EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 298-308

Statistical Significance in the New Tom and the Old Tom: A Reply to Thomas Mayer

Deirdre N. McCloskey[1] and Stephen T. Ziliak[2]

LINK TO ABSTRACT

Tom Mayer, an old friend, wrote in 1980 a pioneering paper making the point we make in *The Cult of Statistical Significance*, namely, that testing for fit is not the same thing as testing for oomph. Tom seems now to have reverted to an erroneous, pre-1980 understanding. The direction of movement from Old Tom to New Tom is unusual. Normally speaking once a man grasps the point that Type I error in the absence of a loss function calculated at R. A. Fisher's conventional level of 5% is neither necessary nor sufficient for scientific and commercial discoveries, he does not lose his grip on it.

The economist Lawrence Officer for instance declares that grasping the insignificance of "Fisher significance" has changed his life (Officer 2010). Now instead of meaningless tests of "significance" he can test the actual hypothesis by doing what most scientists do—by looking at magnitude, by looking at the divergence of evidence from the going rhetoric of the field, that is, by doing what scientists call the test of "interocular trauma." Chemists and geologists, most biologists and historians, and almost all physicists test a hypothesis, whether major or subsidiary, by asking if the difference in magnitude between what they see and what they expect hits them between the eyes. Honest, that's what they do. Most sciences rarely use tests of statistical significance. Funny, then, that we economists are addicted to them.

1. University of Illinois at Chicago, 60607.
2. Roosevelt University, Chicago, IL 60605.

Tom Mayer (2012, abs. and 259) suggests that there has been a "debate" in economics about this old and unoriginal point, that fit is not the same thing as substantive importance. Oh, no, we don't think there has been a debate. The orthodox, who very much want to go on using their statistical training in a mechanical way, will *not* debate. When challenged they revert to a sullen silence.

From the beginning of modern statistics many of its theorists and some of its practitioners have been making 'our' point. Yet repeatedly, despite the devastating implications for the orthodox procedures (such as David Hendry's of "test, test, and test" (1980, 403)), the point has been walked past. No: strode past, with a sneer or a shudder or silence.

In economics the sole exception before New Tom to the defensive silence was a paper by Kevin Hoover and Mark Siegler in 2008, published just before our book came out (Hoover and Siegler 2008a). Kevin and Mark had, like Tom, the scientific integrity to *try* defending the routine of "testing" how important a variable is by examining the sampling variance of its fitted β. Yet Kevin and Mark, like Tom, early in their paper admit our point. True, they proceed to call it names ("jejune," for example). But they say outright that they agree with our point. Unhappily, as we noted in our reply, they don't actually understand it (McCloskey and Ziliak 2008). Sadly now we have to give the same reply to New Tom.

Tom, like Kevin and Mark, focuses on Ziliak and McCloskey. But we showed in the book that our point has been made for nearly a hundred years repeatedly even in the sciences that have come to depend on Fisherian dogma. In economics our critics radically simplify their rhetorical task by attacking the naïfs McCloskey and Ziliak, and giving the silent treatment to the identical point made by Edgeworth, Gosset (aka "Student" of Student's *t*), Egon Pearson, Neyman, Jeffreys, Borel, Wald, Yule, Deming, Savage, de Finetti, Good, Lindley, Feynman, Lehmann, DeGroot, Raiffa, Arrow, Milton Friedman, Mosteller, Tukey, Kruskal, Mandelbrot, Wallis, Roberts, Freedman, Meehl, Cohen, Rothman, Leamer, and Zellner, to name but a few.

