EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 10(1)
January 2013: 97-107

We Agree That Statistical Significance Proves Essentially Nothing: A Rejoinder to Thomas Mayer

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

LINK TO ABSTRACT

We are happy to reply a second time to Tom Mayer, and are flattered that he has devoted so much time to our book. Long before most readers here were born, Professor Mayer was already making major contributions to economics—as recognized by a festschrift in his honor (Hoover and Sheffrin, eds., 1996). Mayer (2012, 2013) strongly agrees with us on four of the five major claims we make in our book about the theory and practice of significance. We are not surprised. As we noted in Ziliak and McCloskey (2004a, b), in reply to other critics from Graham Elliott and Clive Granger (2004) to Jeffrey Wooldridge (2004), Edward Leamer (2004), and Arnold Zellner (2004), leading econometricians find on reflection that they agree with our points, and realize with us that most of modern econometrics has to be redone, focusing on *economic* significance and not on mere *statistical* significance. So we have been saying since 1985 (McCloskey 1985).

Mayer limits his complaints to certain aspects of theoretical and empirical claims we make about significance testing *in economics*, in the 1990s *American Economic Review*. He is ignoring, for example, Ziliak's archival work, especially on William Sealy Gosset and Ronald A. Fisher, and our chapters on the 20th century history, philosophy, sociology, and practice of significance testing in the life and human sciences, from agronomy to zoology and from Gosset to Joshua Angrist (Ziliak and McCloskey 2008, chs. 1, 6-7, 16-24). We wish Mayer took a wider view.

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

Our complaints about statistical significance have been routine in other fields for a century, and are not "sweeping claims" (Mayer 2013, 87). It is quite mistaken, as Mayer asserts without citation, "that [our] point has also been criticized repeatedly for an equally long time" (88). No, it has not been.

No one has refuted our central theses, not ever. In several dozen journal reviews and in comments we have received—from, for example, four Nobel laureates, the statistician Dennis Lindley (2012), the mathematician Olle Häggström (2010), the sociologist Steve Fuller (2008), and the historian Theodore Porter (2008)—no one, Tom Mayer included, has tried to actually *defend* null hypothesis significance testing. True, people sometimes express anger. Given the large implications for science and policy, we understand. But no one has offered real arguments. We praised Kevin Hoover and Mark Siegler (2008) for attempting to use significance tests in step-wise fashion, that is, as a mechanical screening device for sorting possibly important from unimportant relationships (McCloskey and Ziliak 2008, 46-47). But they fumbled the ball, and for the same reason that Mayer (2013, 94) does and must: statistical significance does not and cannot do the job of interpretation and decision-making.

But the main point here is that Mayer mostly agrees with us. We are in fact in complete agreement about four of the five major claims we make in *The Cult of Statistical Significance* (Ziliak and McCloskey 2008, *passim*):

1.) **Economic significance is not the same thing as statistical significance: Each can exist without the other—without, that is, either the presence or the prospect of the other—though economists often confuse and conflate the two kinds of significance.**

Mayer agrees. As he admits in his second assessment of our work by quoting his first assessment:

> 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" ([Mayer 2012,] 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). (Mayer 2013, 88)

2.) **Economists allocate far more time, space, and legitimacy to calculating *statistical* significance than to exploring what they should: *economic* significance.**

Mayer agrees: "[I]n 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" (Mayer 2012, 279, quoted in Mayer 2013, 88). A few pages on, he goes further, fixing on the main point of our agreement: "M-Z claim that a significance test is 'not answering the scientific or policy or other human question at issue'…. Yes, I agree" (Mayer 2013, 90).

3.) **The probability of the hypothesis given the available evidence is not equal to the probability of the evidence assuming the null hypothesis is true. Thus the null hypothesis test procedure is illogical in its primitives, equations, and conclusions. The illogic of the test is an example of what is known as the fallacy of the transposed conditional.**

Mayer again agrees, in this case with mere logic and common sense. As he said both in his first assessment of our work and in the current reply: "They [Z-M] are also right in criticizing the wrong-way-round use of significance tests" (Mayer 2012, 278, quoted in Mayer 2013, 88 and 93; see also Mayer 2012, 271). The null test of significance is illogical because it measures the probability of the evidence assuming a true null hypothesis but then pretends to answer a much more important question: the probability of the hypothesis, given the new evidence. Thus economists do not know the probability of their hypotheses.

