Checklists

In: Encyclopedia of Evaluation

By: Michael Scriven
Edited by: Sandra Mathison
Book Title: Encyclopedia of Evaluation
Chapter Title: "Checklists"
Pub. Date: 2011
Access Date: October 3, 2018
Publishing Company: Sage Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780761926092
Online ISBN: 9781412950558
DOI: http://dx.doi.org/10.4135/9781412950558
Print pages: 54-60

Procedures for the use of the humble checklist, although 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. However, many checklists used in evaluation incorporate a quite complex theory, which we are well advised to uncover—and the process of validating an evaluative checklist is a task calling for considerable sophistication. It is interesting that although this theory is less ambitious than the kind that we normally call program theory, it is often all the theory we need for an evaluation. This entry 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 defined here as a list of factors, properties, aspects, components, criteria, tasks, or dimensions, the presence or amount of which are to be separately considered in the performance of a certain task. There are many different types of checklist, although all have at least one nondefinitional function in common—that of being a mnemonic device. This function alone makes them useful in evaluation, as the nature of evaluation calls for a systematic approach to determining the merit, worth, and so on, of what are often complex entities. Hence, a list of the many components or dimensions of performance of such entities is frequently valuable.

Checklists are of various kinds: At the bottom of the checklist pecking order 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 in a laundry list does not affect the validity of the list: We can start by entering on the list whatever items are at the top of the laundry pile. However, the entry of entities into the right place on the list is crucial to avoid the equivalent of key-boarding errors in empirical data entry. Also, the grouping of items as the list is being constructed is often quite important: For example, shirts with colors that may bleed need to be kept separate from white shirts. Note that a real laundry list is not an evaluative list, but plenty of “laundry lists” are used in evaluation, and one of these is discussed 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, wherein the sequencing (of some or all checkpoints) must be followed 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 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 important, 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 contain the Accuracy standards that were the obvious candidates but the Utility standards, 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 of 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 because that will eliminate the need to spend time checking out the performance on other criteria of those candidates that flunk this requirement. Again, this will be a weakly ordered (sequential) checklist.

An iterative checklist is sequential, in whole or part, but requires—or may require—multiple passes to reach a stable reading on each checkpoint. The Key Evaluation Checklist, one of those provided at the University of Western Michigan's Evaluation Checklist Project Web site, 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 is hard to determine what alternatives to it should be considered. After going farther down the checklist, however, 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 Key Evaluation Checklist is not strongly sequential, but weakly.

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; for example, a troubleshooting list in which the conclusions must necessarily be faultfinding and hence evaluative (e.g., "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"). If the checklist itself is not be evaluative, the context of use may still justify certain types of evaluative conclusions; for example, "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 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 barbecue or farm produce competition; it is what evaluators use—or should be using—for evaluating teachers or researchers or colleges or funding requests and what teachers or researchers use when evaluating evaluations and evaluators. At "the Royal"—the crown of the competitive barbecue season in Kansas City, which 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 (a) appearance, (b) tenderness, and (c) taste, with equal weight to each.

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 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 checklists discussed so far. For example, it is essential that it be complete, or very close to it, meaning that it must include every significant criterion of merit. Otherwise, something that scores well on the comlist 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 the list is to be used as a basis for scoring, to avoid "double counting" in the overlap area.

By now enough examples have been covered to support some general conclusions on the

pragmatic side, worth mentioning before the hard work starts.

THE VALUE OF CHECKLISTS

1.
- Checklists are mnemonic devices; that is, they reduce the chances of forgetting to check something important, and 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. Because 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" (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. Note that checklists do not eliminate the use of holistic considerations, which can be listed as separate criteria of merit.

4.
- Comlists reduce the influence of the Rorschach effect (the tendency to see what one wants to see) in a mass of data. They do this by forcing a separate judgment of each separate dimension and a conclusion based on these judgments.

