somewhat below the entry level of what we call a methodology, let alone a theory. But many checklists used in evaluation incorporate a quite complex theory, or at least a set of assumptions, which we are well advised to uncover—and the process of validating an evaluative checklist is a task calling for considerable sophistication. Interestingly, while the theory underlying a checklist is less ambitious than the kind that we normally call program theory, it is often all the theory we need for an evaluation. This memo covers some of the basic features of checklists and their application in evaluation, but it does not claim to exhaust their logic or methodology.

Basic Concepts

A **checklist** is taken here to be a list of factors, properties, aspects, components, criteria, tasks, or dimensions, the presence or amount of which are to be separately considered, in order to perform a certain task. There are many different types of checklist, although they have at least one nondefinitional function in common—that of being a mnemonic device. This function alone makes them useful in evaluation, since the nature of evaluation calls for a systematic approach to determining the merit, worth, etc., of what are often complex entities. Hence, a list of the many components or dimensions of performance of such entities is frequently valuable, and to judge from the results, even professional evaluators often forget key elements that should be included in systematic evaluations.

Checklists are of various kinds: at the bottom of the checklist peck order, there is the eponymous **laundry list**, which is almost entirely a mnemonic device and nonetheless useful for that. Notice that the order in which one calls on the items on a laundry list does not affect its validity: we can just start by entering on the list whatever items are at the top of the laundry pile. But the entry of entities into the right category on the list is crucial—to avoid the equivalent of keyboarding errors in empirical data entry. And the grouping of items, when constructing the list, is often quite important, e.g., shirts with possibly bleeding colors need to be kept separate from white shirts. Of course, a real laundry list is not an evaluative list, but plenty of "laundry lists" are used in evaluation, and one of these is given later.

Next is the sequential checklist, where the order does matter. The first kind of these is what we might call the **strongly sequential** kind, where the sequencing (of some or all checkpoints) must be followed in order to get valid results. One example of this is the preflight checklist, whose use is compulsory, not merely recommended, for the flight crews on aircraft carrying hundreds of thousands of passengers a day. It is sequential because, for example, the accuracy of the reading of instrument A depends on whether or not the setting on instrument A has been zeroed, so one must do the setting before the reading. The use of the preflight checklist is evaluative because it is designed to provide support for the evaluative conclusion that the plane is (in certain crucial respects) in good enough condition to fly safely. Many sequential checklists, however, are not intrinsically evaluative, although they might nevertheless be used in the course of an evaluation. Flowcharts often imply one or more sequential checklists, but they are often a

better way to represent inference chains that involve extensive conditionals (i.e., "if-then" statements) as well as sequences.

A **weakly sequential** checklist is one where the order is of some importance, but for psychological or efficiency reasons rather than from logical or physical necessity. Example: In the early days of the development of *The Program Evaluation Standards*, Dan Stufflebeam recalls Lee Cronbach making a strong argument that the first group of these standards should not be the Accuracy ones that were the obvious candidates, but the Utility ones, because—as Cronbach saw it—people were getting sick of evaluations that might be accurate but showed every sign of being, and usually turned out to be, useless. Convince them that evaluations were going to be useful, he argued, and you would get their attention when you turned to matters such as accuracy.

Efficiency considerations can also suggest a certain ordering within a checklist. For example, if experience reveals that a required level of performance on a particular dimension of merit—perhaps a certain minimum productivity figure—is the one most commonly failed by candidates in a recurrent competition, efficiency suggests putting it first in the order since that will eliminate the need to spend time and possibly other resources checking out the performance on other criteria of those candidates that flunk this requirement. Again, this will be a weakly ordered (sequential) checklist.¹

An **iterative checklist** is sequential, in whole or part, but requires—or may require—multiple passes in order to reach a stable reading on each checkpoint. The Key Evaluation Checklist (KEC), one of those provided at The Evaluation Center's Checklists Web site,² is iterative. Used for evaluating a program, it places the Cost checkpoint ahead of the Comparisons checkpoint, because until one has determined the cost of something, it's hard to determine what alternatives to it should be considered. But after going further down the checklist, one may be led to think of still further alternatives for the comparison group. This does no harm, by contrast with the situation in the strongly sequential preflight checklist—one can still correct the tentative conclusions on the Comparisons checkpoint. Hence, the KEC is not strongly, but weakly sequential.