4.) **Assessing probabilities of hypotheses (Bayesian, classical, and other), comparing expected loss functions, and demonstrating substantive significance ought to be the central activities in applied econometrics and in most other statistical sciences.**

Here again, Mayer agrees. "M-Z criticize my argument ([Mayer 2012,] 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" (2013, 91). In other words, Mayer agrees that statistical significance—though conventional and even required by modern institutions of science, politics, and finance—is a dangerously broken instrument, one that in most cases needs to be put back on the shelf. Statistical significance proves nothing. Let's get back to science. End of story, off to the pub.

With such overlap in our positions the impartial spectator will begin to suspect that the remaining differences could be not very important ones. But Mayer hesitates, perhaps out of an admirable loyalty to other economic scientists, such as Kevin Hoover and Aris Spanos. He persists in trying to defend another side of the story, as though there was one, kicking at a horse that has long been dead, as Thomas Schelling (2004) has noted. Then he falls back into the main and highly ubiquitous confusion, which our work has sought to repair. Mayer (2013, 89-90)

offers a hypothetical test of $y > x$ in 98 cases out of 100 as evidence of a "significant difference." Wait a minute. Suppose y was 4.2 and x was 4.1 *and there was no scientific or policy reason to care about a difference of 0.1*. "A binomial sign test does tell us something," he writes (90). Yes, but it does not tell us that a difference is important, which is overwhelmingly the scientific point at issue, a point which Mayer elsewhere understands and accepts.

5.) **Applying our 19-item questionnaire to all of the 187 full-length empirical articles published in the 1990s in the *American Economic Review*, we find that about 80% of economists made the significance mistake, equating statistical significance with economic importance, and lack of statistical significance with unimportance.**

Mayer (2012, 279) agrees that the error is "inexcusable" but he disagrees with our estimate (2013, abs. and 93). Notice that of the five main claims we make in *The Cult*, the only real difference of opinion between Mayer and us is not a difference in kind but of degree. This makes our point rather well that quantitative oomph, not a qualitative statement that $p < .05$, is the main point of science.

Mayer did not conduct a replication of our survey of significance testing in *American Economic Review* articles from the 1990s. Neither did he attempt to replicate our survey of 182 articles published in the same journal during the 1980s (Ziliak and McCloskey 2008, chs. 6-7; Ziliak and McCloskey 2004a, b; McCloskey and Ziliak 1996). He did not use our survey at all. And yet he expresses discontent with the estimate that emerges from our application of our survey to the pages of the *AER*.

We find, to repeat, that 80% or more of our colleagues who published in the *AER* in the 1990s made the significance mistake—the mistake of not understanding claim number (1) above—causing incorrect decisions about models, variables, and actual economies. Mayer says he disagrees. We do not accept his alleged evidence, but even if we did we do not believe that a failure rate of 20% (his lower bound) or of 50% (his upper bound), though lower than the 80% rate we found, would be cause for jubilation. A 50% failure rate in significance testing would still be akin to choosing hypotheses, estimates, and economic policies on the basis of random coin-flipping.

Mayer (2013, abs.) wonders in his reply to us why we did not more fully discuss his "estimate." One reason is that Mayer's 2012 "Assessment" of our work is not based on a valid empirical result. Mayer (2012, 266) looked at only 35 of the 369 articles that we studied from *AER* issues between January 1980 and December 1999. We use the phrase "looked at" precisely. He did not use our survey instrument—the 19-item questionnaire—and therefore he is not able to determine whether he and we come to the same empirical conclusions about our colleagues' misuse of significance testing. Our survey questions were culled from best-practice

statistics and econometrics, descending from Francis Edgeworth and Gosset on down to Zellner, Leamer, and the statistician James O. Berger. Yet instead of using quantitative criteria of judgment suggested by best-practice statistics, as we did, Mayer (2012, 267) instead made a vague overall assessment with regard to "the correct takeaway point with respect to significance and oomph" on the 35 articles he selected (we hope randomly) from our population of 369 articles. In an Appendix, Mayer (2012, 285-289) took a closer look at 11 of the 369 articles. And again, rather than using our 19-question instrument for actual replication, he instead assigned overall judgments of "good", "fair", and "poor". Still, in Mayer's most detailed analysis (such as it is), he finds that about 50% of the economists fail when it comes to oomph and probability.

