Weighted Checklists

## Description

Tweet

# A participatory means of measuring complex change

**On this page:**


1. What is a weighted checklist?
2. Where are (participatory) weighted checklists used?
3. What is different about (participatory) weighted checklists?
4. Potential problems
5. How do rubrics compare to weighted checklists?
6. Variable design features of weighted checklists
7. Related resources


## 1. What is a weighted checklist?

A weighted checklist has:

- *A list of items*, each of which describes an attribute of an organisation or an event. The attribute may or may not be present (indicated by a 1 or 0), or it may be present in a degree measured in a simple scale (e.g. 0 to 3).
- A *set of weights,* which describes the relative importance of each item
- A *summary score*, based on the number of items identified as present, but adjusted by their individual weights.

Here is an example of a very simple customer satisfaction survey that is in the form of a weighted checklist (that was sent to me by a firm I used). In this case the survey respondents provided two sets of information (a) their views on the importance of each item, to them (the weights), in the second column, (b) their views on how well the firm was doing on each criteria according to their experience, in the third column.

| | Importance to you<br>1= low, 7 = high | Barr Ellison's performance<br>1=low, = 7 = high |
|---|---|---|
| Quality of work carried out | | |
| Ease of contacting solicitor | | |
| Ongoing communication updates | | |
| Understanding your needs | | |
| Meeting agreed deadlines | | |
| Reception by telephone | | |
| Reception at office | | |
| Value for money | | |
| | | |

Once the responses have been collected, weighted scores for individual respondents then can be calculated, along with an average score for all respondents. The process is as follows:

1. Multiply the importance rating x actual performance rating for each item

2. The sum of these is the *actual raw score*

3. Multiply the importance rating x highest possible performance rating for each item

4. The sum of these is the *highest possible raw score*

5. Divide the actual raw score (2) by the highest possible raw score (4), to get a *percentage score* for the respondent. A high percentage = high degree of satisfaction, and vice versa

6. Calculate the average  percentage score for all the respondents

This is a *participatory* form of weighted checklist, because respondents themselves determine the weights given to different items on the checklist. Other types of checklists use weightings solicited from experts. They are not the focus of the remainder of this paper. Judging from a Google search these expert weighted checklists are mainly used for staff performance appraisal purposes. For more information on these, see:

- Performance Appraisal Tips Help Page by Dexter Hansen.  â??Weighted Checklist. â?? The term used to describe a performance appraisal method where supervisors or personnel specialists familiar with the jobs being evaluated prepared a large list of descriptive statements about effective and ineffective behaviour on jobs.â?•
- Managing Employee Performance and Reward: Concepts, Practices, Strategies, by John Shields. 2007, page 170
- Managing Human Resources in Small & Mid-Sized Companies By Diane Arthur, 1995, page 178
- Handbook of Public Personnel Administration By Jack Rabin, 1995, page 337

## 2. Where are (participatory) weighted checklists used?

#### 1. When the event is complex and difficult to measure with any single indicator

Often people try to measure a change by finding a single measurable indicator that will capture the change. For complex changes, such as improvements in peopleâ??s participation or changes in organisational capacity, finding such an indicator can be a major challenge. Often the chosen indicator seems far too simplistic. Such as using the number of people participating in x type of meetings, as an indicator of participation.

#### 2. Where there may be multiple measures in place, but a single aggregate measure is needed of overall performance

Sometimes Logical Framework descriptions of project designs will include more than one indicator to track a given change that is recognised to be complex. However this response presents a further challenge, of how to aggregate the evidence of change described by multiple indicators.

**3. Where peoplesâ?? views of the significance of what has happened differ**

Users of a health centre may have different views on how well the health service is performing compared to the health centre staff, or to the views of the senior managers of health services

The matrix below can be used to describe where three different methods are most suitable (ordinary indicators, Most Significant Change stories, and weighted checklists)

| Outcomes are | Expected | Unexpected |
|---|---|---|
| Of agreed significance | Use indicators | Use MSC |
| Of disputed significance | Use weighted checklists | Use MSC |

PS: The axes of this table are essentially the same as the â??Stacey matrixâ?•

## 3. What is different about (participatory) weighted checklists?

Weighted checklists separate out *value data* from *observational data*. In the example above, the second column asks about importance to you, the respondent. This is value data. The third column asks you about the companyâ??s performance. This is observational data.

With the use of conventional indicators judgments about importance happen only once, when the choice is made to use a specific indicator or not. This happens at the planning stage, and is set thereafter. It is not possible to change the choice of indicator later on, without losing continuity of the data that has been collected so far. With weighted checklists the same set of *observational data* can be

re-analysed with different sets of *value data*, reflecting the views of different stakeholders, or the views of the same stakeholders that might have changed over time .

