EJW
ECON JOURNAL WATCH
Scholarly Comments on
Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 256-297

Ziliak and McCloskey's Criticisms of Significance Tests: An Assessment

Thomas Mayer[1]

LINK TO ABSTRACT

If economists have natural constants, then the most well-known is 0.05.
—Hugo Keuzenkamp and Jan Magnus (1995, 16)

Significance tests are standard operating procedures in empirical economics and other behavioral sciences. They are also widely used in medical research, where the prevalence of small samples makes them particularly welcome. And—mainly in the form of error bars—they are also at home in the physical sciences (see Horowitz 2004). But they have many critics, particularly among psychologists, who have done much more work on the topic than have economists.[2] Some members of the American Psychological Association even tried to ban their use in all journals published by the Association. That proposal was easily defeated, but some of the editors of those journals moved on their own to discourage significance tests, although with little effect; significance tests still reign in psychology. In medical research, however, the critics appear to have had considerable influence.[3]

[1] University of California, Davis, Davis, CA 95616. I am indebted for excellent comments to Kevin Hoover and Deirdre McCloskey and to three outstanding referees for this journal. An earlier version was presented at the 2010 meeting of the Society of Government Economists.

[2] For telling quotations from critics see Johnson (1999).

[3] Fidler et al. (2004b, 626) explain the spread of the reform in part by a shift from testing to estimation that was facilitated by the medical literature, unlike psychology, using a common measurement scale, to "strictly enforced editorial policy, virtually simultaneous reforms in a number of leading journals, and the timely re-writing [of] textbooks to fit with policy recommendations." But their description of the process suggests that an accidental factor, the coincidence of several strong-willed editors, also mattered. For the classic collection of papers criticizing significance tests in psychology see Morrison and Hankel (1970), and for a more recent collection of papers see Harlow et al. (1997). Nickerson (2000) provides a comprehensive survey of this literature.

Within economics, although significance tests have been occasionally criticized for many years (see for instance White 1967 and Mayer 1980), these criticisms became prominent only in 1985 with the publication of D. N. McCloskey's *The Rhetoric of Economics* and a subsequent series of papers by McCloskey and by Stephen Ziliak that culminated in their 2008 book, *The Cult of Statistical Significance*.[4] There they charge: "Statistical significance is not the same thing as scientific finding. R^2, t-statistic, p-value, F-test, and all the more sophisticated versions of them in time series and the most advanced statistics are misleading at best. … [M]ost of the statistical work in economics, psychology, medicine, and the rest since the 1920s…has to be done over again" (Ziliak and McCloskey 2008b, xv and 18). They are "very willing to concede some minor role to even mindless significance testing in science" (48), but they declare: "Significance testing as used has no theoretical justification" (Ziliak and McCloskey 2004, 527).

Reception of Ziliak and McCloskey's book

The Cult of Statistical Significance has been widely, and in many cases very favorably, reviewed by journals in diverse fields, such as *Science*, *Administrative Science Quarterly*, *Contemporary Sociology*, *SIAM News*, *Nature*, *Medicine*, and *Notices of the American Mathematical Society*, as well as by some economics journals. Few, if any, books by contemporary economists have stirred interest in so many fields.

The main criticisms of Ziliak and McCloskey (henceforth referred to as Z-M) have come from papers in a symposium in the 2004 *Journal of Socio-Economics*, and from Kevin Hoover and Mark Siegler (2008a), Tom Engsted (2009), and Aris Spanos (2008). I will discuss most of the criticisms in the context of specific claims made by Z-M. But some of Spanos' criticisms are too general for that, since they raise fundamental and disturbing questions about the whole approach to significance testing (see Mayo 1996; Mayo and Spanos 2006; Spanos 2008). Thus, one disturbing problem is that significance testers pay insufficient attention to the problem of statistical inadequacy. If the model does not fit the data, e.g., if it is linear when the data embody a log-linear relationship, then the estimated standard errors and t-values are incorrect. Spanos calls this a problem for which no adequate solution has been discovered, and he guesses that, of the applied papers published in the *American Economic Review* over the last 30 years, less than one percent would pass a test of statistical adequacy (Spanos 2008, 163), thus implicitly agreeing with Z-M in rejecting published significance tests in economics.

[4] Following the standard practice of focusing on an author's most recent statement of his or her thesis I will primarily discuss their 2008 book. There they take a more radical position than in their prior papers. As McCloskey (2008) explained, their frustration at having their more moderate statement ignored drove them towards making stronger statements.

Among Spanos' other criticisms are that Z-M play to the galleries, with repetitious and provocative catchphrases. One might reply that this makes for exciting reading, but it can also cover up inadequacies in the analysis. He also argues that Z-M's rejection of Deborah Mayo's (1996) interpretation of significance tests as measures, not of the probable truth of the hypothesis, but of the severity of the test, is based on a misunderstanding, and that their discussion of the relations among sample size, the power of a test, and the probability that the test will reject the null is confused.

