EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 10(1)
January 2013: 87-96

Reply to Deirdre McCloskey and Stephen Ziliak on Statistical Significance

Thomas Mayer[1]

LINK TO ABSTRACT

Deirdre McCloskey and Stephen Ziliak (2012) have graciously replied to my essay titled "Ziliak and McCloskey on Statistical Significance: An Assessment" (Mayer 2012). McCloskey and Ziliak's (M-Z) criticisms can be grouped into fourteen points. I will discuss them in the sequence that M-Z do. Page references, unless otherwise indicated, are to M-Z's reply, or to my paper.

1. M-Z say that although I got it right in 1980 I have now "reverted to an erroneous, pre-1980 understanding" (298). I am not aware of such a change. On re-reading my 1980 article (which devoted only one page to significance tests), I do not see the large difference between the "old Tom" and the "new Tom" that M-Z see, though I did make two points in 1980 that I omitted in 2012. One is that statistics courses warn students about the wrong-way-round use of significance tests; the other is that for some reason significance tests are not used in maximum likelihood methods to see whether the regression that gives the best fit is really better than another regression that attributes a very different value to the strategic coefficient. But, all in all, my 1980 paper is no closer than is my 2012 paper to Ziliak and McCloskey's sweeping claims in *The Cult of Statistical Significance* (2008). Besides, even if the "new Tom" had differed from the "old Tom," changing one's mind is not an offense against the scientific canon.

2. M-Z (299) state that I focus on their 2008 book yet do not do justice to the fact that "we showed in the book that our point has been made for nearly a hundred years repeatedly." That is correct and I acknowledged it (258). When discussing a

1. University of California, Davis, Davis, CA 95616.

claim one should focus on its latest version. And my paper was already, if anything, too long without dealing explicitly with the prior researchers whose thinking M-Z absorbed in their book. Moreover, if M-Z are trying to make an argument from authority, it founders on the fact that their point has also been criticized repeatedly for an equally long time.

3. M-Z criticize my attempt at a balanced reading by attributing it to a vacuous belief that there is always something to be said on the other side. This is wrong. I do not believe that the truth always lies in the middle—see for instance my criticisms of the Fed's thinking on lags in monetary policy (Mayer 1958), of new classical theory (Mayer 1993, chapter 8), of the debate about a monetary growth-rate rule (Mayer 1998), or of naïve interventionist policies (Mayer 2009, chapter 8), or, closer to the topic at hand, my calling the wrong-way-round use of significance tests "inexcusable" (Mayer 2012, 278-279). I do believe, however, that economists who present innovative, unorthodox ideas often do claim more than they should. Here are some examples: Keynesian economics, monetarism, new classical theory, the permanent income theory, Ricardian equivalence, and monopolistic competition theory. And in physics, Newtonian theory provides an example. M-Z (305) cite with approval Iris Murdoch's "Humility is not a peculiar habit of self-effacement, rather like having an inaudible voice, [but] it is selfless respect for reality" (1967, 95). I agree. That is why I did not just repeat platitudes about so many theories having initially made excessive claims. That would have saved much effort and journal space.

4. Having accused me of being wishy-washy, M-Z also accuse me of being strongly on the side of their critics, of giving "almost all the weight to the orthodox side. … Tom does not assemble much of the evidence for a balanced perspective" (299-300). But, far from supporting the orthodox view of significance tests I wrote: "Z-M are right…one must guard against substituting statistical for substantive significance…. They are also right in criticizing the wrong-way-round use of significance tests" (278), and that in tests of maintained hypotheses the latter error "is both severe enough and occurs frequently enough to present a serious—and inexcusable—problem" (279). I then go on to say that this "error is probably much more widespread" when deciding on congruity adjustments (279). I say that "in countering the mechanical way in which significance tests are often used, and in introducing economists to the significance-test literature in other fields, Z-M have rendered valuable services" (279). This is hardly giving "all the weight to the orthodox side."

Turning to specifics, M-Z write (300, italics in original):

> 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. … An example among many…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). … Tom cites [Kevin] Hoover and [Mark] 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*…Tom cites not one. …
>
> Tom repeatedly praises [Aris] Spanos (2008) for his complaint that we do not complain enough about *other* things wrong with significance tests, such as specification error.

M-Z are right in saying that I just mention rather than discuss previous criticisms, but then I was writing an article making specific points about *The Cult*, and not a comprehensive book about it. I made this explicit when mentioning the charge made by the referee that M-Z refer to, saying that I will not discuss it; this is not the same as "giving credence." If, in discussing the Hoover-Siegler vs. Ziliak-McCloskey debate I come down more in favor of the former, this could be due to Hoover and Siegler having had the better of it on these points. It is up to M-Z to show that this is not the case, and they do not do so. That I do not cite the many favorable reviews that *The Cult* received is again a matter of space. I do say that their book "has been widely, and in many cases very favorably, reviewed by journals in diverse fields…. Few, if any, books by contemporary economists have stirred interest in so many fields" (257). I mention Spanos frequently, but nowhere "praise" him.