5.
- The use of a valid comlist eliminates the problem of double weighting.

6.
- Checklists often incorporate huge amounts of specific knowledge about the particular evaluands for which they have been developed. Look at the 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, such as diagnosis and 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 (e.g., aspirin as analgesic).

For these and some other reasons to be developed later, checklists can contribute substantially to (a) the improvement of validity, reliability, and credibility of an evaluation and (b) 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. Criterial status (not mere indicators; see the following)

2. Complete (no significant omissions)

3. Nonoverlapping (if list is used for scoring)

4. Commensurable (explained later)

Also, of course:

5. Clear

6. Concise (mnemonic devices that can themselves be easily remembered score double points)

7. Confirmable (e.g., measurable or reliably inferable)

The first of these requirements is crucial and needs the most explanation. Suppose you are evaluating wristwatches with a view to buying one for yourself or a friend. Depending on your knowledge of this slice of technology, you might elect to go in one of two directions. (a) You could use indirect indicators of merit, such as the brand name or the recommendations of a knowledgeable friend, or (b) you could use criteria of merit, which essentially define the merit of this entity. Such criteria are sometimes called direct indicators of merit or primary indicators of merit. Their epistemo-logical status is superior; but practically, they are less convenient because they refer to characteristics that are both more numerous and less accessible than 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 care only 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 to have settled the merit issue. That guarantee is easily obtained from reputable dealers, 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 (a) a comlist, (b) good access to evidence about the performance of several brands of watch on each checkpoint in the comlist, and (c) a valid way to combine the evidence on the several checkpoints into an overall rating. None of these are easy to get.

Conscientious evaluators can hardly rely on secondary indicators of merit with respect to their principal evaluands. They are obliged to use 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 poor watches by contemporary standards and charges several hundred to 1000% more for them than a competitive brand in terms of merit. What you pay for in a Rolex is a 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 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 are not going to buy the program), and hence we focus 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 by at least 50%, which reduces the resets to about three a year. Idiosyncratically, others will demand something considerably better: As an accuracy of better than a second a century is now available at under $100 (watches radio controlled by the National Bureau of Standards), a minute a year may be considered to be the maximum allowable inaccuracy. The Rolex is certified as a chronometer, an out-of-date standard that is worse than any of those just mentioned.

2.
- Readable dials. Some of Rolex's “jewelry watches” for women are very hard to read.

3.

- Durability of watch and fittings. The watch should be able to survive being dropped onto a wooden floor from 4 feet. The band should survive more than 2 years (leather usually does not).
  4.
- Comfortable to wear. Gold is usually too heavy. A titanium bracelet is best.
  5.
- Flexibility of fit. The band should be easily adjustable, without help from a jeweler. (Fit depends on temperature, diet, etc.)
  6.
- Low maintenance. Batteries should last several years, routine servicing the same. Rolex does not use batteries, and recommended cleaning and servicing is frequent and very expensive.

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

By contrast, an indicator list could be used, 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.

CRITERIA VERSUS INDICATORS

Given that the path of righteousness for evaluators is the path of criteria, not indicators, how do we identify true criteria for an 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, is durable, is comfortable to wear, and so on.

Is this to say that a watch that misses on one of these criteria is by definition not a good watch?

Not quite. 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 themselves absolutely essential.

How does one validate a checklist of criteria of merit? Essentially, by trying to construct hypothetical cases in which an entity has the properties in the proposed comlist but still lacks something that would be required or important to justify an assignment of merit. Looking at the provided 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 is a pretty good argument, and clearly another criterion of merit is needed. 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

6.
- Inexpensively maintainable (batteries, cleaning, repair)

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, and so on. 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 (and there are many more) would extend comlists beyond necessity. We can deal with such cases 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 (Versatility)?

EVALUATIVE THEORIES

The informational content of checklists has already been stressed. 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 is not a theory about how watches work but about what they need to do well to perform their defining function well. That may be just the kind of theory that we need for evaluation purposes.