In their reply Z-M (2008a, 166) claim—without offering any justification—that Spanos "throws up a lot of technical smoke" that hides his basic agreement with them, a claim that is hard to accept. They further state that the "*sole* problem" (166, italic in original) their book deals with is the distinction between statistical and substantive significance, and they blame Spanos for not appreciating that. There is something to this argument (even if their book might not leave the reader with that impression), in that this distinction, simple as it is, needed to be reiterated and that Spanos does not give them sufficient credit for doing so. Spanos, who approaches the matter wearing the spectacles of a mathematician and statistician, is concerned about logical problems with significance tests, such as the failure to ensure that the data conform to the assumptions underlying these tests, while Z-M, wearing the spectacles of applied economists, are concerned with invalid conclusions that appear in the *American Economic Review* (*AER*) and similar journals.

The focus of this paper

I will evaluate Z-M's claims only with respect to economics, even though this may be unfair to Z-M since their criticisms may be more applicable to other fields. (The subtitle of their book is: *How the Standard Error Costs Us Jobs, Justice and Lives*.) I will try to show that although their extreme claims are unwarranted, less extreme versions of some of their claims are correct. In doing so I take a pragmatic, second-best approach, looking only at those errors in the application of significance tests that are likely to cause readers to draw substantially wrong conclusions about substantive issues. As noted by Gerd Gigerenzer (2004) and Heiko Haller and Stefan Krauss (2002), careless statements about significance tests abound. In other words, my orientation is more akin to that of an engineer solving a practical problem in a rough and ready way, than that of a mathematician seeking elegant truths (see Colander 2011).

Since this is not a review of Z-M's book I also leave aside some topics that the book discusses at length, such as the history of significance tests. Their book is as much a history of significance tests as it is a discussion of their current use. For an evaluation that encompasses this and other items that I omit, see Spanos

(2008), who is very critical of Z-M's condemnation of R. A. Fisher. Similarly, I do not discuss 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 over several decades.

I therefore focus on the lessons that applied econometricians should garner from Z-M's work and from various responses to it. This debate has generated much heat and also an all-or-nothing stance on both sides. But with first-rate economists involved on both sides, it is unlikely that either side is totally wrong. I try to show that Z-M have made a substantial contribution, albeit not nearly as great a one as they seem to think. This should not be surprising; economics shows many valuable contributions, such as the permanent income theory, that were initially accompanied by excessive claims.

The emphasis of this paper is therefore more on an evaluative presentation of the debate than on entirely new evidence. Authors of most papers can assume that most of those who read their papers are specialists who are more or less familiar with the prior debate. But this is not the case here; those for whom the paper is relevant include nearly all readers of applied econometrics papers, and presumably most of them have not read Z-M and their critics.

The meaning of significance tests

Discussions of significance tests are not always clear about what they mean by the term. Some seem to mean any standardized statistical measure of whether certain results are sufficiently unlikely to be due to sampling error, including for example Neyman-Pearson methods, while others seem to define the term more narrowly as Fisherian tests. The broader definition seems more common in psychology than in economics. Z-M deal with both definitions, but reserve their special wrath for Fisherian tests.

The literature is also unclear about the source of the variance that underlies any significance test. There are three potential sources, measurement errors, sampling errors, and specification errors. Yet the literature sometimes reads as though the only problem is sampling error, so that with a 100 percent sample significance tests would be meaningless.[5] However, as Tom Engsted (2009) points out, economists generally do not aim for true models in the sense that the only deviation

[5] Thus Z-M ridicule significance tests by pointing out that they are sometimes thoughtlessly used in cases where the sample comprises the entire universe, so that the notion of sampling error is inapplicable. In response, Hoover and Siegler (2008a) argue that although a paper may seem to include the entire universe within its sample, for instance the pegging of long-term interest rates in the U.S. after WW. II, such a paper is really intended to explain what happens *in general* when long-term interest rates are pegged. It is therefore using a sample. But Hoover and Siegler fail to note that it is not likely to be a random sample for all cases when interest rates are pegged, and to that extent standard significance tests are not applicable. However, to the extent that variations between individual observations are due to random measurement errors, t-values regain their meaning.

between their predictions and the data are sampling errors, but for models that are useful. And for that it does not matter if these models result in systematic rather than unsystematic errors. Hence, argues Engsted (2009, 394, italics in original), economists "to an increasing degree hold the view…that we should *not* expect model errors to be unsystematic…. Such models will be statistically rejected at a given significance level if the test is sufficiently powerful. Economists, therefore, to an increasing extent…evaluate economic models empirically using methods that are better suited for misspecified models than [are] statistical hypothesis tests." Specific examples he cites include dynamic stochastic general equilibrium (DSGE) models, linear rational expectations models, and asset pricing models. Spanos (2008) and Walter Krämer (2011), too, argue that Z-M do not treat the problem created by specification errors adequately.

Z-M's criticisms of significance tests

Z-M challenge economists' use of significance tests on four grounds. First, they claim that most economists do not realize that substantive significance, that is, the size of the effect that an independent variable has on the dependent variable (which Z-M call "oomph"), is vastly more important than statistical significance, or what is even worse, economists confound the two. Second, economists often commit the logical fallacy of the transposed conditional; third, they ignore the loss function; and fourth, instead of reporting confidence intervals, they usually present their results in terms of t-values, *p*'s, or F-ratios.

What matters, significance or oomph?