Another type of checklist, one that is sometimes but not always sequential, is based on flowcharts. This is the **diagnostic checklist** that is used by—for example—mechanics, taxonomists, and toxicologists. It typically supports a classificatory kind of conclusion—one that is descriptive, not evaluative—but the conclusion is sometimes evaluative. This may be because the checklist is explicitly evaluative, e.g., a trouble-shooting list whose conclusions are necessarily faultfinding and hence evaluative ("The problem with this engine seems to be that the fuel injector nozzles are seriously worn"; "The culprit in this death seems to be overexertion"). Or the checklist itself may not be evaluative, but the context of use justifies certain types of evaluative conclusions, e.g., "This specimen is too badly damaged to make a final classification possible." It is worth noting that the diagnostic checklist, although it may not itself be couched in theoretical terms, often leads us to causal conclusions because it is often

¹ This point is applied to evaluations using the Fire at the Horses First rule, covered under that heading in *Evaluation Thesaurus* 4e, Sage, 1991, by the present author.

² evaluation.wmich.edu/checklists

theory-based under the surface (e.g., based on a limited theory about the *modus operandi* of a poison).

Probably the most important kind of checklist for evaluation purposes is the **criteria of merit checklist** (hence, COMlist or, here, **comlist**). This is what judges use when rating entries in a skating or barbeque³ or farm produce competition; it's what the evaluator uses—or should be using—for evaluating teachers or researchers or colleges or funding requests. (Or, for that matter, when teachers or researchers are evaluating evaluations and evaluators.) A number of comlists are available from the Checklists Web site, for example, some for evaluating teachers, one for evaluating systems for evaluating teachers, and one for evaluating evaluations. Comlists are widely used as the basis for a particular scoring procedure: the criteria are given weights (e.g., on a 1-5 scale), the candidates are given performance scores on a standard scale (e.g., 1-10), and the sum of the products of the weights (of each criterion by the performance on that dimension) for each candidate is used as the measure of merit. However, comlists can be used with benefit without using this particular scoring procedure (the "numerical weight and sum" or NWS procedure), so their value is (fortunately) not dependent on the known invalidity of that scoring procedure. The comlist is often a tough item to develop and validate: it has to meet some stringent requirements that do not apply to the simpler types of checklist discussed so far. For example, it is essential that it is complete, or very close to it, i.e., that it include every significant criterion of merit. Otherwise, something that scores well on it may be quite inferior because of its poor performance on some missing but crucial dimension of merit. Again, the criteria in a comlist should not overlap if it is to be used as a basis for scoring, to avoid "double counting" of the overlap area.

Before going into more details about the logic of comlists, however, we have by now covered enough examples to support some general conclusions on the pragmatic side, worth mentioning before the hard work starts. (In what follows, the term "evaluand" is occasionally used to refer to whatever is being evaluated.)

The Value of Checklists

1. Checklists are mnemonic devices, i.e., they reduce the chances of forgetting to check something important. They reduce errors of omission.
2. Checklists in general are easier for the lay stakeholder to understand and validate than most theories or statistical analyses. Since evaluation is often required to be credible to stakeholders as well as valid by technical standards, this feature is often useful for evaluators.
3. Checklists in general, and particularly comlists, reduce the influence of the halo effect, i.e., the tendency to allow the presence of some highly valued feature to overinfluence one's judgment of merit. Checklists do this by forcing the evaluator to consider separately and allocate appropriate merit to each of the relevant dimensions of possible merit. Notes: (i)

³ At "the Royal"—the crown of the competitive BBQ season in Kansas City, where only winners of the major regionals are eligible to enter—the judges use one of the simplest examples of a decision-controlling comlist. All entries (called "Qs") are rated on (1) Appearance, (2) Tenderness, and (3) Taste, with equal weight to each.

they do not eliminate the use of holistic considerations, which can be listed as separate criteria of merit; (ii) halo effect is still possible, so order should be considered carefully to reduce it. This is a further reason for (weak) ordering of checklists.

