Erasmus Journal for Philosophy and Economics,
Volume 1, Issue 1,
Autumn 2008, pp. 165-170.
http://ejpe.org/pdf/1-1-br-2r.pdf

Science is judgment, not only calculation: a reply to Aris Spanos's review of *The cult of statistical significance.*

STEPHEN T. ZILIAK
Roosevelt University

DEIRDRE N. MCCLOSKEY
University of Illinois at Chicago

Over the past century the usual (and the conveniently mechanical) procedure devised by the great statistician, geneticist, and racial eugenicist R. A. Fisher has been shown to be scientifically silly again and again and again. Rarely has anyone actually defended NHST (null hypothesis significance testing). That is because it is logically indefensible. Statistical significance is neither necessary nor sufficient for substantive scientific significance. Everyone knows this, once they stop regressing for a minute and actually think.

We have noticed two peculiar features of the rare defenses, exhibited in Aris Spanos's (2008) review. For one thing, when mounted by people sophisticated in statistics, such as Spanos, or his allies Kevin Hoover and Mark Siegler (2008), the defenses are never defenses. They begin on the first page by admitting that NHST does *not* give mechanical assurances that its alleged findings are scientifically important. Spanos acknowledges the salience of this "long-standing problem of statistical vs. substantive significance". It is certainly "long-standing"—the error of mixing one with the other, as we show, dates to the foundation of the journal *Biometrika*, in 1901. Unhappily, though, and every time, the defenders promptly lose sight of their concession. On the second page they re-assert, as for example both Spanos and Hoover/Siegler do, that NHST offers the scientist a way of making a scientific judgment without regard to what is persuasive to other scientists.

For another thing, the defenders are always angry. Ignorant sneering, personal insult, and irrelevant indignation are judged acceptable when defending NHST. We think the anger comes from a psychological tension. The defenders realize uneasily that it is strange to depend for scientific judgment on a sampling statistic without a persuasive context—failing to ask how big is big, which is the only scientific context relevant to a real scientific test. But they have been thoroughly

indoctrinated in NHST, and belong to a professional club in which $t > 2.0$ or $p < .05$ or whatever is substituted for scientific judgment. The mechanical procedure of their profession is under attack. So they get angry. They have no reply. So they shout and bluster.

Spanos throws up a lot of technical smoke that has the effect of obscuring the plain fact that he agrees with us. (The mathematics in his piece is irrelevant to anything of importance. The reader may omit it.) His technical smoke billows. For example, he calls NHST "the Fisher-Neyman-Pearson approach". The terminology is conventional, but expresses a revealing historical error. Jerzy Neyman and Egon S. Pearson were in fact enemies of Fisher (true, *anyone* who disagreed with Fisher became instantly his enemy for life, especially if he or she was not academically powerful). The young men, Neyman and Pearson, with the encouragement of William Gosset (aka "Student"), were to be precise *criticizing* Fisher's one-criterion test of significance, from 1928 on. Although they did not then introduce the loss functions that later became routine in statistical and econometric theorizing (despite Fisher's fierce and irrational opposition), they did for example in 1933 emphasize that "how the balance should be struck" between Type I and Type II errors (false positive and false negative errors) "must be left to the investigator" (Neyman and Pearson 1933, 296).

That is a big improvement over elevating Type I error to the only criterion, $t > 2.0$, and pretending that judgment and persuasion therefore do not need to be the crucial last step in any scientific test. Statistical significance according-to-Fisher translated every quantitative question into a probability about the data assuming the truth of the singular hypothesis. It collapsed the scientific world into a Borel space, p (0, 1.0)—a procedure, by the way, that the mathematical statistician Émile Borel himself emphatically rejected. Borel (1871–1956), though a master of abstract imagination, was deeply interested in the substantive side of testing, and in Paris in the 1920s helped convert a young Jerzy Neyman to a life of substantive significance (Reid 1982, 68-70).