There are several issues under this rubric. One is: What it is that significance tests do? The second is whether the mere existence of an effect—as distinct from its size—is a legitimate scientific question. The third is the frequency with which economists and others focus on statistical instead of on substantive significance. A fourth, raised by a referee of this paper, is that we do not know how to measure oomph properly, that a "point estimator is clearly not a good measure of 'oomph'" and that "nobody knew how to address the problem!" (Anonymous 2012). But the world's work has to be done, and I will therefore arbitrarily assume that the point estimate is a sufficiently good measure of oomph, while admitting that we have here another reason for being modest in our claims.

What do significance tests tell us?

In at least some of their writings Z-M give the impression that significance tests only tell us (directly in the case of confidence intervals, and indirectly for t-values and *p*'s) the spread around a point estimate (or a sample mean) whose correct value we already know (see Horowitz 2004, 552). They suggest the following mental experiment:

> Suppose you want to help your mother lose weight and are considering two diet pills with identical prices and side effects. You are determined to choose one of the two pills for her. The first pill, named Oomph, will on average take off twenty pounds. But it is very uncertain in its effects—at plus or minus ten pounds. … Oomph gives a big effect, you see, but with a high variance. Alternatively the pill Precision will take off five pounds on average. But it is much more certain in its effects. Choosing Precision entails a probable error of plus or minus a mere one-half pound. … So which pill for Mother, whose goal is to lose weight? (Z-M, 2008b, 23)

This mental experiment is flawed (Hoover and Siegler 2008b, 15-16; Engsted 2009, 400.) It 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: Given the existence of sampling error, how confident can we be about these point estimates? Suppose your sample consists of only two cases for each pill. Shouldn't you then warn mother not to place much importance on what you told her about the two means? (See Wooldridge 2004.) But sample size alone does not tell her how much credence to give your means; variance also matters. So why not couch your warning in terms that combine sample size and variance, such as t-values, *p*'s or confidence intervals? It is Z-M's unwarranted assumption that we know the mean of the universe, rather than just the mean of the sample, that allows them to dismiss significance tests as essentially useless.

This failure to acknowledge that statistical significance is often needed to validate a paper's conclusions about oomph underlies Z-M's rejection of testing for both statistical significance as well as for oomph. Thus they write: "Statistical significance is *not* necessary for a coefficient to have substantive significance and therefore *cannot* be a suitable prescreen" (Z-M 2008b, 86, italics in original). Yes, statistical significance is not necessary for substantive significance, but that is not the issue. Statistical significance is needed to justify treating the coefficient generated by the sample as though it had been generated by the universe, i.e., as a sufficiently reliable stand-in for the true coefficient.

Part of Z-M's argument against the importance of statistical significance is couched as an attack on what they call "sign econometrics" and "asterisk econometrics", by which they mean placing asterisks next to coefficients that are significant and have the right sign, because researchers mistakenly believe that what makes a variable important is its significance along with the right sign, and not its oomph.[6] But this is not necessarily so. Putting asterisks on significant variables does not necessarily imply that they are more important than others; the importance of a variable can be discussed elsewhere. And readers should be told for which coefficients there is a high likelihood that their difference from zero is not just the result of sampling error. Mayo (1996) provides a credible argument that one should interpret a t-value, not as an attribute of the hypothesis being tested, but as an attribute of the severity of the test to which it has been subjected. And there is nothing wrong with using asterisks to draw attention to those hypotheses that have passed a severe test.[7]

Is existence a scientific question?

There is also a flaw in Z-M's claim that significance tests only tell us whether an effect exists, and that this is a philosophical and not a scientific question. But existence *is* a scientific question, in part because it may tell us whether we are using the right model (see Elliott and Granger 2004) and because it makes little sense for scientists to try to measure the size of something that does not exist. It would be hard to obtain an NSF grant to measure the density of ether. And we do observe natural scientists asking about existence.[8] Hoover and Siegler (2008b) cite a classic test of relativity theory, the bending of light near the sun, as an example where oomph is irrelevant while significance is crucial.[9] Recently there was much excitement about neutrinos allegedly traveling faster than light because relativity theory prohibits that, even if it is only trivially faster. Wainer (1999) lists three other examples from the natural sciences where oomph is not needed: (a) the speed of

[6] Z-M do, however, allow for exceptions to their condemnation, writing: "*Ordinarily* sign alone is not *economically* significant unless the magnitude attached to the sign is large or small enough to matter" (2008b, 70, first italic added, second in original).

[7] By "severity" I mean the probability that the test would reject the null even if it were true.

[8] Thus Hoover and Siegler (2008a, 27) show that, contrary to Z-M's claim, physical scientists also use significance tests. In their reply McCloskey and Ziliak (2008, 51-52) concede this, but surmise that they do so much less frequently than economists do. However, *if* physical scientists typically have larger samples than economists (perhaps because they can rerun their experiments many times) they might have less need for significance tests (see Stekler 2007).

[9] Z-M (2008b, 48-49), however, reject this interpretation of that test because statistical significance did not play a role in it. But the issue here is whether existence matters in science, and not whether significance tests happen to be used to establish existence.

light is the same at points moving at different speeds; (b) the universe is expanding; (c) the distance between New York and Tokyo is constant. Horowitz (2004) cites additional examples from physics.)