4. Comlists reduce the influence of the Rorschach effect, i.e., the tendency to see what one wants to see in a mass of data. They do this by forcing a separate judgment on each separate dimension and a conclusion based on these judgments.
5. The use of a valid comlist eliminates the problem of double weighting when using an informal list.
6. Checklists often incorporate huge amounts of specific knowledge about the particular evaluands for which they have been developed. Look at the Stufflebeam checklist for evaluation contracts, for example: it is based on, and manifests, a huge amount of experience. Roughly speaking, this amount is inversely proportional to the level of abstraction of the items in the checklist. (Example: the preflight checklist for any aircraft is highly type-specific.) Hence, checklists are a form of knowledge about a domain, organized so as to facilitate certain tasks, e.g., diagnosis, evaluation.
7. In general, evaluative checklists can be developed more easily than what are normally described as theories about the domain of the evaluand; hence, we can often evaluate (or diagnose, etc.) where we cannot explain. (Example: yellow eyes and jaundice.) This is analogous to the situations where we can predict from a correlational relationship, although we cannot explain the occurrence of what we predict. (Example: aspirin as analgesic.)

For these and some other reasons to be developed later, checklists can contribute substantially to (i) the improvement of validity, reliability, and credibility of an evaluation and (ii) our useful knowledge about a domain. Now, we return to some further development of the logic of the comlist.

Requirements for Comlists

Most of the following are self-explanatory, and refer to the criteria or checkpoints that make up a comlist:

1. The checkpoints should refer to **criteria** and not mere indicators (explained below).
2. The list should be **complete** (no significant omissions).
3. The items should be **nonoverlapping** (if list is used for scoring).
4. The criteria should be **commensurable** (explained below).

And of course,

5. The criteria should be **clear** (a.k.a. comprehensible, applicable).
6. The list should be **concise** (to assist its mnemonic function).

7. The criteria should be **confirmable** (e.g., measurable or reliably inferrable).

The first of these requirements is crucial and needs the most explanation. Suppose you are evaluating wristwatches in order to buy one for yourself or a friend. Depending on your knowledge of this slice of technology, you might elect to go in two directions. (i) You could use indirect indicators of merit, such as the brand name or the recommendations of a knowledgeable friend; or (ii) you could use criteria of merit, which essentially define the merit of this entity. They are sometimes called direct indicators of merit or primary indicators of merit. Their epistemological status is superior, but practically they are often less convenient, because they refer to characteristics which are both more numerous and less accessible than many indirect or secondary indicators.

For example, many people think that the brand name Rolex is a strong indicator of merit in watches. If you do believe that (or if you only care how the gift is perceived, not how good it is in fact), you just need a guarantee that a certain watch is a genuine Rolex in order to have settled the merit issue. That guarantee is fairly easily obtained by getting a reputable dealer to examine the interior of the watch (the amateur is easily misled), leaving you with only aesthetic considerations to get you to a purchase decision. However, if you want to get to the real truth of the matter without making assumptions, you will need to have (i) a comlist; (ii) good access to evidence about the performance of several brands of watch on each checkpoint in the comlist; and (iii) a valid way to combine the evidence on the several checkpoints into an overall rating. None of these are easy to get.

However, a conscientious evaluator can hardly rely on secondary indicators of merit with respect to the principal evaluands they are tasked to report on. They are obliged to go the route using criteria of merit, so they typically need to be good at developing (or finding and validating) comlists. This approach has its own rewards: for example, it quickly uncovers the fact that Rolex makes extremely poor watches by contemporary standards of e.g., time-keeping accuracy, or durability, or nocturnal readability—"extremely poor" means getting less than 10 percent of easily achievable standards of merit on each of these—and charges several hundred to several thousand percent more for the watches than a brand that is competitive on merit. What you pay for in a Rolex is their massive advertising campaign and the snob value. Apart from the waste of money in buying one, in terms of true merit, there is also the fact—a good example of a bad side-effect—that you considerably increase the chance of being robbed or carjacked.

A comlist for wristwatches, or anything else you are thinking of buying, begins with what we can call the core comlist, defining the general notion of merit in wristwatches, to which we can add, as a guide to purchase, any personal or special-group preferences such as affordability, aesthetic, or snob-value considerations—the "personal criteria of merit." In evaluating programs for some agency, the professional evaluator's typical task, the personal criteria have no place (you're not going to buy the program, and you're probably not going to even use its services), and hence we focus more closely on the core comlist. When *Consumer Reports* is evaluating wristwatches or other consumer products, they similarly deal only with the core comlist, leaving the rest up to the reader. Now, what does a core comlist look like for wristwatches?