But of course that is the *sole* problem we are concerned with in *The cult*, the Fisherian mistake of supposing that *statistical* significance is just the same thing as *substantive*, *scientific*, *economic* significance. Spanos ends by claiming that we have ignored specification errors (which is false: we speak of them, and of twenty-something other errors of statistical and scientific experiments. But in the book we did not want to be distracted from observing the main and elementary problem of

lack of scientific substance). That specification errors, and sample-selection bias, and biases of the auspices, and the rest, are *also* problems with the usual mechanism of NHST does not (of course) somehow repair the simpler problem that we and hundreds of other critics since the 1920s have drawn attention to.

The problem is always ignored in econometrics. Arthur Goldberger gives the topic of "statistical vs. economic significance" one page of his *A course in econometrics* (1991), quoting a little article by McCloskey in 1985. Goldberger's lone page was flagged as unusual by someone in a position to know. Clive Granger reviewed four econometrics books in the March 1994 issue of the *Journal of Economic Literature* and wrote: "when the link is made [in Goldberger between economic science and the technical statistics] some important insights arise, as for example the section [well… the page] discussing 'statistical and economic significance', *a topic not mentioned in the other books*" [by R. Davidson and J. G. MacKinnon, W. H. Greene, and W. E. Griffiths, R. C. Hill, and G. G. Judge] (Granger 1994, 118, italics supplied).

Not mentioned in the other books. *That* is the standard for educating young people on the statistical/substantive distinction in econometrics and statistics at the advanced level. We wonder if Professor Spanos does better for his own students. The three stout volumes of the *Handbook of econometrics* contain a lone mention of the point, unsurprisingly by Edward Leamer (Griliches and Intriligator 1983, I, 325). In the 732 pages of the *Handbook of statistics* (Maddala, Rao, and Vinod 1993) there is one sentence (by Florens and Mouchart on p. 321). In his own impressive *Probability theory and statistical inference* (1999) Spanos himself tried to crack the Fisherian monopoly on advanced econometrics. But even Spanos looks at the world with a sizeless stare (Spanos 1999, 681-728).

The main point of Spanos's piece is that Ziliak and McCloskey do not offer guidance on how to address substantive scientific significance. Yet even if we had not, it would not be a fault. NHST is intellectually bankrupt, as Spanos agrees it is, and it should be abandoned. If you earn your living robbing banks, you should stop, right now, at once. You should not complain, "But how am I now to earn my living?" Go get honest work. And the honest work in the present case is the exercise of scientific judgment, quantified by relevant magnitudes that the best scientists find persuasive. It is quite false that Ziliak and McCloskey offer no such guidance. On the contrary, in scores of places in the book,

especially on the economic matters, we offer ideas about what constitutes an oomph-ful, scientifically relevant judgment, on, say, an experiment in paying companies to hire the unemployed. Of course, we have more intelligent suggestions about economics than about psychology or medicine. We are economists, after all. But that is the main point. *There is no discipline-independent criterion for importance, calculable from the numbers alone*. Read that again. *There is no discipline-independent criterion for importance, calculable from the numbers alone*. Scientific judgment is scientific judgment, a human matter of the community of scientists. As vital as the statistical calculations are as an input into the judgment, the judgment cannot be made entirely by the statistical machinery.

That is really what Spanos craves: a machine for making scientific judgments. He is scornful of Bayesians (on the usual illogical and Fisherian grounds that judgment cannot be exercised in scientific decisions, or on the anti-economic and Fisherian grounds that cost and benefit in persuasion are irrelevant). We are rather fond of Bayesians. If Thomas Kuhn and his numerous children and grandchildren in the history, sociology, and philosophy of science have taught us anything it is that science is a community of mutual—preferably honest and logical—persuasion. That is what Bayesians say, and it seems a sensible reminder that science must always entail judgment, not merely calculation.