Within economics, Giffen goods provide an example. Economists are interested in knowing whether such goods exist, regardless of the oomph of their coefficients. In Granger tests, too, what counts is the existence of an effect, not its oomph (Hoover and Siegler 2008a). And that is so also when we use a likelihood ratio to choose between a restricted and a more general model (see Robinson and Wainer 2002). Even when we are interested in oomph, it does not always matter more than existence does. Consider the hiring policy of a firm. Suppose a variable measuring race has a much smaller oomph in a regression explaining the firm's employment decisions than an education variable does. 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?

One can go beyond such individual cases by dividing hypotheses into two classes. One consists of hypotheses that explain the causes of observed events. For these oomph is generally important. If we want to know what causes inflation, citing declines in strawberry harvests due to droughts will not do, even if, because of the great size of the sample, this variable has a t-value of 2. But there is also another type of model (and for present purposes one need not distinguish between models and hypotheses), one that Allan Gibbard and Hal Varian (1978) called a "caricature model", which tries to bring out important aspects of the economy that have not received enough attention. And these aspects may be important for the insight that they provide (and hence for the development of new causally-oriented hypotheses), even though the variables that represent these aspects have little oomph in a regression equation.

For cause-oriented hypotheses, regression tests can be categorized as direct or indirect. Direct tests ask about whether the variable has a large oomph, or if it explains much of the behavior of the dependent variable. If we find neither oomph nor statistical significance we consider the hypothesis unsatisfactory. But we frequently also use indirect tests. These are tests that draw some necessary implication from the hypothesis and test that, even though this particular implication is of no interest on its own. Here oomph does not matter, but the sign and significance of the coefficient do. The additional opportunities for testing that these indirect tests provide are important parts of our toolkit because we frequently lack adequate data for a direct test.

For instance, Milton Friedman (1957) encountered a problem in testing the permanent income theory directly because no data on permanent income were then available. He therefore drew implications from the theory, for example, that at any given income level farm families have a lower marginal propensity to consume than urban families, and tested these implications (see Zellner 2004). Surely, few readers of Friedman's *A Theory of the Consumption Function* found the relative size of these propensities to consume interesting for their own sake, and therefore did not have any interest in their oomph, but the sign of the difference and its significance told them something about the validity of the permanent income theory. Friedman presented eight tests of this theory. Seven are indirect tests.[10] Standing by itself, each of these seven tests is only a soft test because it addresses only the direction of a difference, and a priori, without the availability of the theory, there is a 50 percent chance that the difference will be in the predicted direction. But if all of these seven tests yield results in the direction predicted by the permanent income theory, then one can invoke the no-miracles argument.

Or suppose you test the hypothesis that drug addiction is rational by inferring that if the expected future price of drugs rises, current drug consumption falls. And you find that it does. To make sure that you have a point, you need to check whether this reduction is large enough so as not to be attributable to sampling error. But it does not have to account for a substantial decline in drug use. This does not mean that oomph never matters for indirect tests; in *some* cases it may.

There are also in-between cases where oomph matters for some purposes but not for others. Take the standard theory of the term structure of interest rates. It seems logically compelling, but it does not predict future changes in the term structure well. 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 to those who wonder why the standard theory predicts badly, regardless of the oomph of the new variable. Similarly, it would be useful to have a variant of the Fisher relation that predicts movements of exchange rates better, even if it means adding a variable with a low oomph.

Finally, the role of the coefficient's sign and its significance level are enhanced when one considers papers not in isolation, but as part of an ongoing discussion where the purpose of a paper may be to counteract a previous paper. For that it may suffice to show that in the previous paper, once one corrects for some error or uses a larger sample, the crucial coefficient has the wrong sign, or

[10] And the one direct test Friedman provided did not support his theory against the rival relative income theory of Duesenberry and Modigliani. Hence, by concluding that his evidence supported the permanent income theory Friedman put more weight on the indirect tests than on the direct test. For a detailed discussion of Friedman's tests see Mayer (1972).

loses significance, and never mind its oomph. For example, it was widely believed that, prior to the Glass-Steagall Act, banks that underwrote securities had a conflict of interest that allowed them to exploit the ignorance of investors. The evidence cited was that in the 1930s securities underwritten by banks performed worse than others. By showing that at the five-percent significance level the opposite was the case, Randall Kroszner and Raghuram Rajan (1994) challenged this hypothesis without having to discuss oomph.

None of the above denies that in many cases oomph is central. But it does mean that Z-M's claim that existence is generally unimportant, while oomph is generally important, is invalid. As the first column of the Table based on a sample of 50 papers in the *AER* shows, oomph was neither required or very important in at least eight (16 percent) and arguably in as many as 12 (24%) of the papers. But while this result rejects Z-M's strong claim, it still means that, at least in economics, the oomph of strategic coefficients usually deserves substantial emphasis.

How often do economists confuse statistical significance with substantive significance?