New Tom's rhetoric throughout is of "a more balanced reading" (2012, abs.). He has come to believe that all sides must have *some* merit (though he gives almost all of the weight to the orthodox side). He is employing, quite in the character of this judicious and fair-minded scholar (true of both Old and New Tom), a Lemma of Compromise. His repeated talk of 'on the one hand, on the other' (as for example in his treatment of Aris Spanos' 2008 review of our work) suggests that he is committed to balance in scholarship. Good. But his version of balance reminds us of a bill before the Indiana Legislature long ago proposing to square the circle and set the value of pi to, for example, precisely 3.00000, *in order to make calculation easier* (and some commentators believe) to honor the rationality of the Holy Trinity (Singmaster 1985). If pi is actually an irrational number estimated

roughly at 3.14159, then for most practical purposes (such as wheel turning) it is *not* a good idea to compromise on, say, 3.07080, halfway between 3.00000 and the correct approximation, because that is the social convention and is easier. Tom remarks mildly that "it is unlikely that either side is totally wrong" (2012, 259). We honor the tone. But in science we try to arrive at mutually exclusive positions and then decide which position is better. Social grace is not the same thing as good science, and square wheels can't carry us to the market. Effect size matters all the way down.

Tom's devotion to balance leads him to quote charges against us without troubling to read the historical evidence, or to think through the charges, or to make his own judgment on the evidence and logic involved. It's like the journalist who is told by his editor to *always* get both sides of the story. There are *always* two sides (or more). Fine. It is one reason Tom edges closer to 3.00000 than to 3.14159. An example among many scattered through the paper is giving credence to "a referee's charge that when Z-M complain about the confusion of statistical with substantive significance they merely reiterate what many psychologists and statisticians have said for several decades" (259). The referee had it seems not read our book. No worries: we do not insist that the whole world read it. But Tom *is* supposed to have read it. Roughly a hundred and fifty of the book's pages, after all, cite, quote, praise, admire, amend, extend, and celebrate "what many psychologists and statisticians have said for several decades" (closer, by the way, to ten decades).[3] Tom cites Hoover and Siegler (2008a, 2008b) with approval on some fifteen occasions. He cites our crushing reply to Hoover and Siegler twice only, once to quarrel with it. Of the thirty or so positive reviews of *The Cult*, ranging from approving notices to raves, in journals from *Nature* to the *Notices of the American Mathematical Society*, Tom cites not one—not even the long and erudite retrospective of our work by Saul Hymans, Director Emeritus of the National Bureau of Economic Research, published in the *Journal of Economic Literature* (Hymans 2009). New Tom does not assemble much of the evidence for a balanced perspective.

Tom repeatedly praises Spanos (2008) for his complaint that we do not complain enough about the *other* things wrong with significance tests, such as specification error. When one of us presented a talk to the Harvard graduate students trying to get across our much more elementary point (Thomas Schelling described it in a blurb for our book as "this very elementary, very correct, very important argument"), an influential economist replied, "Yeah, sure. But *another* serious error is such-and-such," mentioning some fashionable point of textbook econometrics. We don't think such an attitude is a good one. If an engineer uses 3.00000 rather than 3.14159 in practical applications of pi, she is going to get wrong

3. Ziliak and McCloskey (2008b, 1-22, 123-237, 265-287).

results—for some purposes not *too* bad, for others, disastrous. One should get the elementary concepts of our science right before rushing off to apply still another regression technique to confounded data.

Here's "elementary." Tom (2012, 264) speaks of Milton Friedman's great book on the consumption function as finding that "at any given income level farm families have a lower marginal propensity to consume than urban families." But "lower" is not a feature of the numbers themselves. It is always and everywhere an economic and human matter of how low is low. Similarly, Tom imagines that one might hypothesize that as "the expected future price of drugs rises, current drug consumption falls. And you find that it does" (264). But "it does" is always and everywhere a question of how big is big. You have to say *how much* rise of price causes how much fall in consumption, along a scale of big and small meaningful for some human question at issue. The question is not to be answered *inside the numbers themselves*, without a scale. If you ask whether the outside temperature is hot today you are supplying implicitly some standard, some scale meaningful for a human question. Hot for golfing, hot for September, hot for interstellar gas.