Recently a few other studies, which do use quantitative criteria, have come to our attention. At the University of Leuven, Sophie Soete completed a master's thesis, "The Use and Abuse of Statistical Significance." Soete (2012, 26) finds that:

> Mayer did not use Ziliak and McCloskey's questionnaire to assess the use of statistical and economic significance in the *AER* papers, but examined each paper based on the following criterion: Would a harried reader who is not watching for the particulars of significance testing obtain the correct take-away-point with respect to statistical and economic significance? Mayer thus focuses on the main thesis of a paper, and classifies a paper as discussing economic significance if it mentions the magnitude of the key variables but does not discuss effect sizes of other variables. …
>
> Mayer's results contrast sharply with those of Ziliak and McCloskey—this is not surprising however, as the studies use a very different methodology.

Soete continues (28):

> The (mis)use of statistical significance has also been examined for papers published in the *German Economic Review*. In an unpublished paper, Krämer (2011) examines all the articles that have been published in the *German Economic Review* between 2000 and 2011. The results seem similar to the results of the studies of Ziliak and McCloskey, but appear to be slightly less dramatic. It is impossible to compare both studies in detail, however, as Krämer does not provide exact information about his methodology. Of the 110 papers that report using significance tests, 56% confuse economic and statistical significance.

Thus our critics Mayer and Walter Krämer[3] join Hoover, Siegler, Spanos (2008), and others, in expressing strong opinions about our survey instrument and our empirical findings without actually testing them. No replication, no results.

In 2012 at the University of Brasília, Carlos Cinelli completed a dissertation on "The Use and Abuse of Tests of Significance in Brazil." Cinelli did not attempt to replicate our results with the *American Economic Review*. But he did use our survey instrument to analyze 94 articles in the *Revista Brasileira de Economia*:

> The empirical chapter discusses the literature about the subject specifically in economics. We show the evidence found in other countries like the United States—McCloskey and Ziliak (1996), Ziliak and McCloskey (2004a, [2008])—and Germany—Krämer (2011): 70 and 79% of the papers published in the *American Economic Review*, in the 80's and the 90's, respectively, and between 56 to 85% of the papers published in the *German Economic Review* conflate statistical and economic significance. We, then, quantify the problem in Brazil, taking a sample of all 94 papers published in *Revista Brasileira de Economia*, between 2008 and 2011, and carefully analyzing all 67 that used significance tests. Among other numbers, the main results are: 64% of them confused statistical significance with economic significance; more than 80% ignored the power of the tests; 97% did not discuss the significance level; 74% showed no concern about specification or statistical adequacy; 40% did not present descriptive statistics; more than half did not discuss the size of the coefficients; also more than half did not discuss the scientific conversation within which a coefficient would be judged large or small. (Cinelli 2012, abs.)

Cinelli credits several of our critics for offering comments and advice to him: "Agradeço a…Aris Spanos, Deborah Mayo e Walter Krämer pelas informações prestadas e dúvidas esclarecidas" (2012, acknowledgments).

In 2006 at the University of Pretoria, Walter H. Moldenhauer completed a master's thesis, "Empirical Analysis in South African Agricultural Economics and the R-Square Disease." Moldenhauer applied our 19-item questionnaire to research articles published in *Agrekon*, a top journal in Africa of applied agricultural economics. The size matters/how much question is more often addressed in subfield journals, we've found, as against general-interest journals such as the *AER* (Ziliak and McCloskey 2008, 45). Yet even in the leading South African journal of

3. Krämer, who is cited favorably by Mayer, openly admits in his paper that its purpose "is to put the Ziliak-McCloskey view into perspective" (Krämer 2011, 3).

agricultural economics Moldenhauer finds that "Almost 33 percent of the papers in stratum three remarked on the sign not the size of the coefficients. The figure is more or less in line with the 48 percent reported by Ziliak and McCloskey (1996) for their survey of papers published during the 1980s" (Moldenhauer 2006, 113).