Very often, say Z-M. Having surveyed 369 full-length *AER* articles from January 1980 to December 1999 that contain significance tests, they claim: "Seventy percent of the articles in… the 1980s made no distinction at all between statistical significance and economic or policy significance… Of the 187 relevant articles published in the 1990s, 79 percent mistook statistically significant coefficients for economically significant coefficients." (Z-M 2008b, 74, 80).[11] Even though there are many cases, such as reduced form usage of vector-autoregressions (VARs), where the magnitude of a particular coefficient is of little interest (Engsted 2009, 400), Z-M's results are still surprising, particularly since plotting confidence intervals is the default setting of frequently used econometric software packages (Hoover and Siegler 2008a, 20). Moreover, Engsted (2009) points to several flourishing areas of economic research, such as DSGE models and return predictability, where statistical and substantive significance are clearly *not* confounded.

How then did Z-M obtain their dramatic results? They did so by giving each paper a numerical score depending on its performance on a set of nineteen

[11] Altman (2004, 523) in a less quantitative way supports Z-M's claim of frequent confusion, writing: "As both a researcher and a journal editor, I have been struck by the insistence of [*sic*] the use of tests of statistical significance as proxies for analytic significance. More often then not the writers are not aware of the distinction between statistical and analytic significance or do not naturally think of the latter as being of primary interest or importance."

questions, e.g., whether the paper refrains from using the term "significant" in an ambiguous way, and whether in the conclusion section it keeps statistical and economic significance separated (Z-M 2008b, 72-73). Such grading requires judgment calls. What is ambiguous to one reader may be unambiguous to another. Moreover, suppose a paper uses "significant" in an ambiguous way in its introduction, but in the concluding section uses it in a clear way. Should it be docked for the initial ambiguity? And what grade does a paper deserve that does not distinguish between statistical and economic significance in the conclusion, but does so at length elsewhere? It is therefore not surprising that Hoover and Siegler (2008a, 5) criticize Z-M for frequently making dubious judgments as well as for using a hodge-podge of questions, including some questions that indicate good practice, some that indicate bad practice, and some that are redundant. Moreover, as Hoover and Siegler point out, some questions duplicate others, which results in double counting and hence an arbitrary weighting.[12] And Hoover and Siegler found entirely unconvincing the five case studies that Z-M provide. Similarly, Jeffrey Wooldridge (2004, 578) wrote: "I think [Z-M] oversell their case. Part of the problem is trying to make scientific an evaluation process that is inherently subjective. It is too easy to pull isolated sentences from a paper that seem to violate ZM's standards, but which make perfect sense in the broader context of the paper." Spanos (2008, 155) calls "most" of Z-M's questions "highly problematic". Appendix A lists and evaluates each of Z-M's questions.

An alternative procedure is not to use predetermined questions to assess specific sentences in a paper, but to look at a paper's overall message and ask whether it is polluted by a failure to distinguish statistical from substantive significance. Anthony O'Brien (2004) selected a sample of papers published in the *Journal of Economic History* and in *Explorations in Economic History* in 1992 and 1996 and asked whether their conclusions were affected by an inappropriate use of significance tests. In 23 out of the 118 papers (19 percent), significance tests were used inappropriately, but in only eight of them (7%) did "it matter for the paper's main conclusions" (O'Brien 2004, 568). This low percentage, he suggested, may explain why Z-M have had so little success in moving economists away from significance tests.

In my own attempt to replicate Z-M's results I followed O'Brien, as did Hoover and Siegler, by looking not at the specific wording of particular sentences but at a paper's overall Gestalt. I used a sample of 50 papers, thirty-five of them taken from Z-M's sample (17 from the 1980s and 18 from the 1990s), and, to update the sample, fifteen papers from 2010. Specifically, I asked whether a harried

[12] For Z-M's reply and Hoover and Siegler's rejoinder, see McCloskey and Ziliak (2008) and Hoover and Siegler (2008b).

reader not watching for the particulars of significance testing would obtain the correct takeaway point with respect to significance and oomph.[13] Admittedly, this procedure requires judgment calls, and other economists may evaluate some papers differently from the way I do.[14] But Z-M's criteria also require judgment calls. So that readers can readily judge for themselves which procedure is preferable, Appendix B provides summaries of the eleven papers in my sample that Z-M give a low grade.

The second column of the Table shows the results. A "yes" means that the authors do give the oomph. Since a yes/no dichotomy often does not capture the subtlety of a discussion, dashes and footnotes indicate in-between cases. The Table treats as a "yes" cases where the authors do not discuss oomph in the text but do give the relevant coefficients in a table. It may seem necessary to discuss oomph in the text and not just in a table, because that allows the author to tell readers whether the coefficient should be considered large or small, something that may not always be obvious, particularly if the regression is in natural numbers rather than logs. For example, if we are told that by changing the bill rate by ten basis points, the Fed can change the five-year rate by one basis point, does this mean that it has sufficient or insufficient control over the latter? But even in a brief discussion in the text it may sometimes be hard to say that a coefficient is "large" or "small", because the results of the paper may be relevant for several issues, and what is a large and important oomph with respect to one issue may not be so for another.[15] And even for the same issue it may vary from time to time: When the bill rate is five percent it does not hinder the Fed as much if it requires a change in the bill rate of 30 basis points to change the five-year rate by 10 basis points as it does when the bill rate is 0.25 percent. A requirement that oomph be discussed in the text would therefore not be a meaningful criterion by which to distinguish between those who use significance tests correctly and those who don't. I have therefore used a minimal criterion, that the coefficient be given, so that readers can make up their own minds.