Like everyone who makes the "significance" mistake, Tom shifts immediately to the *other* question—to the amount of sampling error on the estimate of a coefficient. Statistical economists after Lawrence Klein first started the routine in economics are always shifting immediately to the question of the sampling error on the estimate of a coefficient, because it gives them a mechanical test, a yes/no or on/off switch, though a test not answering the scientific or policy or other human question at issue: the question of how much. Tom says, "If someone, by adding an additional variable develops a variant that does predict well, this will be of interest to many economists, both to those who want to predict future rates, and those who wonder why the standard theory predicts badly, regardless of the oomph of that variable" (Mayer 2012, 264). But predicting "well" and predicting "badly" are themselves matters of oomph, not fit. If the additional variable said to be relevant to explaining the term structure of interest rates gave us a better explanation in the magnitude of one hundredth of a basis point, the operator of a super millisecond arbitrage engine might care, but the Federal Reserve Board would not. If someone came up with a very large sample that showed the additional variable to be nonetheless "significant" at the 5% level (setting aside power, sample biases, misspecification, and so forth), the other economists would judge her to be naïve—*if* it had occurred to them already that good fit is, after all, not the same thing as quantitative relevance to how big is big. Our point is that it has *not* on the whole occurred to most economists, who are over-trained in Fisherian econometrics and under-trained in the economic approach to uncertainty originated by William Sealy Gosset at the Guinness Brewery in 1904.

The point is made vividly in a recent interocular experiment by Emre Soyer and Robin Hogarth (2012) who tested the forecasting abilities of 257 well-published econometricians (as discussed by Ziliak 2012). When the econometricians were given conventional output such as *t*-tests and R^2s, over 70% of the predictions were wrong (again, by some human scale of mattering). When the econometricians were shown only a scatterplot of data relative to a theoretical regression line, 3% of the predictions were wrong. New Tom needs to hear about the Soyer-Hogarth experiment.

Tom attacks our thought experiment (Mayer 2012, 261, concerning Ziliak and McCloskey 2008b, 23-25, 43-44) about a choice of pills for mother's weight loss, pill Precision versus pill Oomph, writing that "it [the experiment] assumes that we already know the means for the two pills, and it thereby bypasses the need to ask the very question that significance tests address." Wait a minute. Mom and we and New Tom do know for each variety of pill the sample mean, which is after all an unbiased estimate of the population mean. We stipulated the fact, and stipulation or not a sample mean stands for what Mom needs to know. Our thought experiment presupposes large enough samples and good enough experiments to be confident in taking the sample mean to be the best available evidence on the population mean. So knowing the means is not the issue. Yet Tom, driven by his admirable desire to compromise, wants the level of significance ($p < .05$) to decide the choice of weight loss pill. He very much wants to find a middle ground between us and, say, Hoover and Siegler, and more generally the conventional practices of economists. So as usual in the literature he slides over into the *different* question—sometimes a relevant question, usually in science not the main one—of how *certain* we are of the estimated means (modulo sampling theory alone, as though sampling variation were the only source of what Gosset called "real error"; see Ziliak 2011). "Given the existence of sampling error," Tom writes, "how confident can we be about these point estimates?" (2012, 261). He writes again, "Statistical significance is needed to justify treating the coefficient … as a sufficiently reliable stand-in for the true coefficient" (261). But in our example of pill Precision versus pill Oomph, your mother—whose goal is to lose weight, not to publish in economics journals dominated by a mistaken routine for science—is given the best available knowledge about variances, too. She knows the means *and* the variances. She has to choose and, if she is wise, she will use oomph to guide her to the best choice, not the probability of a Type I error in the absence of a loss function. The loss function is weight lost, not satisfaction of a referee with a feeble grasp of statistical theory. To say it yet again: *there are two separate questions*—one question is about oomph, the magnitudes of which Mom and most scientists seek. The other question is about the error terms around such magnitudes, which Mom

has already assessed in the two pills, and is anyway irrelevant to her goal of losing weight.

And, more deeply, how would you know how "true" something is without some standard of truth *within the realm of human importance, beyond the numbers*? That is, how would you know what advice to give as to degrees of belief *without a loss function*? As the Old Tom writes (he occasionally pops up here), "What we need, but do not have, is an explicit loss function" (Mayer 2012, 273). Precisely. That is what we note in numerous examples throughout *The Cult of Statistical Significance*.