And to return briefly to our objection that Mayer skips over the converging evidence from other fields on the size and robustness of our estimate, that statistical significance is being widely confused with substantive significance, and vice versa, it's important to understand that the confusion is just as widespread in applied fields of science far removed from economics. Consider for example a study of significance testing in the field of conservation biology, one that cites the Ziliak and McCloskey (2004a) questionnaire:

> In 2000 and 2001, 92% of sampled articles in *Conservation Biology* and *Biological Conservation* reported results of null-hypothesis tests. In 2005 this figure dropped to 78%. There were corresponding increases in the use of confidence intervals, information theoretic, and Bayesian techniques. Of those articles reporting null-hypothesis testing—which still easily constitute the majority—very few report statistical power (8%) and many misinterpret statistical nonsignificance as evidence for no effect (63%). Overall, results of our survey show some improvements in statistical practice, but further efforts are clearly required to move the discipline toward improved practices. (Fidler et al. 2006, 1539)

Likewise in the leading journals of criminology:

> We find that most researchers provide the basic information necessary to understand effect sizes and analytical significance in tables which include descriptive statistics and some standardized measure of size (e.g., betas, odds ratios). On the other hand, few of the articles mention statistical power and even fewer discuss the standards by which a finding would be considered large or small. Moreover, less than half of the articles distinguish between analytical significance and statistical significance, and most articles used the term "significance" in ambiguous ways. …
>
> To address these four issues, we adapt the instrument used in economics by McCloskey and Ziliak (1996) and Ziliak and McCloskey (2004). We used this instrument to code 82 articles in criminology and criminal justice selected from three sources: *Criminology*, the flagship journal of the American Society of Criminology, *Justice Quarterly*, the flagship journal of the Academy of Criminal Justice Science, and a review piece by Farrington and Welsh (2005) on experiments in criminal justice.

> In each case our goal is to focus on outlets representing the best practice in a particular area of the field.
>
> We find very similar results across the outlets. … On the other hand, only 31% of the articles mention power and fewer than 10% of the articles discuss the standards by which a finding would be considered large or small. None of the articles explicitly test statistical power with a specific alternative hypothesis. It is not surprising, therefore, that only 40% of the articles distinguish between analytical significance and statistical significance, and only about 30% of the articles avoid using the term "significance" in ambiguous ways. In large part, research in this field equates statistical significance with substantive significance. Researchers need to take the next step and start to compare effect sizes across studies rather than simply conclude that they have similar effects solely on the basis of a statistically significant finding in the same direction as previous work. (Bushway, Sweeten, and Wilson 2006, 1, 4)

Whether the malfeasance rate is closer to Mayer's impressions or to our detailed test yielding an 80% figure, we are all glad to know that our shared scruples are finding widespread support and application. A Columbia University statistician, Andrew Gelman, concludes in his popular blog that "A sieve-like approach seems more reasonable to me, where more complex models are considered as the sample size increases. But then, as McCloskey and Ziliak point out, you'll have to resort to substantive considerations to decide whether various terms are important enough to include in the model. Statistical significance or other purely data-based approaches won't do the trick" (Gelman 2007). And to repeat, as Mayer himself admits: "M-Z criticize my argument ([Mayer 2012,] 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" (2013, 91).

In other words, there is little in Mayer's articles to change beliefs about the theory or fact of oomphless economics. Suppose instead of economists and statisticians we were talking about dentists and toothpaste. Agreement on four out of five virtues of toothpaste would be broadcast on national television; the scientific consensus would be used as a marketing slogan, and four out of five dentists surveyed would agree. Likewise, consensus sometimes occurs in other branches of science, economics included. (Perhaps in this way economists can make good on Keynes's dream of economists acting less like prophets and more like dentists.)[4] We are encouraged by the new and growing consensus on statistical

4. See the final sentences of *Essays in Persuasion* (Keynes 1931).

significance—that as a formula for science and decisions, it is weak from roots to crown.