Value data is *meta-information*: information about information. This can be of different kinds.  In the simple example shown in the table above, respondents are asked about their *preferences*. Another survey could ask people which items they thought were basic *rights*, which all people should have access to. This is the basis of the design of the Basic Necessities Survey (BNS). Or, a survey could ask which items would be the most important *cause* of an overall outcome e.g. improved community health. This was the subject of my posting on â??Checklists as mini theories-of-changeâ?• Because of these choices available participatory processes used to elicit checklist weightings should always be clear on what type of judgments are being sought.

Value data can be worthwhile analyzing in itself. Different groups of stakeholders will usually vary in the extent that their views agree with each other. We could measure and monitor this degree of alignment by looking at how participantsâ?? ratings in the second column of the example above *correlate* with each other, using Excel. Social Network diagrams could also be produced using the same data (in a participants x item ratings matrix) to show in more detail how various stakeholder groups are aligned with each other in their views. Of special importance in development project settings will be how the alignment of views between stakeholders changes over time. Is there a stronger consensus developing or not?

Changes over time are likely to be important in other ways as well. If a survey asks for information about the perceived importance of different health services (as well as their actual availability) the increased expectations over time might be as important an indicator of development as any increased availability of services.  Knowledge that the public were expecting more could affect the responsiveness of local politicians to their constituencyâ??s concerns. Differences between what people report as available and what is reported to be available through other sources of information could also be informative. It could highlight a lack of public knowledge of what is available, or raise questions about the validity of officialsâ?? claims about what is available.

## 4. Potential problems

#### 1. Constructing the checklist

A common challenge to the use of methods like the BNS is â??But who constructed the checklist? Surely the contents of this list, and what it omits, will affect the overall findings?â?• There are two ways of addressing this potential problem. The first is to ensure that the checklist contents are developed through a consultative process involving a  range of stakeholders, especially those whose performance is being assessed. The other is to ensure that the checklist is long enough. The BNS checklist had around thirty items. The larger the checklist the less vulnerable the aggregate score will be to the accidental omission of individual items that could be important. But there will also need to be some limits to the size of the list, because respondentsâ?? interests are likely to wane towards the end of a long list.

Long lists of items in a checklist can also present another challenge, of how to assign weightings to all of them. One way around this problem is to group the items into nested categories, and then proceed

with weightings in two stages: (a) for the main categories first, then for the individual items within each category.

#### 2. Interpreting the checklist

In a survey form it will not be easy to elicit *reasons* why people have rated one item on a checklist as more important than another. But in workshop settings this can be easier. One way of eliciting these explanations from participating stakeholders is to do pair comparisons, asking â??Why is this category of activities more important than this one?â?• Answers to this question help provide insight into people as consumers of services or citizens with rights or managers with theories-of-change about how their intervention should work.

The same problem exists with understanding the observational data, especially where there are rating scales rather than yes/no answer options. With yes/no options the main requirements it that the items on the list are clearly defined entities or events. With ratings there is the possibility of significant differences in response styles, how people use the ratings available. One common strategy is to provide the respondent with guidance on what would constitute a 0, 1, 2, of 3 on a rating scale.

#### 3. Transparency issues

There is some debate however about whether the weightings of items should be made visible to respondents before a survey, or only made visible later on, when results have been aggregated. This would not be an issue where the weightings themselves are obtained from the respondents, as is the case with the BNS survey. But it could have an influence on the survey results where weights are decided before the survey is implemented. It could lead to respondents, say health centre staff, deciding to improve one aspect of their service more than another, because they know it receives a higher weighting in the checklist. But that response may not necessarily be a bad, thing, if those aspects of service are really more important than others. Being open about the weightings could give health centres some choices about how to improve their service, in contrast to performance measurement relying on one key indicator.

Where weightings are obtained from stakeholders (including respondents) via a workshop event after the survey the effects might be less easy to predict. Participants might be inclined to argue for higher weightings for items they know they have done well on, and vice versa. Making their raw checklist scores visible during the workshop discussion could help make this tendency evident, but it is not likely to eradicate it. Structuring a debate around the proponents of different weightings might help force any apparent self-interest proposals to be justified.

## 5. Variable design features of weighted checklists

- **Who** provides the **weights**
  - Participants (whose behavior is being assessed)
  - Others
    - Experts re the behavior or events of interest
    - Different stakeholder groups
- **Who** makes **descriptions**

    - Participant (whose behavior is being assessed)
    - Others
        - Experts
        - Different stakeholder groups
- **How** are **weights** assessed
    - Binary (yes/no checklist)
    - Rating or ranking scale
- **How** are **descriptions** assessed
    - Binary (yes/no checklist)
    - Rating or ranking scale
    - Existing data

## 6. How do rubrics compare to weighted checklists?

This question is discussed in detail in this April 2020 â??Rick on the Roadâ?• blog: Rubrics? Yes, butâ?¦. That blog posting was a response to Tom Astonâ??s blog posting: Rubrics as a harness for complexity. I have reproduced it in full, hereâ?¦