In the end we are reminded of what the American philosopher William James said about the three stages of a theory's reception: "First, you know, a new theory is attacked as absurd; then it is admitted as true, but obvious and insignificant; finally it is seen to be so important that its adversaries claim that they themselves discovered it" (James 1907, 198). Spanos has examined no archives on the history of statistics, but claims (stage 1) that our theory of how NHST arose from Fisher's disputes is absurd, and that we are silly to reject NHST for model validation in econometrics. Anyway (stage 2), everyone knows that "significance" is not the same thing as scientific importance. The point, he says, is obvious and insignificant: misspecification is what matters. Yet, by-passing our large-scale empirical work on the *American Economic Review*, Spanos offers his own claim to have discovered what we discovered (stage 3): "One wonders how many applied papers published in the *American Economic Review* over the past thirty years

are likely to pass the statistical adequacy test; I hazard a guess of less than 1%" (Spanos 2008, 163).

Here is our challenge. If you think, like Spanos, that you have a valid defense of NHST, offer it. Spanos, like Hoover/Siegler, and Anthony O'Brien (2004), have tried. They have failed. But at least they are serious about their intellectual commitments, and *believe* (given their Bayesian priors) that NHST is defensible. It is not.

REFERENCES

Florens, Jean-Pierre, and Michel Mouchart. 1993. Bayesian testing and testing Bayesians. In *Handbook of statistics, Vol. 11*, eds. G. S. Maddala, et al. Amsterdam: North Holland, 303-391.

Goldberger, Arthur. 1991. *A course in econometrics*. Cambridge: Harvard University Press.

Granger, Clive W. J. 1994. A review of some recent textbooks of econometrics. *Journal of Economic Literature*, 32 (1): 115-122.

Griliches, Zvi, and Michael D. Intriligator (eds.) 1983, 1984, 1986. *Handbook of econometrics.* Vols. I, II, and III. Amsterdam: North-Holland.

Hoover, Kevin D., and Mark Siegler. 2008. Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology*, 15 (1): 1-37.

James, William. 1907. *Pragmatism: a new name for some old ways of thinking.* New York: Longmans, Green.

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. The standard error of regressions. *Journal of Economic Literature*, 34 (1): 97-114.

McCloskey, Deirdre N., and Stephen T. Ziliak. 2008. Signifying nothing: a reply to Hoover and Siegler. *Journal of Economic Methodology*, 15 (1): 57-68.

McCloskey, Deirdre N. 1985. The loss function has been mislaid: the rhetoric of significance tests. *American Economic Review*, Supplement 75 (2): 201-205.

Neyman, Jerzy, and E. S. Pearson. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference, Part I and Part II. *Biometrika*, 20A (1-2)*:* 175-240, 263-294.

Neyman, Jerzy, and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, A*, 231: 289-337.

O'Brien, Anthony P. 2004. Why is the standard error of regression so low using historical data? *Journal of Socio-Economics*, 35 (5): 565-570.

Spanos, Aris. 1999. *Probability theory and statistical inference*. Cambridge: Cambridge University Press.

Spanos, Aris. 2008. Review of S. T. Ziliak and D. N. McCloskey's *The Cult of Statistical Significance. Erasmus Journal for Philosophy and Economics*, 1 (1): 154-164.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2004a. Size matters: the standard error of regressions in the *American Economic Review*. *Journal of Socio-Economics*, 33 (5): 527-46.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2008. *The cult of statistical significance: how the standard error costs us jobs, justice, and lives.* Ann Arbor (MI): The University of Michigan Press.

Stephen T. Ziliak is professor of economics at Roosevelt University and his research fields are: welfare and poverty; economic history, rhetoric, and philosophy; and history and philosophy of science and statistics. Contact e-mail: <sziliak@roosevelt.edu>

Deirdre N. McCloskey is Distinguished professor of economics, history, English, and communication, at the University of Illinois at Chicago. Contact e-mail: <deirdre2@uic.edu>
Website: <www.deirdremccloskey.com/>