[13] I assume a harried reader because, given the great volume of reading material that descends upon us, it seems unlikely that most papers receive a painstaking reading. Further, I focus on a paper's main thesis, and therefore do not penalize it if it fails to discuss the oomph of a particular variable that is not strategic, even if this oomph is interesting for its own sake.

[14] In fact, in some cases I changed my mind when I reviewed an earlier draft.

[15] Moreover, even with respect to any one issue the magnitude of oomph may not answer the question of interest, because (leaving the causality issue aside) all it tells you is by how much *y* changes when *x* changes by one unit. It does *not* tell you what proportion of the observed changes in *y* is due to changes in *x*, because that depends also on the variance of *x*—a statistic that should be given, but often is not.

TABLE. Uses and misuses of significance tests in 50 *AER* papers

Papers:	(1) Oomph required or very important for topic	(2) Paper provides oomph	(3) Paper provides correct takeaway point with respect to oomph	(4) Significance tests used wrong-way-round for testing maintained hypothesis	(5) Significance tests used wrong-way-round for congruity adjustments[a]
1980s:					
Acs & Audretsch (1988)	Yes	Yes	Yes	Yes	No
Blanchard (1989)	Yes	Yes	Yes	No	--[b]
Bloom & Cavanagh (1986)	Yes	Yes	Yes	No	Yes
Borjas (1987)	Yes	Yes	Yes	--[c]	No
Carmichael & Stebbing (1983)	Yes	Yes	Yes	No	Yes[d]
Darby (1982)	Yes	Yes	Yes	No	No[e]
Evans & Heckman (1984)	No	No	Irrelevant	No	No
Froyen & Waud (1980)	No	Yes	Yes	No	No
Garber (1986)	Yes	Yes	Yes	Yes	No
Johnson & Skinner (1986)	Yes	Yes	Yes	No	No
Joskow (1987)	Yes	Yes	Yes	--[f]	No
LaLonde (1986)	Yes	Yes	Yes	No	No
Mishkin (1982)	No[g]	Yes	Yes	--[h]	Yes
Pashigian (1988)	--[i]	Yes	--[j]	Yes	No
Romer (1986)	No	Yes	Yes	Yes	No
Sachs (1980)	Yes	Yes	Yes	No	No
Woodbury & Spiegelman (1987)	Yes	Yes	Yes	No	No
1990s:					
Alesina & Perotti (1997)	Yes	Yes	Yes	No	No
Angrist & Evans (1998)	Yes	Yes	Yes	No	No
Ayres & Siegelman (1995)	Yes	Yes	Yes	No	No
Borjas (1995)	Yes	Yes	Yes	No	No
Brainard (1997)	Yes	Yes[k]	Yes	No	No
Feenstra (1994)	Yes	Yes	Yes	--[d]	No
Forsythe (1992)	No	Yes	Yes	Yes	No
Fuhrer & Moore (1995)	Yes	Yes	Yes	No	Yes[d]
Gali (1999)	No[m]	Yes	Yes	No	Yes
Ham et al. (1998)	Yes	Yes[n]	Yes	No	No
Hendricks & Porter (1996)	Yes	Yes	Yes	No	No
Hoover & Sheffrin (1992)	No	Yes	Irrelevant	Yes	No
Kroszner & Rajan (1994)	No	Yes	Yes	Yes	No
Mendelsohn et al. (1994)	Yes	Yes	Yes	No	No
Pontiff (1997)	Yes	Yes	Yes	No	No
Sauer & Leffler (1990)	No	Yes[n]	Yes	No	No
Trejo (1991)	Yes	Yes	Yes	No	No
Wolff (1991)	Yes	Yes	--[n]	No	No

Papers:	(1) Oomph required or very important for topic	(2) Paper provides oomph	(3) Paper provides correct takeaway point with respect to oomph	(4) Significance tests used wrong-way-round for testing maintained hypothesis	(5) Significance tests used wrong-way-round for congruity adjustments[a]
2010:					
Artuç et al. (2010)	Yes	Yes	Yes	No	No
Bailey (2010)	Yes	Yes	Yes	Yes	No
Bardhan & Mookherjee (2010)	No	Yes	Yes	No	Yes
Chandra et al. (2010)	Yes	Yes	Yes	No	No
Chen et al. (2010)	Yes	Yes	Yes	No	No
Conley & Udry (2010)	Yes	Yes	Yes	No	No
Dafny (2010)	No[o]	Yes	Yes	No	No
Ellison et al. (2010)	Yes	Yes	Yes	Yes	No
Fowlie (2010)	Yes	--[p]	Yes	No	No
Harrison & Scorse (2010)	Yes	Yes	Yes	No	No
Landry et al. (2010)	Yes	Yes	Yes	No	No
Lerner & Malmendier (2010)	Yes	Yes	Yes	--[q]	No
Leth-Petersen (2010)	Yes	Yes	Yes	No	No
Mian et al. (2010)	Yes	Yes	Yes	Yes	No
Romer & Romer (2010)	Yes	Yes	Yes	No	No