Even in gloriously second or third moments, that is, numbers do not contain their own interpretation. Our handful of critics in economics, joined now by New Tom, want technical tricks with the numbers to substitute for scientific considerations of how big is big. Tom thinks that "there is nothing wrong with using asterisks to draw attention to those hypotheses that have passed a severe test" (262), as though sampling error is the chief test that a hypothesis must pass. It's not. Consider. Your doctor has in her possession a conventional regression of health status on two variables, height and weight. Though the standard error on height is delightfully low, that on weight is embarrassingly high, not passing what Mayer (following Deborah Mayo) calls a "severe test" (Mayer 2012, 262). So when you go for a checkup your doctor says, "Your problem is not that you're too heavy, it's that you're too short."

On the basis of a fleeting reference to a paper by our former colleague Joel Horowitz (who well understands the problem, and teaches 'our' point to his students), Tom asserts that tests of statistical significance "are at home in the physical sciences" (Mayer 2012, 256). That is quite wrong. Tom must not have looked into the matter empirically. We realize that economists will be surprised to hear the news, but we say again: physicists and chemists and geologists almost never use such tests. A very minor use is restricted, as Tom (262) concedes, to informing readers about the sampling variation in a measurement of, say, the speed of light, but never as a test of the scientific hypothesis in question, such as that nothing can move faster. If you do not believe us, wander over some afternoon to, say, the physics department and spend an hour or two paging through any of the four versions of *The Physical Review*. You will not have to understand the physics (thank heaven) or the mind-boggling mathematics (double thanks) to notice that the *t*-tests that proliferate in economics journals and the *p* values in psychological journals are not in evidence. The lack of asterisks and R^2s is not because physical scientists do not face samples similar to those economists face (though the physical scientists do exhibit a clearer understanding that a sample must be a sample, not a universe; economists are frequently found insisting on the mathematics of sampling theory when they have the urn of nature spilled out before them). Economists in their modest way usually assume that people who don't use our

wonderfully sophisticated econometric methods of testing (no one, by the way, including us, is complaining about *estimation*, which is very useful compression of data) must be sadly lacking in the math to do so. But the hypothesis won't survive an hour or two with *The Physical Review*.

Tom misunderstands our distinction between philosophical existence questions and scientific magnitude questions. He cites (262) for example the mistaken assertion by Hoover and Siegler that the 1919 debate in physics about the bending of light around the sun was a matter of existence, not magnitude. On the contrary, within the limits of exactitude in the instrumentation it was a question of magnitude, as anyone knows who has actually read the literature on the matter (see, for example, Jeffreys 1919). Without an *important* bending, Einstein fails (and his exact prediction did fail for some years until a bias in the instruments was cleared up). Triviality is a matter of oomphility. It is not merely some property of the numbers themselves, independent of a scale of judgment along which to measure it.

Philosophy, theology, and mathematics care only about existence, not magnitude. But none of those admirable departments of the intellect is an empirical science. In the Department of Mathematics if *just one* even number is found that is not the sum of two primes, Goldbach's Conjecture will be discarded forever. The Conjecture remains unproven, in the mathematician's sense of the term. Yet it has been shown to be true by calculation up to very, very large numbers. For engineering or physics the numbers are large enough to treat the Conjecture as true—say, for purposes of making a computer lock. The trouble is that most economists have learned their math in the Department of Mathematics, not the Department of Engineering. So they think that in an empirical science you can test for existence separately from magnitude. They give the usual, philosophically naïve justification that "it makes little sense for scientists to try to measure the size of something that does not exist" (Mayer 2012, 262). That's not how actual science works. In actual science one wants to know how big is big, every time. There's no use for an existence of infinitesimal size ε. Not in the vast bulk of science at any rate (that is, in $[1 - \varepsilon][100]$ percent of it).

