# Asymmetrical Tests

Michael Scriven

*Western Michigan University—The Evaluation Center*

It is often thought, or presupposed, that investigative tests done for evaluative purposes should be symmetrical—that is, they should be equally capable of giving positive and negative results. This is an error in the logic of evaluation, and it arises from confusion between the *asymmetrical* ability of a test to yield information and its propensity to yield *biased* information. Although 'asymmetrical' does connote one-sidedness and we often use 'one-sided' to mean biased, we can and should use the terms more carefully. Some tests can only yield information about the faults of a program (or product, or person, etc.)—or only about its virtues—while others can yield one of these more reliably than the other, and yet others are symmetrical in their treatment of merits and demerits. All of them gather relevant information for evaluative purposes, and one cannot conclude that any of them are biased simply because they are asymmetrical. The point is important because in many situations, one may only have access to asymmetrical tests and this does not support the claim of a biased approach *as long as* there is more than one test in the battery used to evaluate, and one test's asymmetry is balanced out by the other test(s).

An important example from personnel evaluation concerns classroom visits, especially pre-announced visits, done as part of an evaluation of teaching. If you see a very polished and knowledgeable presentation, you *cannot* conclude that this is an indication of general high quality, since it may have been specially prepared for, or stimulated by, your presence. But if the content presented (or the set of answers to several questions) is seriously defective, one *can* conclude with reasonable probability—subject to independent confirmation—that the teacher is not competent in the subject-matter. Similarly, if the classroom is chaotic, one could, with a somewhat lower probability, infer that the teacher is pedagogically incompetent. In each case, there are obvious further tests that can be made fairly easily to confirm the *prima facie* interpretation, e.g., by ruling out the possibility that anxiety due to your presence caused the error or chaos, or that some deep purpose was served by the apparent flaws.

This 'evaluative asymmetry' of a test should be distinguished from 'formal asymmetry' which is present when the response scales are different in length on the upside and the downside. For example, in evaluating teaching one may use a scale like this: Poor, Fair, Good, Very Good, Superb, where there are three or four positive and only two negative anchors. This may be desirable if prior experience shows that the ratings on a symmetrical five point scale run into headroom problems, just where one needs to spread the candidates in order to provide room for improvement, or to select someone for a teaching award. So the collection of useful information is then facilitated by spreading the topside of the scale. Similar examples can be given from program and product evaluation; the point is quite general. (Of course, if you now dropped the bottom *two* anchors, the scale *would* be biased!).

In general, then, neither evaluative asymmetry nor formal asymmetry is an intrinsic flaw in a test or instrument (within limits), and in particular, neither shows bias. Each may be thought to deserve some justification, to ward off the common concern about asymmetric instruments, and at least the first will normally require some compensation in the rest of the design.[1]

One should also note the existence of what might be called 'contextual (evaluative) asymmetry' where the test itself is not intrinsically asymmetrical but becomes so in a certain context of use. An example of this is the use of lists of publications (even if supplemented by citation indices for each of them) used as tests of research merit in the usual context of evaluating candidates for college positions, promotions, tenure, or research funding. In the common context where the review panel has: (i) no time to read the listed articles or books, or call on experts who have; and (ii) limited or zero knowledge about the quality of the journals in which the articles appear or about the publishers of (some of) the books; and (iii) no time or skill in deciding whether the citation indexes have been jiggered in one or more of the many common ways of doing this (self-reference, etc.), the list cannot provide evidence of high quality research. But if the list contains nothing at all, or just a couple of brief book reviews, in the multi-year period under consideration, then it provides excellent *prima facie* evidence of low quality research performance. (A quick check might be made for references in the documentation to a *magnum opus* under development.).

---

[1] In practical evaluation, we are often concerned with credibility as well as validity. For survey audiences brought up on the white bread diet of Likert scales, an asymmetrical test may seem biased.

In the general logic of evaluation, the asymmetrical test is analogous, although not precisely equivalent, to the use of experts who are known to have strong views about X, on panels that are to judge applicants or applications relating to X, some of which may be associated with the opposite position. Overzealous attorneys for the responsible agency sometimes try to disbar such experts, *tout court*; other attorneys may argue for their presence as evidence of bias, in a suit attacking the decisions made. But such an expert may be completely correct, since strong convictions are sometimes well justified. Indeed, such an expert may be the only true expert on the panel. The proper position is to consider whether the panel as a whole is biased (as well as knowledgeable), not whether its members are all undecided. After all, one might say, truth is evaluatively asymmetrical.