1. Accuracy. Roughly speaking, this can be taken to require, at a minimum, accuracy within less than a minute a month. Most busy people will prefer to cut this in half, which reduces the resets to about three a year (background considerations of fact include the margin of error in making connections at airports, catching trains that run a close schedule, etc.).

Idiosyncratically, others will demand something considerably better, since an accuracy of better than a second a century is now available at under \$30 (in watches radio-controlled by the National Bureau of Standards). Many Japanese and German as well as Swiss movements can now manage a second a month without radio control, so one should really consider a minute a year to be the maximum allowable inaccuracy for anything that is to qualify as a good watch by modern standards. The Rolex is proudly advertised as officially certified as a chronometer, a standard from pre-quartz crystal days that is far worse than any of those just mentioned.

2. Readable dials. Some of Rolex's "jewelry watches" for women are very hard to read. No Rolexes meet modern standards, since their luminous paint fades after an hour or two. The Luminox breakthrough means that watches readable at night from the other side of the room are available at mid-level prices (around \$100).
3. Durability (of watch and fittings). Should survive dropping onto wooden floor from 4 feet, tile or concrete from 3 feet, the most common accidents. Band should survive more than 2 years (leather usually does not). Case should be heavy rain-proof.
4. Comfortable to wear. Gold is usually too heavy and steel comes close. A titanium bracelet is best.
5. Band should be easily adjustable, without help from a jeweler. (Fit depends on temperature, diet, etc.)
6. Low maintenance. Batteries should last several years, if used; routine servicing the same. Rolex does not use batteries, but recommended cleaning and servicing is frequent and extremely expensive. The Seiko self-winder uses no batteries and requires negligible and cheap maintenance.

Each of these claims requires some data gathering, some of it quite difficult to arrange. To these criteria of merit, we would often need, for personal use, to add idiosyncratic requirements about appearance, and features, e.g., luminous hands, stopwatch or alarm functions, subsurface waterproofing, and cost.)

By contrast, we could use an indicator list (*indlist*) with items like this:

1. Made by Rolex

Evidence for this, easy to get, would be that it was sold by an authorized Rolex dealer, who guaranteed it in writing and by serial number. The validity of this indicator, as of any secondary indicator, is (roughly) the correlation between it and the cluster defined by the first set of six indicators. The hints provided make it clear that this correlation is low. However, before getting too set on the high horse, it's worth remembering that there are many occasions when you can't get at criteria of merit but you can get at indicators for them; and that even when you can get both, the indicators may be much easier and/or less costly to get. But keep in mind that indicators are easily corrupted and once it becomes known that they are being used as an indicator for something important in an evaluation, people are very ingenious about faking the score on them; you can't do this with a criterion of merit, by definition.

Criteria vs. Indicators

Given that the path of righteousness for evaluators is the path of criteria, not indicators, how do we identify true criteria for the evaluand X?

The key question to ask is this: What properties are parts of the concept (the meaning) of "a good X?" Note: In general, you will not get good results if you start by identifying the defining criteria for X itself, and try to go from there to the criteria for "good X." Thus, in our example, to call something a good watch is to say that it tells the time accurately, is easy to read, durable, comfortable to wear, etc. You won't get to a good comlist by starting off with the comlist for "watch," which starts off with: "wearable on the wrist," since extremely small is extremely wearable but not extremely desirable in a watch.

Is this to say that a watch that misses on one of these criteria is by definition not a good watch? Not quite.⁴ A watch that is rather fragile, for example—enough so that one would not call it durable—but excels on the other criteria, would probably be called "good but not great." Still, that failing raises some doubt about whether we should really call it a good watch, and any more shortcomings would make us hesitate even more. A criterion of merit is one that bears on the issue of merit, sometimes very heavily (so that a failure on that criterion is fatal) but often just in the sense of being one of several that are highly relevant to merit, although not—in itself—absolutely essential.

How does one validate a checklist of criteria of merit? Essentially, one begins by trying for completeness, i.e., by trying to construct hypothetical cases in which an entity has all the properties in the proposed comlist but still lacks something that would be required or important in order to justify an assignment of merit. Looking at the above checklist for a watch, for example, one might say, "Well, all that would get you a watch that ran well if you stayed home all the time . . . but suppose you have to fly from one part of the country to another. That will require you to reset the time, and there are watches where that is a virtually impossible task unless you carry an instruction book with you (e.g., the Timex Triathlon series). Surely, that flaw would lead you to withhold the assignment of merit?" That's a pretty good argument, and I think it shows we need to add one more criterion of merit. So we now have the following: (Can you see other loopholes? There is at least one minor one.)

