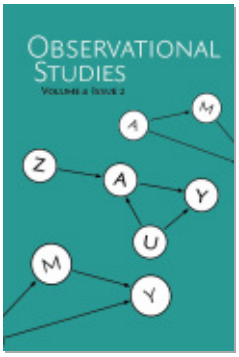
Statistical Criticism, Self-Criticism and the Scientific Method

David Rindskopf



Observational Studies, Volume 4, Issue 2, 2018, pp. 61-64 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2018.0007*

Observational Studies 4 (2018) 61-64 Submitted 12/17; Published 1/18

Statistical Criticism, Self-Criticism and the Scientific Method

David Rindskopf **drindskopf@gc.cuny.edu**
CUNY Graduate Center
New York, NY 10016, U.S.A.

I have long admired Bross's article on statistical criticism, and in my mind it has much broader implications than those Bross chose to present. In fact, others have discussed much the same points in the context of all of scientific methodology.

I also hope that Bross would say that his rules should be applied to self-criticism as well as criticism by others. There would be less need for criticism by others if there were better self-criticism in the first place. Cochran (1965) is cited by Rosenbaum:

> When summarizing the results of a study that shows an association consistent with the causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him.

Such advice is valuable, but alas mere mortals (including me) are usually deficient in self-criticism. What sometimes helps is to put a piece away for a while after writing it, and then going back specifically to criticize it before putting it on display for others to criticize.

In spite of the difficulty of self-criticism, the author is best placed to criticize, having full access to all the data of the study. As data sets more often become publicly available, this advantage will diminish. Even so, many data sets have restricted access and in those cases it will remain difficult for critics to test alternative theories.

Chamberlain (1890, reprinted 1964) and Platt (1965) have similar views as Bross, but applied more broadly to scientific methods as a whole.

Chamberlain discussed the affection a scientist feels for his or her ideas, and how unconscious bias ("tubular vision" in Bross's terms) takes over:

> As soon as this parental affection takes possession of the mind, there is a rapid passage to the adoption of the theory. There is an unconscious selection and magnifying of the phenomena that fall into harmony with the theory and support it, and an unconscious neglect of those that fail of coincidence. The mind lingers with pleasure upon the facts that fall happily into the embrace of the theory, and feels a natural coldness toward those that seem refractory. Instinctively there is a special searching- out of phenomena that support it, for the mind is led by its desires. There springs up, also, an unconscious pressing of the theory to make it fit the facts, and a pressing of the facts to make them fit the theory. When these biasing tendencies set in, the mind rapidly degenerates into the partiality of paternalism. The search for facts, the observation of phenomena and their interpretation, are all dominated by affection for the

> favored theory until it appears to its author or its advocate to have been overwhelmingly established. The theory then rapidly rises to the ruling position, and investigation, observation, and interpretation are controlled and directed by it. From an unduly favored child, it readily becomes master, and leads its author whithersoever it will. The subsequent history of that mind in respect to that theme is but the progressive dominance of a ruling idea.

Chamberlain suggests that instead researchers should develop alternative theories so that they do not maintain an allegiance to any one of them.

Platt (1964) thought that the best way to make progress in science was to eliminate alternative theories with each experiment. In that context, he said:

> In its separate elements, strong inference is just the simple and old-fashioned method of inductive inference that goes back to Francis Bacon. The steps are familiar to every college student and are practiced, off and on, by every scientist. The difference comes in their systematic application. Strong inference consists of applying the following steps to every problem in science, formally and explicitly and regularly:
>
> 1) Devising alternative hypotheses;
> 2) Devising a crucial experiment (or several of them), with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses;
> 3) Carrying out the experiment so as to get a clean result;
> 1') Recycling the procedure, making subhypotheses or sequential hypotheses to refine the possibilities that remain; and so on.

In the case of the statistician, the alternative hypotheses are alternative artifacts that might plausibly account for the findings. In parallel with Platt's reasoning for experiments, the statistician should control the most plausible alternatives first.

Platt cites one paper that contains the following line of reasoning:

> Our conclusions...might be invalid if...(i)...(ii)...or (iii)...We shall describe experiments which eliminate these alternatives.

Statisticians can use the same method, substituting "analyses" for "experiments".

Note that Platt and Chamberlain's views contradict the way science is commonly taught. Students are often taught to develop a theory or hypothesis, and then test it. They are not taught to think of alternative hypotheses until it is too late (when their experiment has failed); or in case it is not contradicted, they believe their hypothesis was confirmed, when, in fact, plausible rival hypotheses may have made the same prediction.

Donald T. Campbell often discussed plausible rival hypotheses, even devising checklists of general categories of these that one should consider in evaluating causal conclusions from a study. In the first such instance I can find, Campbell and Stanley, 1963, discussing pre-test post-test design said:

> Between O_1 [observation at time 1] and O_2 [observation at time 2] many other change-producing events may have occurred in addition to the experimenter's X [treatment]. If the pretest (O_1) and the posttest (O_2) are made on different days, then the events between may have caused the difference. To become a ***plausible*** **rival hypothesis** [italics in original], such an event should have occurred to most of the students in the group under study, say in some other class period or via a widely disseminated news story.

Note that Campbell and Stanley did not say one should defend against all rival hypotheses. They, like Bross, saw that as too broad; only plausible rival hypotheses need be checked. This is more reasonable than requiring all (an infinite number, and therefore rather difficult to execute) to be checked, but does leave some wiggle room: What is plausible?

Campbell and Stanley (1963) also stated:

> In 1923, W. A. McCall published a book entitled *How to Experiment in Education.* The present chapter aspires to achieve an up-to-date representation of the interests and considerations of that book, and for this reason will begin with an appreciation of it. In his preface McCall said: "There are excellent books and courses of instruction dealing with the statistical manipulation of experimental data, but there is little help to be found on the methods of securing adequate and proper data to which to apply statistical procedure."

Campbell and Stanley emphasized the phrase "securing adequate and proper data" in discussing this quote. This suggests a possible (sometimes probable) reason why critics could have problems making a case against a proponent: They don't have the data that would be adequate or proper to back their claims. The proponent has much more control over the data; the critic always has less control, and sometimes has little or none.

This suggests that Bross's requirement for critics may sometimes be too strict. What data are available for the critic to investigate the plausible rival hypotheses? Are they sufficient to address the necessary issues? If so, were they used correctly by the critic? If not, did the critic spell out what data would be necessary to confirm the counterhypothesis?

Rosenbaum (2002) presented an interesting quote from Fisher, who stated that criticism should not be accepted merely because it is an authority making it; an example he gives is "His controls are totally inadequate", without any elaboration of what adequate controls might be. This is interesting because Bross cites Fisher for making a similar statement in the arguments about smoking and cancer.

Bross provided a valuable service to the field by suggesting that there should be standards for criticism. We might argue about exactly how those standards should be applied, and whether they can be as strictly adhered to as he would like, but requiring critics to do more than sling random criticisms without some backing for their statements was, and still is, a reasonable standard to meet.

References

Bross, I. D. J. (1960). Statistical criticism. *Cancer*, 13, 394-400.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science* 15:9296. (reprinted in *Science* 148: 754759 [1965]).

Cochran, W. G. (1965) The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, 128, 134-155.

Platt, J. R. 1964. Strong inference. *Science* 146:347353.

Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, Springer-Verlag, New York.