We say reluctantly that we find New Tom's small sample finding that economists are in fact *not* misled by mistaking statistical for substantive significance, … well, incredible. Surveys of the sort he and we did of *AER* papers involve judgment calls, and we suppose he reckons that if half of the profession gets it (which he claims based on his little sample claiming to re-test our AER studies), that's good enough. Our own experience—and the experience of by now hundreds of critics of null hypothesis significance testing in dozens of fields of science—is that we find the mistake in eight or nine out of every ten empirical papers in economics and other life and human sciences, from medicine to psychology.

We're glad that Tom brings up court cases. He argues that:

> As long as the racial variable has the expected sign and is significant, you have bolstered a claim of racial discrimination. By contrast, suppose you find a substantial oomph for the racial variable, but its *t* is only 1.0. Then you do not have as strong a case to take to court. Z-M might object that this merely shows that courts allow themselves to be tricked by significance tests, but don't courts have to consider some probability of error in rendering a verdict? (Mayer 2012, 263)

Well, it turns out that a court, the Supreme Court of the United States, has weighed in on the question. In "Brief of *Amici Curiae* Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents" (2010) before the U.S. Supreme Court, an eminent New York law firm drew upon our book and our articles to argue an important case of biomedical and securities law. In March 2011 the U.S. Supreme Court decided in 'our' favor—nine to zero (*Matrixx Initiatives, Inc. v. Siracusano* 2011). Statistical significance, the Court decided, is *not* necessary to prove biomedical, investor, and economic importance. It is the law of the land, and should be of econometrics.

Tom is right that the "debate" is unlikely to end unless people stop the silent treatment or the angry shouting, and start actually listening to each other. As the novelist and philosopher Iris Murdoch wisely put it, "Humility is not a peculiar habit of self-effacement, rather like having an inaudible voice, [but] it is selfless respect for reality and one of the most difficult and central of all virtues" (Murdoch 1967, 95). A question, then, in all humility has to be answered:

Is statistical significance a wise and wonderful tool with which economic science has made great strides? Your local econometrician will affirm that it is, and will press the graduate committee to insist on still more *t*-testing econometrics, instead of educating the students in quantitative methods used in progressive sciences such as engineering, physics, scientific brewing, agriculture, cosmology, geology, and history—the methods of experimentation, simulation, surveys, narrative, accounting, canny observation, interocular trauma, Bayesian decision analysis, the comparative method, natural experiments, primary documents, and the economic approach to the logic of uncertainty, lacking in contemporary econometrics.

Econometricians have been claiming proudly since World War II that significance testing is the empirical side of economics (most youngsters in economics think that "empirical"—from the Greek word for "experience"—simply *means* "collect enough data to do a significance test"). Tjalling Koopmans's influential book of 1957, *Three Essays on the State of Economic Science*, solidified the claim. But if you take the con out of econometrics (and the "trics," and the "me" too) you are not left with much (actually, just the *cri de cœur* "eo!"). What major scientific

issue since the War has been decided by tests of statistical significance? Yes, we understand: your *own* view by your *own* tests of monetarism or the minimum wage or whatever. Then why don't those misled other economists agree with you? We said "decided."

Geologists decided in the 1960s that plate tectonics was correct, after resisting it for fifty years. Historians decided in the 1970s that American slavery was profitable in a pecuniary sense, after resisting that claim for a hundred years. Mayan archaeologists decided in the 1980s that Mayan script was not mainly ideographic, after resisting for forty years. Physicists decided in the 1990s that most of matter and energy in the universe was "dark," after declaring for decades that they were close, so, so close, to a Theory of Everything, and needed only some hundreds of billions of dollars to get to it. These are examples of actual progress in science. If the rhetoric of significance made any sense, all these advances could have been based on the level of statistical significance. None actually were. And so too in economics.

The loss-functionless test of statistical significance is a tool of stagnant anti-progress in economics and a few other sciences. Let's bury it, and get on to empirical work that actually changes minds.

References

Hendry, David F. 1980. Econometrics—Alchemy or Science? *Economica* 47: 387-406.