1. Accurate
2. (Easily) Readable
3. Durable
4. Comfortable
5. (Easily) Adjustable for fit
6. Highly Autonomous (Batteries, Cleaning, Repair)

⁴ See "The Logic of Criteria" Michael Scriven, *The Journal of Philosophy*, Vol. 56, October 22, 1959, pp. 857-868.

7. (Easily) Settable

Some things are taken for granted in these lists. For example, we could add the requirements that the watch does not emit evil radiation, does not induce blood poisoning or skin eruptions, etc. We simply put those into the general background for all consumer products, not thereby belittling them—there are documented cases of radiation damage from the early days of luminous dials. But these possibilities—there are many more—would extend comlists beyond necessity. We can deal with such cases as context, and in detail only as and when they arise.

There are other interesting issues, which we pass over here. For example, should luminous dials be taken as an extension of readability, as an idiosyncratic preference, or as an entry under an additional heading: 8. Versatility. Should some standards of modern design that transcend issues of function be incorporated, and if so, what standards: i.e., should there be a Checkpoint 9. Aesthetics?

Evaluative Theories

We've already stressed the informational content of checklists. For example, the watch checklist exhibits knowledge of the components of watches; the contracting checklist exhibits considerable knowledge of the process whereby organizations approve contracts. Now what theory underlies the watch comlist? It's not a theory about how watches work, but about what they need to do well in order to perform their defining function well. And that may be just the kind of theory, and it may be the only kind of theory, that we need for evaluation purposes.

These "evaluative theories" are not as ambitious as an explanatory theory of the total operation (including dysfunction) of the evaluand, something that is more than anyone can manage with many complex evaluands such as large educational institutions. But it's not so hard to say what such an institution has to do in order to be regarded as meritorious: not a trivial task, but much easier. One attraction about an evaluative theory is thus that it's much easier to give good evidence for its acceptability than it is to demonstrate the truth of an explanatory theory.

Those who favor an outcomes-based approach to program evaluation will perhaps be particularly attracted to this kind of theory, because of the emphasis on performance. However, it can easily include process variables—such as comfort in wearing a watch.

It is true that evaluative theories—in a sense, the underpinnings of comlists—are not particularly versatile at generating explanations and recommendations, where program theories are supposed to excel, if you are lucky enough to have a valid one. But they do have a trick up their sleeves under this heading: they are outstandingly good at one valuable aspect of formative evaluation—identifying the areas of performance that need attention.

Criteria, Subcriteria, and Explanatory Text

The richness and value of a comlist is often greatly increased by unpacking some of the criteria. In particular, their value in formative evaluation can be greatly improved by this procedure. Here are the main headings from the comlist for evaluating teachers advocated in another paper of mine available on this site:

1. Knowledge of subject matter
2. Instructional competence
3. Assessment competence
4. Professionalism
5. Nonstandard but contractual duties to school or community (e.g., chapel supervision)

Not too controversial, but also not tremendously useful. It's still a long way from the trenches; more at home in the general's tent than the drill-sergeant's playbook. Let's look at how one might expand the second entry here, so that we'd have something that can really make distinctions between the better and the weaker teachers.

2. Instructional competence
 1. Communication skills (use of age-appropriate vocabulary, examples, inflection, body language)
 2. Management skills
 1. Management of (classroom) process, including discipline
 2. Management of (individual student's educational) progress
 3. Management of emergencies (fire, tornado, earthquake, flood, stroke, violent attack)
 3. Course construction and improvement skills
 1. Course planning
 2. Selection and creation of materials
 3. Use of special resources
 1. Local sites
 2. Media
 3. Specialists
 4. Evaluation of the course, teaching, materials, and curriculum

Now we can see more clearly what's being included. And now we're much close to being able to apply the checklist. However, in the publication where this appeared, experience led us to add 8,000 words of more specific detail, some for each sub-criterion, in order to complete a working checklist. This points up one feature of the use of checklists that has to be kept in mind: the balance between ease of and value added via applicability, on the one hand, and length on the other. Brevity is desirable; but clarity is essential—especially, of course, when people's careers or other highly important matters are at stake.

