Policy Studies Review, May 1986, Vol. 5, No. 4

David L. Weimer

COLLECTIVE DELUSION IN THE SOCIAL SCIENCES: PUBLISHING INCENTIVES FOR EMPIRICAL ABUSE

There is a skeleton in the social science closet. Almost all of us have heard it rattled at one time or another.[1] Nevertheless, we collectively ignore it. Some ignore it because they do not think it is very scary; others because its putting to rest would require a major restructuring of our scholarly journals. Perhaps others even fear that admitting its existence would undermine belief in the empirical progress[2] the social sciences have apparently enjoyed since the widespread diffusion of statistical training and the ready availability of the computer. I will argue that the skeleton, the bias social science journals exhibit for publishing articles reporting statistically significant results, is dangerous not so much for the obvious reason, but rather for the perverse incentives it provides to researchers. I will also suggest some changes in editorial policy that might improve the system of incentives.

Let us begin with an illustration of the problem as it is commonly perceived. Imagine 100 researchers who randomly draw data samples from a distribution of outcomes from some social process. Suppose there is a popular theory that implies the mean of the distribution should be greater than zero. Our researchers would formulate a null hypothesis that the mean is zero which they would test against the alternative hypothesis that the mean is greater than zero. They would then apply a statistical test to their respective samples such that the null hypothesis would be rejected in favor of the alternative only alpha per cent of the time when the null hypothesis is in reality true. Assume that the null hypothesis is in fact true; that is, the theory in question is false. We would nevertheless expect about alpha of our researchers to reject the null hypothesis in favor of the alternative. Now all the researchers submit their findings to journals. Perhaps out of concern that the statistically nonsignificant findings may be due to a lack of power[3] of the tests used, or perhaps because the verification of theories is intrinsically more interesting than their refutation, the editors and reviewers are more likely to react favorably to the researchers reporting rejection of the null hypothesis (most often viewed by the crude rule of thumb that t-statistics be greater than two), which we refer to as statistically significant results. The scarcity of journal space, coupled with a preference for originality, might very well result in the publication of the work of one of the researchers who found significant results, and publication of none of the others. This process leads to our collective delusion about the validity and predictive power of our theories.

How serious is the bias? Ask yourself: How often have I recommended publication of empirical papers that report no theoretically confirming and statistically significant findings? A casual look at the journals suggests that a lot of us would answer "not very often."[4] Of course, it is possible

My views on this topic have been influenced by Michael Barron, Eric Hanushek, George Benston, and Aidan Vining. However, they should not be held accountable for my assertions and conclusions.

that very few papers without statistically significant findings are ever submitted to journals. It may be that we have more powerful theories and adequate data for testing them. It may also be that researchers regularly search through their data sets until they find results that are apparently theoretically relevant and apparently statistically significant.[5] When this occurs, the negative impact of the publication bias is greatly magnified. Our theories are not just being wagged by the tail of the distribution, but by the tail of the tail.

The opportunity for searching is usually great. Often theories permit hypotheses to be specified in a variety of ways. In the multivariate context, for example, the researcher must decide which variables must be included to "hold other things equal," which functional forms are most appropriate, and for which data the theory is most relevant. By searching through the various combinations of assumptions that can reasonably be made, the researcher may be able to find a statistically significant result even when the null hypothesis is true and not rejected with the most straightforward specification. The problem with this approach is that it violates the basic assumption of classical statistical inference that the tests of hypotheses are specified before looking at the data. Once we look at the data with a particular test of hypothesis, the levels of significance we estimate in subsequent tests with standard statistical techniques are no longer valid. The more we pick through the data, the more likely it is that we will be fooled by idiosyncrasies of our particular sample of data. The increasing availability of interactive statistical packages makes the temptation for such abuse ever greater.

With the notable exception of Leamer (1978), few methodologists have attempted to develop statistical procedures that can validly be applied in succession to a data sample. Unfortunately, Leamer's approach requires us to embrace the complexities of Bayesian analysis. What can a researcher do, short of becoming an overt Bayesian, when theory is weak or the possibilities of specification are many? We were taught in our statistics classes that data samples can be randomly split so that we can search for structure in one subsample and then make a valid test of our discovered specification on the other subsample. However, this approach reduces the power of our tests of significance, and it is often impractical with time series data. Another approach, suggested by Leamer (1983), is to present the various specifications considered in a manner that allows the reader to make assessments about the validity of reported statistical tests. Unfortunately, editors are stingy about space for such discussions.