5. M-Z (301) reject my example of Milton Friedman's (1957) use of the differences between the income elasticity of consumption (along with differences in the average propensity to consume) of urban and of rural (or farm) families as an example of sign mattering regardless of the size of the difference. Here M-Z are right and I was wrong, badly wrong. At many points Friedman reports the extent of the differences, and I should not have cited his tests as support for the proposition that sometimes size does not matter.[2] However, the invalidity of a particular argument for a claim does not invalidate that claim, and I can defend my claim in a way that does not draw on Friedman's discussion. Suppose one claims that $y > x$ without being able to quantify the difference. Suppose further that in

2. My careless error is explained—but not excused—by the fact that, having worked extensively on the permanent income theory many years ago (see Mayer 1972), I thought I remembered Friedman's book well enough not to have to read it again. I was wrong; my apologies to the readers of my article.

a random sample of 100 cases, $y > x$ in 98 cases. A simple binomial test tells us that it is highly likely that $y > x$ also holds in the universe, that is, that we have here a significant difference. To be sure, it would be better if we could quantify the difference; a smaller sample would then suffice, or our degree of confidence could be greater. But a binomial sign test does tell us something.

In defense of citing not significance but size M-Z write that "'lower' is not a feature of the numbers themselves. It is always and everywhere an economic and a human matter of how low is low" (301). In saying this they illustrate an illuminating difference between a logical or statistical point of view and an historical one. To an historian, and Deirdre McCloskey is, of course, an eminent economic historian, it does matter whether per capita GDP in, say, 1950, was twice as large in the U.S. as in Britain or only five percent larger. That is clearly a human matter, which anyone interested in the economic structure of the two countries would want to know. But a statistician testing a model whose logic implies only that per capita GDP was higher in the U.S. does not. This difference mirrors a choice between "thick" and "thin" theories.

A similar reasoning applies to M-Z's discussion of my example about a rise in the expected future price of drugs affecting current new addiction rates. That this happens is an implication of the theory of rational behavior, and it can be used to test the theory almost without regard to the question of how much addiction rates change. I say "almost" because there is the common-sense qualification that if the change were really trivial, say, 0.001 percent, we could not trust our data to tell us the correct sign of the change, and also because a theory of rational behavior that generates only so weak an effect is not an interesting theory.

6. M-Z claim that a significance test is "not answering the scientific or policy or other human question at issue" (301). Yes, I agree. But it does partially answer the question whether the results obtained should be given scientific credence, and thus it is a question worth asking before we ask the more interesting questions about scientific, policy, or human implications. Let us imagine that a research assistant hands some computer output to Deirdre and tells her: "I am not sure, but I may well have made a big error in key-punching the data." Wouldn't it make sense for Deirdre not to look at the data until she has checked them? And isn't checking for whether—were it not for sampling error—the scientific, policy, or human true value of the coefficients could well be zero or far removed from what the output shows, similar to checking for key-punch errors? M-Z rightly castigate researchers for overemphasizing the importance of t-values, but then proceed to do so themselves.[3]

3. In their transformation into confidence intervals t-values tell us more, but that is because they then embody oomph.

7. M-Z criticize my argument (264, third paragraph) that a variable may be important regardless of its oomph. In this they are right; my argument was muddled, and I withdraw it with apologies to the readers.

8. M-Z then discuss their thought experiment of advising Mother which weight-control pill to take. I had argued that Mother should be told not just the mean weight loss for each pill, but also how confident we can be about these means. In their reply (302), they add two elements to this thought experiment. One is that Mother was given not only the means, but also the variances. The second is that it presupposes large enough samples. Given these two new elements and interpreting "large enough" as "very large" (see Mayer 2012, 272), my criticism is, indeed, off-base. But saying that you do not need significance tests if you know the variances and have a large enough sample is a much more modest claim than the one they make in their book. And it substantially changes their thought experiment. In any case, a major use of significance tests in academic economics is to see whether a given result should be accepted into the inventory of journal-certified research, and not to make decisions. M-Z do not think much of such a distinction (and I have some sympathy with them), but that is another issue.

9. M-Z object to my saying that attaching asterisks to those coefficients that have passed a severe test is permissible. They object because a significance test is not the most important test of a hypothesis. And in this they are right; many other problems, such as collinearity or conceptually inappropriate data, may result in errors that could be much more serious than sampling errors. But the danger resulting from the latter are much easier to flag than those other errors, so *perhaps* highlighting them by asterisks is appropriate. I say "perhaps" because M-Z are right to worry that such treatment will result in *some* economists overemphasizing their importance.