The second matter that can perhaps be illuminated using this example is the criterion (for checklists) of commensurability. What this means is that headings at one level of a checklist have to be at roughly the same level of generality. In the present example, there are four levels of headings. Looking at any one set, in its location under a higher-level heading, one can see that they are all of the same level of specificity. The other side of the commensurability coin is that one must pay some attention to the function of the checklist in grouping and naming subheadings. For example, in the laundry list itself, if the function is to control the actions of the laundry person, colored articles need to be listed separately from the white ones. But if the

function is simply to make a record of what went to the laundry, the color of the shirts is irrelevant.

Another matter that requires close attention when building checklists into one's methodology is intelligence, including thoughtfulness, in the application of checklists. Daniel Stufflebeam reports on a pilot whose considered judgment was that some pilots he had flown with focused on covering the preflight checklist in the sense of checking items off on it, but not on the meaning of the checkpoints, thereby creating serious risks.

The Use of Comlists for Profiling and Scoring Purposes

Possibly the most important use of checklists in evaluation involves using them as the basis for assessing and representing the overall merit, worth, or importance of something. In rating decathletes, for example, we can simply set up a graph in which each of the ten merit-defining events is allocated a half-inch of the horizontal axis, while their best score in each event is represented by a (normalized) score in the range 1-10 on five inches of the vertical axis. Using this kind of bar graph is called profiling and is a very useful way to display achievement or merit, especially for formative evaluation purposes. However, it will not (in general) provide a ranking of several candidates; for that, we need to amalgamate the subscores into an overall index of some kind. In the decathlete case, this is easily done: we allot equal weight to each performance (since that is how the decathlon is scored) and add up the normalized performance scores. The athlete with the top score is the best selection; the second highest score identifies the runner-up, etc.

But in program evaluation and most personnel evaluation, matters are not so easy. One often feels that different criteria of merit deserve different weights, but it's very hard to make a case for a quantitative measure of that difference certainly for a precise measure. Worse, the use of a single weight for each criterion of merit is an oversimplification. It is often the case that a certain level of performance on criterion N is much more important than a certain level of performance on criterion M, but that increments above that level on N are no more important than increments on M. In other words, the value or utility function is not a linear function of performance. If that is so, what kind of function is it? Evaluators might begin to feel out of their depth at this point. The following remarks may be helpful.

1. Do not abandon equal weighting without overwhelming evidence. In the first place, it may not be exactly right, but it may be the best approximation. In the second place, even if it's not the best approximation, results based on this assumption may be highly correlated with results based on the correct function/weighting, and if you can't determine the latter, so it's this way or the highway.
2. If you are certain that N is more important, throughout its range, than M, make a simple intuitionistic estimate of the difference, as the basis for a trial exploration of its effect. But do this very cautiously. At first, consider whether to use the factor 1.5 rather than 2, and almost never go beyond the ratio of 2. It is extremely hard to justify a higher ratio than 2, to others: "If 3, why not 4?" is hard to refute.
3. If the ratio you pick seems not to apply constantly across the whole range of performance on a particular criterion, try varying it *for a certain interval*.

4. Testing your attempts to set differential weights requires some judgment about whether the results show it to have been a success or failure. Do this by inventing and considering a range of hypothetical cases, in order to see whether they lead to implausible results, i.e., look into the robustness of your weights. You are likely to find out quickly that large differences in weights allow easy counter-examples to be created.
5. A procedure that combines qualitative weighting with minimalist quantitative procedures, called the Qualitative Weight and Sum (QWS) is set out in the present author's *Evaluation Thesaurus* (Scriven, 1991) and refined somewhat in E. Jane Davidson's (2004) *Evaluation Methodology Basics*.

Conclusion

Laundry lists, sequential checklists, and comlists all serve important roles in evaluation. A basic logic, covering only some of their properties, has been set out here, in the hope it may lead to increased attention and improved utility of checklists. Suggestions for improvement and expansion, as well as good examples, would be much appreciated (send to scriven@aol.com). Some will be incorporated into later posted and dated editions, with acknowledgments.

References

- Davidson, E. J. (2004). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.
- Scriven, M. (1991). *Evaluation thesaurus*. Newbury Park, CA: Sage.