These “evaluative theories” are not as ambitious as an explanatory theory of the total operation of the evaluand, something that is more than anyone can manage with many complex evaluands, such as large educational institutions. However, it is not so hard to say what such an institution has to do to be regarded as meritorious—it is not a trivial task, but it is at least much easier. One attraction of an evaluative theory is thus that it is much easier to demonstrate its truth than it is to demonstrate the truth of an explanatory theory.

Those who favor an outcome 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—the underpinnings of comlists—are not particularly adept at generating explanations and recommendations; program theories are supposed to excel at exactly this, if you are lucky enough to have a valid one. Evaluative theories do have a trick up their sleeves, however: 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 when some of the criteria are unpacked. In particular, the value in formative evaluation can be greatly improved by this procedure. Here are the main headings from the comlist for evaluating teachers, which can be found at the Western Michigan University Evaluation Center 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 too useful. It is still a long way from the trenches. The following demonstrates how the second entry here might be expanded so that the comlist could really make distinctions between better and weaker teachers.

Instructional Competence

1.
- Communication skills (use of age-appropriate vocabulary, examples, inflection, body language)

2.
- Management skills
    a.
    - Management of (classroom) process, including discipline

    b.
    - Management of (individual students' educational) progress

    c.
    - Management of emergencies (fire, tornado, earthquake, flood, stroke, violent attack)

3.
- Course construction and improvement skills
    a.
    - Course planning

    b.
    - Selection and creation of materials

    c.
    - Use of special resources
        (1)

- Local sites
  (2)
- Media
  (3)
- Specialists

4.

- Evaluation of the course, teaching, materials, and curriculum

Now what is being included is much clearer, and we are much closer to being able to apply the checklist. However, in the publication where the original list appeared, experience led the authors to add 8000 words of more specific detail, some for each subcriterion, 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 understanding and length. 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 be illuminated from this example is the criterion (for checklists) of commensurability. What this means is that headings at one level of a checklist have to be of 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 all items in the set are 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, if the function is to control the actions of the laundry person, colored articles need to be listed separately from white ones. If, however, 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 thought-fulness 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 10 merit-defining

events is allocated half an inch of the horizontal axis, and the decathlete's best score in each event is represented by a normalized score in the range 1 to 10 on 5 inches of the vertical axis. Using this kind of bar graph is called profiling, and it 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: Equal weight is allotted to each performance (as that is how the decathlon is scored) and the normalized performance scores are added up. The athlete with the top score is the best selection, the second highest score identifies the runnerup, and so on.

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 is very hard to make a case for a quantitative measure of that difference. 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 in criterion N is much more important than a certain level of performance in criterion M, but increments above that level in N are no more important than increments in 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 is not the best approximation, results based on this assumption may be highly correlated with results based on the correct function or weighting, and you cannot determine the latter, so it is this way or no way.

2.
- If you are certain that N is more important, throughout its range, than M, make a simple intuitive estimate of the difference as the basis for a trial exploration of its effect. Do this very cautiously: At first, consider whether to use 1.5 as the factor rather than 2, and almost never go beyond that ratio. 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, to see whether they lead to implausible results. You are likely to find out quickly that large differences in weights allow for the easy creation of counterexamples.

5.

- A procedure that combines qualitative weighting with minimalist quantitative procedures is set out in the fourth edition of the Evaluation Thesaurus.

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 that it may lead to increased attention to and the improved utility of checklists.

Michael Scriven

http://dx.doi.org/10.4135/9781412950558.n79
10.4135/9781412950558.n79

Further Reading

Evaluation Checklist Project, Evaluation Center, Western Michigan University.(2004) Evaluation checklists.Retrieved May 4, 2004, from http://www.wmich.edu/evalctr/checklists/

Scriven, M. The logic of criteria. *Journal of Philosophy*56857–868(1959)

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