Hoover, Kevin, and Mark Siegler. 2008a. Sound and Fury: McCloskey and Significance Testing in Economics. *Journal of Economic Methodology* 15(Mar.): 1-38.

Hoover, Kevin, and Mark Siegler. 2008b. The Rhetoric of "Signifying Nothing": A Rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology* 15(Mar.): 57-68.

Hymans, Saul. 2009. Review of *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, by Stephen T. Ziliak and Deirdre N. McCloskey. *Journal of Economic Literature* 47(2): 499-503.

Jeffreys, Harold. 1919. On the Crucial Test of Einstein's Theory of Gravitation. *Monthly Notices of the Royal Astronomical Society* 80(Dec.): 138-154.

Koopmans, Tjalling. 1957. *Three Essays on the State of Economic Science*. New York: McGraw Hill.

Mayer, Thomas. 1980. Economics as an Exact Science: Realistic Goal or Wishful Thinking? *Economic Inquiry* 18(Apr.): 165-178.

Mayer, Thomas. 2012. Ziliak and McCloskey's Criticisms of Significance Tests: An Assessment. *Econ Journal Watch* 9(3): 256-297. **Link**

McCloskey, Deirdre N., and Stephen T. Ziliak. 2008. Signifying Nothing: Reply to Hoover and Siegler. *Journal of Economic Methodology* 15(Mar.): 39-56.

McCloskey, Deirdre N., and Stephen T. Ziliak. 2010. *Brief of* Amici Curiae *Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents*, ed. Edward Labaton, Ira A. Schochet, and Christopher J. McDonald. No. 09-1156, Matrixx Initiatives, Inc., et al. v. James Siracusano and NECA-IBEW Pension Fund. November 12. Washington, D.C.: Supreme Court of the United States.

Murdoch, Iris. 2001 [1967]. *The Sovereignty of Good*. London: Routledge.

Officer, Lawrence. 2010. Personal communication, University of Illinois-Chicago.

Singmaster, David. 1985. The Legal Values of Pi. *Mathematical Intelligencer* 7: 69-72.

Soyer, Emre, and Robin Hogarth. 2012. The Illusion of Predictability: How Regression Statistics Mislead Experts. *International Journal of Forecasting* 28(3): 695-711.

Spanos, Aris. 2008. Review of Stephen T. Ziliak and Deirdre N. McCloskey's *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. *Erasmus Journal for Philosophy and Economics* 1(Autumn): 154-164. **Link**

Ziliak, Stephen T. 2011. W. S. Gosset and Some Neglected Concepts in Experimental Statistics: Guinnessometrics II. *Journal of Wine Economics* 6(2): 252-277.

Ziliak, Stephen T. 2012. Visualizing Uncertainty: Is a Picture Worth a Thousand Regressions? *Significance* (Royal Statistical Society) 9(5): forthcoming.

Ziliak, Stephen, and D. N. McCloskey. 2008a. Science Is Judgment, Not Only Calculation: A Reply to Aris Spanos's Review of *The Cult of Statistical Significance*. *Erasmus Journal of Philosophy and Economics* 1(Autumn): 165-170.

Ziliak, Stephen, and D. N. McCloskey. 2008b. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.

Cases Cited

Matrixx Initiatives, Inc., et al. v. Siracusano, et al., 563 U.S. ___ (2011).

About the Authors

Deirdre N. McCloskey is UIC Distinguished Professor of Economics, History, English, and Communication, University of Illinois at Chicago, and Professor of Economic History, University of Gothenburg, Sweden. For more information about her books, articles, essays, and interviews, visit her website and blog at **http://deirdremccloskey.org**. Her email address is deirdre2@uic.edu.

Stephen T. Ziliak is Trustee and Professor of Economics at Roosevelt University Chicago. For more information about his books, articles, essays, and interviews, visit his websites at **http://sites.roosevelt.edu/sziliak** and **http://stephentziliak.com**. His email address is sziliak@roosevelt.edu.

Go to Archive of Economics in Practice section
Go to September 2012 issue