References

Bushway, Shawn D., Gary Sweeten, and David B. Wilson. 2006. Size Matters: Standard Errors in the Application of Null Hypothesis Significance Testing in Criminology and Criminal Justice. *Journal of Experimental Criminology* 2: 1-22.

Cinelli, Carlos L. K. 2012. *Inferência estatística e a prática econômica no Brasil: os (ab)usos dos testes de significância*. Master's diss., Programa de Pós-Graduação em Economia, Universidade de Brasília.

Elliott, Graham, and Clive W. J. Granger. 2004. Evaluating Significance: Comments on "Size Matters". *Journal of Socio-Economics* 33(5): 547-550.

Fidler, Fiona, Mark A. Burgman, Geoff Cumming, Robert Buttrose, and Neil Thomason. 2006. Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. *Conservation Biology* 20(5): 1539-1544.

Fuller, Steve. 2008. 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. *Times Higher Education*, April 3. **Link**

Gelman, Andrew. 2007. Significance Testing in Economics: McCloskey, Ziliak, Hoover, and Siegler. *Statistical Modeling, Causal Inference, and Social Science*, October 5. **Link**

Häggström, Olle. 2010. 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. *Notices of the American Mathematical Society* 57(9): 1129-1130. **Link**

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

Hoover, Kevin, and Steven Sheffrin, eds. 1996. *Monetarism and the Methodology of Economics: Essays in Honour of Thomas Mayer*. Cheltenham, UK: Edward Elgar.

Keynes, John Maynard. 1931. Economic Possibilities for Our Grandchildren. In *Essays in Persuasion*, 358-373. London: Macmillan.

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**

Leamer, Edward E. 2004. Are the Roads Red? Comments on "Size Matters". *Journal of Socio-Economics* 33(5): 555-557.

Lindley, Dennis V. 2012. Correspondence to Stephen T. Ziliak, November 14.

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

Mayer, Thomas. 2013. Reply to Deirdre McCloskey and Stephen Ziliak on Statistical Significance. *Econ Journal Watch* 10(1): 87-96. **Link**

McCloskey, Deirdre N. 1985. The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests. *American Economic Review* 75(2): 201-205.

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: Reply to Hoover and Siegler. *Journal of Economic Methodology* 15(1): 39-55.

Moldenhauer, Walter H. 2006. *Empirical Analysis in South African Agricultural Economics and the R-Square Disease*. Master's diss., Faculty of Economic and Management Sciences, University of Pretoria.

Porter, Theodore M. 2008. Signifying Little [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]. Science 320(5881): 1292.

Schelling, Thomas C. 2004. Correspondence [on "Size Matters: The Standard Error of Regressions in the *American Economic Review*", by Stephen T. Ziliak and Deirdre N. McCloskey]. *Econ Journal Watch* 1(3): 539-540. **Link**

Soete, Sophie. 2012. *The Use and Abuse of Statistical Significance: The Case of "The Spirit Level."* M.A. thesis, Faculteit Economie en Bedrijfswetenschappen, Katholieke Universiteit Leuven.

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**

Wooldridge, Jeffrey M. 2004. Statistical Significance is Okay, Too: Comment on "Size Matters". *Journal of Socio-Economics* 33(5): 577-579.

Zellner, Arnold. 2004. To Test or Not to Test and If So, How? Comments on "Size Matters". *Journal of Socio-Economics* 33(5): 581-586.

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): 527-546. (This article was also published, with permission of the journal just cited, in *Econ Journal Watch* 1(2): 331-358 (**link**).)

Ziliak, Stephen T., and Deirdre N. McCloskey. 2004b. Significance Redux. *Journal of Socio-Economics* 33(5): 665-675.

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 Authors

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.

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.

Go to Archive of Economics in Practice section
Go to January 2013 issue