We thus face a compounded problem: the selection bias of the journals encourages researchers to ransack their data for apparently significant results that in actuality are largely invalid. How might we change our editorial policies to reduce our collective delusion? Three suggestions follow:

1. The results of statistical tests should not be submitted to journals until after articles have been accepted for publication.[6] Editors and reviewers would base their decisions on discussions of the theory under consideration, the specific hypotheses to be tested, and the data sample to be used.[7] This would divert attention from the final result to the a priori specification of hypotheses and the appropriateness of data and statistical technique for testing them. Unfortunately, this approach would not work unless adopted by all the relevant journals. If a single journal adopted it, researchers

who already knew they were not going to have statistically significant results would have an incentive to submit their papers to the journal using the result-blind review procedure. As a result, the journal would tend to have a disproportionate number of articles reporting insignificant results. Would it be able to retain its readership?

2. Any journal that publishes an article with statistical models should be required (by the disciplines?) to provide a page of space to anyone who wants to report the results of applying the authors' models to different sets of data. This approach would provide a less biased sampling of research results. It would also allow us to better gauge the robustness of our theories.[8] Some journals now provide some space for verifications; doing so should be standard editorial practice.

   An apparently similar approach would be to guarantee space to the presentation of alternative models using the same data as the published article. While this approach might help to expose obviously "cooked" findings, it would have the undesirable effect of encouraging even more ransacking. Consequently, considerable editorial discretion would be needed in implementation to make this policy on net worthwhile.

3. Researchers should be required to submit a statement with their articles indicating whether or not the model being presented is the one actually first estimated from the data. A general description of statistical techniques would also be included in the submission. Is the presented model the first specification or is it the end product of stepwise regression or some other questionable data search? Where appropriate, the editor would affix a warning along the following lines next to results:

   WARNING! THE EDITOR HAS DETERMINED THAT THE STATISTICAL TESTS REPORTED ARE BASED ON AD HOC PROCEDURES THAT VIOLATE THE ASSUMPTIONS OF CLASSICAL STATISTICS. REPORTED SIGNIFICANCE LEVELS ARE SUSPECT -- INTERPRET THEM WITH CARE!

If we assume that researchers are basically honest, the major implementation problem is that the warning would be so common that we would all soon come to ignore it.

NOTES

[1]I first heard this problem discussed by Aaron Wildavsky. The earliest published reference I could find is Sterling (1959).

[2]For general discussions of whether or not we are making empirical progress see, for example, David F. Hendry (1980) and Fischer Black (1982).

[3]Power is one minus the probability of failing to reject the null hypothesis when in fact it is false. Estimating power is difficult because it

requires assumptions to be made about the variances of the underlying distributions involved. Generally, the larger the size of the data sample, the greater the power of statistical tests.

[4]Bozarth and Roberts (1972) reviewed 1,046 empirical research articles published in journals of clinical psychology and counseling. Of the 86 percent that used tests of significance, 94 percent reported rejection of the null hypothesis. Less than one percent were replications of previously published studies.

[5]The earliest reference to this problem appears to be Walster and Cleary (1971).

[6]Feige (1975) has suggested this approach for economic journals. Although the editors of the *Journal of Political Economy* rejected his suggestion as impractical, they responded with the creation of a "Confirmations and Contradictions" section.

[7]The only journal that I am aware of that has ever followed this policy is *Representative Research in Social Psychology*, edited by psychology graduate students at the University of North Carolina. I have been unable to discover when or why the journal abandoned its "early review option."

[8]For a discussion of the desirability of greater replication, see Thomas Mayer (1980). This approach was suggested by Tullock (1959).

REFERENCES

Black, F. (1982). The trouble with econometric models. *Financial Analysts Journal, 38*(2), 29-37.

Bozarth, J.D., & Roberts, R., Jr. (1972). Signifying significant significance. *American Psychologist, 27*(8), 774-775.

Feige, E.L. (1975). The consequences of journal editorial policies and a suggestion for revision. *Journal of Political Economy, 83*(6), 1291-1295.

Hendry, D.F. (1980). Econometrics--Alchemy or science? *Economica, 47*(188), 387-406.

Leamer, E.E. (1983). Let's take the con out of econometrics. *American Economic Review, 73*(1), 31-43.

Leamer, E.E. (1978). *Specification searchers: Ad hoc inference with nonexperimental data*. New York, NY: John Wiley & Sons.

Mayer, T. (1980). Economics as hard science: Realistic goal or wishful thinking. *Economic Inquiry, 18*(2), 165-178.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance -- or vice versa. *Journal of the American Statistical Association, 54*(285), 30-34.

Tullock, G. (1959). Publication decisions and tests of significance -- A comment. *Journal of the American Statistical Association, 54*(287), 593.

Walster, G.W., & Cleary, T.A. (1971). A proposal for a new editorial policy in the social sciences. *Representative Research in Social Psychology, 2*(1), 5-10.