10. M-Z write: "On the basis of a fleeting reference to a paper by…Joel Horowitz, 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" (303). M-Z are right that, rather than braving the risk of misunderstanding articles in physics journals, I relied on the statement of a trained physicist.[4] And I do not see why it should take more than "a fleeting reference" since Joel Horowitz was unequivocal. Turning to the lesser breeds of the natural sciences, Douglas Johnson (1999, 763) reports that "Statistical testing of hypotheses in the wildlife field has increased dramatically in recent years", while Tyler VanderWeele (2010) writes about "The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research." The statisticians Peter Guttorp and

4. As M-Z call Horowitz someone "who well understands the problem, and teaches 'our' point to his students" (303), he makes a particularly effective witness for me.

Olle Häggström (2011, 1) inform us that: "Throughout most or all of the empirical sciences, significance testing is one of the most widely used statistical procedures."

And even if significance tests are not used in physics, such tests might still be useful in economics. Economists can ill afford the snobbery of saying: "what isn't good enough for physicists is not good enough for us." Where would that leave rational-agent models? Since physics has much better opportunities than economics to isolate the effects of critical variables from their background noise, good economics need not look like physics in every way.

11. M-Z say (304) that I do not understand the difference between mathematics and philosophy—subjects that are concerned with whether an effect exists—and the sciences—subjects that deal with how big the effect is. This is not an adroit division. In the late seventeenth century scientists tried to explain fire by the presence in the air of a substance called phlogiston. Science then progressed, not by measuring the size of phlogiston or its effects better and better, but by showing that it does not exist. By contrast, someone who asks whether patriotism is a nobler virtue than is a cosmopolitan outlook is asking a question about size, but she is a philosopher, not a scientist.

In my paper I cited several examples of scientists asking about existence. M-Z question only one, the bending of light rays predicted by relativity theory, and leave the others standing. Rather than argue about light rays, I will withdraw this example and substitute another one. Suppose it were shown that teleporting can occur. Scientists would be greatly interested, and their interest would be about the same regardless of whether it was one gram that was teleported one millimeter, or 100 kilograms teleported one kilometer.

12. We come to the major issue of how frequently economists misuse significance tests by interpreting t's as though they measure oomph, or by using significance tests the wrong way round. What is radical about Ziliak and McCloskey's book is not their recognition of the distinction between significance levels and oomph, or the distinction between right-way-round and wrong-way-round significance tests—something probably most economists know, in principle. What is radical about their book is their claim that most economists confuse these in their day-to-day work. This should therefore be the main battleground. Ziliak and McCloskey based their claim on their analysis of all relevant *American Economic Review* articles in the 1980s and 1990s. They reject in just a single paragraph the quite different results from my own analysis of *AER* articles. Here is this paragraph:

> 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. (304, italics in original)

M-Z here make one claim that has some, but limited justification, and another that is completely unjustified. The former is that, for the mundane reason that evaluating *AER* papers is time consuming, and that I had no funding to hire a research assistant, I used a relatively small sample—but a sample of fifty is not all that small. Moreover, M-Z ignore that in addition to my own *AER* sample I also cited Anthony O'Brien's (2004) sample of 118 papers from economic history journals, and also Walter Krämer (2011), who had analyzed all relevant papers in the *German Economic Review* since 2000. O'Brien's results were more or less similar to mine, while Krämer faulted 56 percent of his papers, well below the 80 or 90 percent claimed by M-Z. Sample size is therefore not a credible explanation for the disparities in results. What is a more likely explanation is that Ziliak and McCloskey (2008) applied more severe criteria than the rest of us did. To help the reader decide whose criteria are more reasonable I had provided two appendices: Appendix A (Mayer 2012, 280-284) discusses Z-M's criteria, while Appendix B (285-289) provides relevant passages from the eleven papers that are in both Z-M's and my sample, and that Z-M classified as "poor" or "very poor." Beyond that, since Z-M and I both listed the individual papers we evaluated, interested readers can also compare Z-M's and my evaluations of a sample of other papers.

