Book Review/Compte rendu

Stephen T. Ziliak and **Deirdre N. McCloskey**, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press, 2008, 325 pp. $US 24.95 paper (978-0-472-05007-9), $US 75.00 hardcover (978-0-472-05007-7)

In this book, the economists Ziliak and McCloskey mount a passionate plea against the widespread practice of using tests of statistical significance as the primary (or even sole) criterion for assessing the plausibility of a scientific argument. They start with two observations. The first, enshrined in the fallacy of the transposed conditional, is that it is logically not possible to get from "the probability of the data given the (null) hypothesis" to "the probability of the hypothesis given the data." The first probability is provided in tests of statistical significance while the latter probability is the proper concern of empirically based arguments. The second is a distinction between philosophical/metaphysical and scientific questions. Philosophy is concerned with existence, science with magnitude. Since statistical significance assesses the probability of the existence of an effect, but provides no information on its magnitude, it is, by Ziliak and McCloskey's definition, "almost valueless, a meaningless parlor game" for addressing scientific questions (p. 2). In short, statistical tests of significance are neither necessary nor sufficient conditions for establishing substantive significance — which they label "oomph." To develop credible arguments, they insist, at a minimum equal attention must be focussed on the magnitudes of effects (effect sizes) and the power of one's tests to detect effects of various magnitudes. None of these points should be controversial, since, as Ziliak and McCloskey note, there is no credible defense for the use of Fisher's "rule of two" — that an effect exists if it is at least twice its own sampling error distant from the (null) hypothesized value (and vice versa, that it does not exist if it does not).

The book is written in a nontechnical, almost conversational, manner that often lightly pokes fun at both the authors and others. In developing their history of how tests of significance developed into a cult, they pit as their main characters an arrogant, domineering, and ruthless professor (Ronald A. Fisher) against a modest and practical experimental brewer of Guinness stout (William S. Gosset, immortalized as the anonymous

"student" from Student's *t*). The chronicle of this battle relies on the numerous letters written between Gosset and Fisher and their respective allies. Unlike the David and Goliath saga, the institutional might and voracious ego of Fisher prevailed.

Given that no credible defense of Fisher's rule of two has been given, one could be forgiven for believing that this practice is not widespread, at least not in a discipline like economics that prides itself in its statistical prowess. To assess its prevalence, Ziliak and McCloskey developed a 19 item assessment index for proper treatment of tests of significance (an example item is: "Does the article refrain from reporting t- or F-statistics or standard errors even when a test of significance is not relevant?). They apply their assessment instrument to all quantitatively based full-length articles published during the 1980s in the *American Economics Review* — the flagship US journal for economists. The results were disheartening to them, with 70 percent of the articles making no distinction between statistical and substantive significance, for example, and less than 10 percent eschewing the reporting of sampling errors when these were irrelevant. The authors presented their results at conferences and in a journal article but often received the response that things have improved substantially since the 1980s. So for this book they repeated their exercise on articles published in the same journal in the 1990s. The results were equally disheartening, if not more so. On some crucial measures, such as mistaking statistical significance for economic significance, the situation had deteriorated.

The authors next consider whether the fetish of tests of significance is basically harmless — something that is just a ritual that perhaps makes quantitative research appear more scientific than is warranted. For the fields on which they concentrate (mostly economics, followed by psychology and medicine) their answer is contained in their subtitle: "How the standard error costs us jobs, justice, and lives." Can the same be said for sociology? Certainly our quantitative research is equally preoccupied with statistical tests of significance, despite the fact that our discipline questioned such practices several decades ago (see Morrison and Henkel's *The Significance Test Controversy*, 1970). Our reliance on Fisher's rule of two is perhaps more innocuous, as policy decisions are less likely to be based on sociological research. In my estimation, the damage is more to the discipline itself when contradictory findings are produced that are merely artefacts of our reliance on tests of statistical significance.

Let me take one example recently published in this journal. In their solid quantitative study "Family structure histories and high school completion: Evidence from a population-based registry," Strohschein, Roos,

and Brownell (*CJS* 34(1):83–103) examined whether parental bereavement had less deleterious consequences on children's educational attainment than parental divorce, as well as whether subsequent remarriage ameliorated any negative consequences. Their data consisted of the total population of all children born or adopted into two-parent households in Manitoba in 1984. They failed to find statistically significant differences in the odds of high school graduation between children whose parents divorced compared to those who lost a parent through death. They therefore conclude that their finding "contradicts" previous studies that found that divorce had greater negative effects than bereavement. Further, focussing just on those children who had experienced the loss of a parent through either death or divorce, they found that subsequent remarriage of the parent had no statistically significant effect on the odds of graduating. Out of this is born "the provocative finding that parental loss may be more important for high school completion than family instability" (Strohschein, Roos, and Brownell 2009:97). However, their calculated odds of completing high school (relative to families in which there were no marital transitions are: .39 (divorce, no remarriage), .46 (death of a parent, no remarriage), and .61 (parental death/divorce with subsequent remarriage). These differences cannot be attributed to sampling fluctuation, since their data constitutes the total population. So the findings are neither contradictory nor provocative; it is the inappropriate reliance on Fisher's rule of two that makes it appear so. Strohschein and her colleagues are clearly competent social analysts; like other quantitative researchers (including me) who wish to get published, they are victims of the tyranny of tests of statistical significance.

If this book were a novel, dramatic justice would dictate that the evil Fisher ultimately be vanquished and the cult of statistical significance be destroyed. However, the book is not a novel, and so the prognosis for victory of estimating effect sizes, the necessity of numerous replications, the careful scrutiny of nature and types of measurement error, and the importance of the power of tests is slim.

Dalhousie University Victor Thiessen

Victor Thiessen is Professor Emeritus in the Department of Sociology and Social Anthropology, Dalhousie University, and Academic Director of the Atlantic Research Data Centre. His work in the past twenty years has focussed on youth transitions, an area in which he has published extensively. His current investigations focus on the various pathways along which young Canadians navigate their way from schooling to employment. Together with Jörg Blasius (University of Bonn) he is writing a book for Sage Publications entitled *A Guide to Using and Understanding Secondary Data: Assessing the Quality of Survey Data*.

Victor.Thiessen@Dal.Ca