â??I have just reviewed an evaluation of the effectiveness of policy influencing activities of programs funded by HMG as part of the International Carbon Finance Initiative.  In the technical report there are a number of uses of rubrics to explain how various judgements were made.  Here, for example, is one summarising the strength of evidence found during process tracing exercises:

- **Strong support** â?? smoking gun (or DD) tests passed and no hoop tests (nor DDs) failed.
- **Some support** â?? multiple straw in the wind tests passed and no hoop tests (nor DDs) failed; also, no smoking guns nor DDs passed.
- **Mixed** â?? mixture of smoking gun or DD tests passed but some hoop tests (or DDs) failed â?? this required the CMO to be revised.
- **Failed** â?? some hoop (or DD) tests failed, no double decisive or smoking gun tests passed â?? this required the theory to be rejected and the CMO abandoned or significantly revised.

Another rubric described in great detail how three different levels of strength of evidence were differentiated (Convincing Plausible, Tentative).  There was no doubt in my mind that these rubrics contributed significantly to the value of the evaluation report.  Particularly by giving readers confidence in the judgements that were made by the evaluation team.

Butâ?¦ I canâ??t help feel that the enthusiasm for rubrics seems to be out of proportion with their role within an evaluation.  They are a useful measurement device that can make complex judgements *more* transparent and thus *more* accountable.  Note the emphasis on the â??*more*â??â?¦ There are often plenty of not necessarily so transparent judgements present in the explanatory text which is used to annotate each point in a rubric scale.  Take, for example, the first line of text in Tom Astonâ??s first example here, which reads â??Excellent: Clear example of *exemplary* performance or *very good* practice in this domain: no weaknessâ?•

As noted in Tomâ??s blog it has been argued that rubrics have a wider value i.e. â??*rubrics are useful when trying to describe and agree what success looks like for tracking changes in complex phenomena* â?•. This is where I would definitely argue â??Buyer bewareâ?• because rubrics have serious limitations in respect of this task.

The first problem is that description and valuation are separate cognitive tasks. Events that take place can be described, they can also be given a particular value by observers (e.g. good or bad). This dual process is implied in the above definition of how rubrics are useful. Both of these types of judgements are often present in a rubrics explanatory text e.g. *Clear* example of *exemplary* performance or very good practice in this domain: no weaknessâ?•

The second problem is that complex events usually have multiple facets, each of which has a descriptive and value aspect. This is evident in the use of multiple statements linked by colons in the same example rubric I refer to above.

So for any point in a rubricâ??s scale the explanatory text has quite a big task on its hands. It has to describe a specific subset of events and give a particular value to each of those. In addition, each adjacent point on the scale has to do the same in a way that suggests there are only small incremental differences between each of these points judgements. And being a linear scale, this suggests or even requires, that there is only one path from the bottom to the top of the scale. Say goodbye to equifinality!

So, what alternatives are there, for describing and agreeing on what success looks like when trying to track changes in complex phenomena? One solution which I have argued for, intermittently, over a period of years, is the wider use of weighted checklists. These are described at length here.

Their design addresses three problems mentioned above. Firstly, description and valuation are separated out as two distinct judgements. Secondly, the events that are described and valued can be quite numerous and yet each can be separately judged on these two criteria. There is then a mechanism for combining these judgements in an aggregate scale. And there is more than one route from the bottom to the top of this aggregate scale.

â??The proof is in the puddingâ?•. One particular weighted checklist, known as the Basic Necessities Survey, was designed to measure and track changes in household-level poverty. Changes in poverty levels must surely qualify as â??complex phenomena â??. Since its development in the 1990s, the Basic Necessities Survey has been widely used in Africa and Asia by international environment/conservation organisations. There is now a bibliography available online describing some of its users and uses. https://www.zotero.org/groups/2440491/basic_necessities_survey/library

## 6. Related resources

- I recommend THE LOGIC AND METHODOLOGY OF CHECKLISTS by Michael Scriven Claremont Graduate University, updated in 2007. In the opening para he says

  â??*The humble checklist, while no one would deny its utility in evaluation and elsewhere, is 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. Indeed, while the theory underlying a checklist is less ambitious than the kind that we normally call a program theory, it is often all the theory we need for an evaluation.*â?•

This is a great paper, informative and a pleasure to read. Amongst other things, it gives a wider background to the use of checklists than I have provided above.

- I also recommend The synthesis problem: Issues and methods in the combination of evaluation results into overall evaluative conclusions. By Michael Scriven, Claremont Graduate University and E. Jane Davidson, CGU & Alliant University.A demonstration presented at the annual meeting of the American Evaluation Association, Honolulu, HI, November 2000

**Category**

1. Uncategorized

**Date**
05/11/2025
**Date Created**
24/12/2008
**Author**
admin