The completely unjustified claim is to attribute to me the belief that half the profession gets it right and that this is good enough. The claim is unjustified for two reasons. First, in my replication more than half got it right (I refrain from giving a single number, since "right" is a matter of degree and of interpretation). Second, I did not claim that that is good enough, writing:

> [W]e should…reject the idea that, at least in economics, all is well with significance tests. … Z-M are right in saying that one must guard against substituting statistical for substantive significance and that economists should pay more attention to substantive significance. They are also right in criticizing the wrong-way-round use of significance tests. In the testing of maintained hypotheses this error is both severe enough and occurs frequently enough to present a serious—and inexcusable—

> problem. And the error is probably much more widespread when [making congruity adjustments]. (Mayer 2012, 278-279)

13. M-Z (305) point to a recent Supreme Court case in which the Court unanimously held that statistical significance is not necessary to prove discrimination, thus answering in the negative my question: "don't courts have to consider some probability of [sampling] error in rendering a verdict?" (Mayer 2012, 263). That the Court rejected formal significance testing is hardly surprising, given the importance in jurisprudence of justice to a particular individual, rather than in deciding what deserves to be treated as a scientific finding, and also the difficulty of knowing what p-value to use. That does not mean, however, that the Court does not consider probability in making its decisions. Suppose a firm is sued because, despite its employing the same number of men and women at the entry level, it has recently promoted to the next level three men, but only one woman. The employer argues that with such a small sample idiosyncratic factors could easily account for it having promoted one more man and one fewer woman than strict equality would call for. Wouldn't the Court be more sympathetic to that firm than to a much larger firm that makes a similar argument for having promoted thirty men but only ten women? If so, the Court is informally taking the probability of sampling error into account, which is all that I had suggested.

14. M-Z ask whether significance testing is a "wise and wonderful tool with which economic science has made great strides" and "What major scientific issue since the War has been decided by tests of statistical significance?" (305-306). Significance testing deals with just one of many problems that confound econometric work (see Mayer 2007), and it is often overemphasized. It may not be possible to point to any breakthroughs it has brought about since the War, but it probably kept out of the literature a number of seeming 'findings' that were based only on the vagaries of sampling and therefore deserved to be kept out. There is value in asking how far we fall short of wonderful. But there is also value in pondering how far we may fall if some of our less-than-wonderful practices are faulted too severely or discarded too readily. Our analysis should be framed as comparison of institutions, whether our topic is the economy or the empirical investigation of the economy.

References

Friedman, Milton. 1957. *A Theory of the Consumption Function*. New York: Columbia University Press.

Guttorp, Peter, and Olle Häggström. 2011. Comment on "Significance Tests in Climate Science" [by M. H. P. Ambaum]. Working paper. **Link**

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

Hoover, Kevin, and Mark Siegler. 2008b. The Rhetoric of "Signifying Nothing": A Rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology* 15(1): 57-68.

Johnson, Douglas. 1999. The Insignificance of Statistical Significance Testing. *Journal of Wildlife Management* 63(3): 763-772.

Krämer, Walter. 2011. The Cult of Statistical Significance: What Economists Should and Should Not Do to Make Their Data Talk. *RatSWD Working Papers* 176. German Data Forum (Berlin). **Link**

Mayer, Thomas. 1958. The Inflexibility of Monetary Policy. *Review of Economics and Statistics* 40(4): 358-374.

Mayer, Thomas. 1972. *Permanent Income, Wealth, and Consumption*. Berkeley: University of California Press.

Mayer, Thomas. 1980. Economics as a Hard Science: Realistic Goal or Wishful Thinking? *Economic Inquiry* 18(2): 165-178.

Mayer, Thomas. 1993. *Truth Versus Precision in Economics*. Aldershot, UK: Edward Elgar.

Mayer, Thomas. 1998. Monetarists and Keynesians on Central Banking: A Case Study of a Flawed Debate. In *Economics and Methodology: Crossing Boundaries*, eds. Roger E. Backhouse, Daniel M. Hausman, Uskali Mäki, and Andrea Salanti, 254-302. New York: St. Martin's Press.

Mayer, Thomas. 2007. The Empirical Significance of Econometric Models. In *Measurement in Economics: A Handbook*, ed. Marcel Boumans, 321-340. Amsterdam: Elsevier.

Mayer, Thomas. 2009. *Invitation to Economics: Understanding Argument and Policy*. New York: Wiley-Blackwell.

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. 2012. Statistical Significance in the New Tom and the Old Tom: A Reply to Thomas Mayer. *Econ Journal Watch* 9(3): 298-308. **Link**

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

O'Brien, Anthony. 2004. Why Is the Standard Error of Regressions So Low Using Historical Data? *Journal of Socio-Economics* 33(5): 565-570.

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(1): 154-164. **Link**

VanderWeele, Tyler. 2010. The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research. *European Journal of Epidemology* 25: 843-845.

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: University of Michigan Press.

About the Author

Thomas Mayer is emeritus professor of economics, University of California, Davis. His main fields are the methodology of economics and monetary policy. His latest book is *Invitation to Economics*. His email address is tommayer@lmi.net.

Stephen Ziliak and Deirdre McCloskey's reply to this article
Go to Archive of Economics in Practice section
Go to January 2013 issue