Notes:
a. Includes tests for breaks in series, such as tests for unit roots, lag length, breaks in time series, etc. As discussed in the text, the frequency with which decisions about data adjustments have been made on the basis of wrong-way-round significance tests is probably understated.
b. Blanchard uses a wrong-way-round test in defending his assumption of stationary, but this is mitigated by his openness about the problem and his stating that theory suggests stationarity, as well as his saying: "as is…well known, the data cannot reject other null hypotheses. … [T]he results…must be seen as dependent on a priori assumptions on the time-series properties of the series" (Blanchard 1989, 1151). I believe that this absolves Blanchard of the charge of misusing the significance test.
c. Not clear how serious the problem is here.
d. Only at an unimportant point.
e. Arguably "yes" as Darby uses results from another paper in which insignificant variables were eliminated.
f. Wrong-way-round significance test used only at a minor point.
g. Oomph not required because the paper obtains negative results for the hypothesis it tests.
h. Wrong-way-round significance test used in auxiliary, but not in the main regressions.
I. Not really needed, but would be useful.
j. Provides oomph, but at one point makes the error of interpreting the size of a regression coefficient as a measure of the extent to which this regressor accounts for changes in the dependent variable; that depends also on the variance of that variable and on the variance of the dependent variable.
k. Provides oomph sometimes, but not frequently enough.
l. Main message of paper is qualitative.
m. Required only for comparison with other hypotheses that are mentioned only briefly.
n. On most, but not all points.
o. Only to the extent that oomph is not trivial.
p. Mainly, but not completely.
q. At one point uses what seems like a wrong-way-round test, but that is ameliorated by using it only to state that this test "provides no evidence" for a difference between two coefficients.

The results shown in the second column of the Table are in sharp contrast to Z-M's. There is no case where oomph is required but is totally ignored, and in only four cases might one reasonably say that it should have been given more emphasis. Further (though this is not shown in the Table) there is no evidence anywhere of a confusion of statistical significance and magnitude. As column (3) shows, for none of the fifty papers is their takeaway point *unequivocally* wrong due to the authors having confused statistical significance with oomph, or having failed to notice the importance of oomph, and in only two cases are the proffered takeaway points *arguably* wrong.

Krämer (2011, 9) examined all empirical papers in *The German Economic Review* since its inauguration in 2000. He found in 56 percent of them: "Confusion of economic and statistical significance of estimated coefficients or effects ('significant' used for both? [or] [m]uch ado about statistically significant but economically small coefficients or effects?)" In addition, 28 percent of the papers discarded (wrongly, he believes) "economically significant and plausible effects…due to lack of statistical significance." These results seem more discouraging than mine, but that may be explained by his criteria being more stringent. My tabulation does not penalize a paper for using the term "significant" for both types of significance. Nor does it ring an alarm bell when a variable with a substantively large oomph is disregarded because its t-value is less than 2.

Neither O'Brien's results nor mine (and certainly not Krämer's) should be read as a wholesale rejection of Z-M's contention. One reason is that in papers in some lesser-ranked journals—as well as in journals outside of economics—the confusion of statistical with substantive significance could well be more common.[16] Second, even if this confusion occurs only occasionally, that is too much. Given the vast number of papers that use significance tests, even a one percent error rate would mean many errors. Hoover and Siegler (2008a) criticize Z-M for making an already well known point in fussing about the distinction between statistical and substantive significance. They are right that the point is well known in an abstract sense, but if it is ignored, even only occasionally in actual practice, it is a point well worth making. Moreover, the error rate should be zero, since the distinction between the two types of significance is an elementary issue.

[16] Mayo and Spanos (2006, 341) call the confusion between statistical and substantive significance the "[p]erhaps most often heard and best known fallacy" in using significance tests.

Wrong-way-round significance tests

Suppose you test your hypothesis that large banks benefit from scale economies. Not quite, says your computer, the relevant coefficient has the right sign, but a t-value of only 1.2. Can you now publish a paper showing that there are *no* scale economies for large banks? Only, Z-M tell us, if editors and referees are not doing their job. (See also Cohen 1994; Krämer 2011; Mayer 1980, 1993, 2001; Mayo 1996.) Except in the case of very large samples (discussed below), treating a low t-value as evidence that some effect does *not* exist is an error; Z-M refer to it as "the error of the transposed conditional."[17] Jacob Cohen (1994, 998, italics added) illustrates this error by contrasting the following two syllogisms:

> If the null hypothesis is correct, then this datum (D) can not occur.
> It has, however, occurred.
> Therefore, the null hypothesis is false.

And

> If H_0 is true this result (statistical significance) would *probably* not occur.
> This result has occurred.
> Then H_0 is probably not true and therefore formally invalid.

The second syllogism, unlike the first, is errant. For the first, since the premises are true the conclusion is also true. But in the stochastic world of the second we cannot be sure of the conclusion. To argue from the failure to disconfirm to the probability of confirmation one needs to look at the power of the test, that is, at the probability that if the hypothesis were true the test would not have disconfirmed it (see Mayo and Spanos 2006). But tests of power are unusual in economics.

An intuitively simple way of showing why one should not treat failure to disconfirm a hypothesis as confirmation of its converse is to point out that the results obtained when testing a null hypothesis fall into one of three bins: "confirmed", "disconfirmed", and "cannot tell". If the coefficient of *x* has a t-value of, say, 1.2, all this means is that the null hypothesis that *x* has no predictive value for *y* cannot be placed into the confirmed bin, but it does not authorize us to place it into

[17] For a discussion of the transposed conditional from a Bayesian viewpoint see Cohen (1994), and from a frequentist viewpoint see Mayo and Spanos (2006). For a comprehensive survey of this problem in the psychological literature see Nickerson (2000), who describes it as an example of the logical fallacy of affirming the consequent.

the disconfirmed bin. If it did, it would be easy to disconfirm any hypothesis—just use a small enough sample.[18]

Sample size plays a fundamental role here, such a fundamental role that it might be useful to call these tests "sample-size" tests instead of significance tests, which would reduce the temptation to use them the wrong way round. As Edward Leamer (quoted in Leamer 2004, 557) has remarked: "Meaningful hypothesis testing requires the significance level to be a decreasing function of sample size." Just what does a significance test reject when $t < 2$? Is it the claim that the coefficient of the regressor exceeds zero for reasons other than sampling error, or is it the adequacy of the sample? (See also Altman 1980.)

And herein lies the kernel of validity in the use of wrong-way-round significance tests. Suppose, working with a sample of 100,000, one finds that a variable that the hypothesis predicts to be positive is positive but not significant. With such a large sample we have a strong expectation that a coefficient that is significant in the universe is also significant in our sample, so its insignificance speaks against the hypothesis. A power test, if available, would help in deciding. If not, we have to make a subjective judgment.

Two further arguments against wrong-way-round significance tests

My discussion seems to contradict the familiar Popperian principle that, due to the problem of induction, data can never prove a hypothesis but can only fail to disconfirm it. And if, again and again, independent and severe tests fail to disconfirm it, that justifies tentatively accepting it, at least as a working hypothesis.[19] This may seem to imply that when in a series of independent severe tests the beta coefficients of the relevant variable all have low t-values, we can treat the hypothesis as disconfirmed. But when philosophers speak of "failure to disconfirm" they mean failure to provide *any* evidence against the hypothesis. And

[18] Moreover, the assumption that failure to disconfirm implies that the converse has been confirmed has an unwelcome implication. Suppose an economist tests the hypothesis that $y = x$ and finds that, though in his data set $x = 5$ and $y = 6$, he cannot reject at the five-percent level the null hypothesis that this difference is due only to sampling error. He therefore concludes that the data do not disconfirm his hypothesis, and that this increases its plausibility. His sister tests the contrary hypothesis that $y < x$, and since she uses the same data also finds that $x = 5$ and $y = 6$. Since in her test the difference between the predicted and the actual coefficients is again not significant, she, too, claims that the data confirm her hypothesis. Who is right?

[19] For this the tests have to be hard, and not in Blaug's (1989, 256) classic description of much econometric testing, as "playing tennis with the net down", so that it would take a highly implausible combination of circumstances for a false hypothesis to have passed all of these tests. As Hoover (2011) has suggested the null hypothesis is therefore often a poor foil to the maintained hypothesis.

even if a coefficient with the right sign is significant only at, say, the 40 percent level, it still provides *some* evidence in favor of—and not against—the hypothesis that the coefficient is positive. In Mayo's formulation, it has failed a severe test, but it *has* passed a less severe test.

Congruity adjustments

My own survey (see the Table) distinguishes between significance tests that deal directly with a maintained substantive hypothesis and those that deal with whether a hypothesis needs certain adjustments—which I will call "congruity adjustments"—to make it congruent to the probability model that generated the data. For example, the data might have a log normal distribution while the hypothesis was initially formulated in terms of natural numbers. Other examples include tests for unit roots, serial correlation, heteroscedasticity, and breaks in the data.[20] Standard regression packages provide options for such adjustments, but their appropriateness in terms of the underlying hypothesis needs to be considered. The prevailing procedure is to make congruity adjustments only if one can reject at the five-percent level the null hypothesis that no adjustment is needed. But it is hard to see why the burden of the proof is thus placed on the hypothesis that an adjustment is needed. Brandishing Occam's razor will not do because not adjusting is only computationally and not philosophically simpler than adjusting. What we need, but do not have, is an explicit loss function. In testing the maintained hypothesis the usual justification for the five-percent level is that a Type II error is more damaging to science than a Type I error. But is this the case when one decides whether to adjust for, say, serial correlation? *Perhaps* a *p* value of 0.50 would be more appropriate. Unless we can decide on the appropriate loss function we should, when feasible, require our results to be robust with respect to these potential adjustments.

Frequency of wrong-way-round tests

As column (4) of the Table shows, even if one leaves congruity adjustments aside and looks only at tests of substantive hypotheses, 10 papers (20 percent) fall into the trap of assuming that failure to confirm a hypothesis at the five-percent level is equivalent to treating its negation as confirmed. And, as discussed in the Table notes, there are five additional papers that fall into the trap if one applies a stricter standard than I did, giving a potential total of 30 percent. Previously (Mayer

[20] The same problem arises when deciding on the appropriate lag length by truncating when lagged coefficients become insignificant.

2001), I looked at papers in all the 1999 and 2000 issues of the *American Economic Review* and the *Review of Economics and Statistics* and found six cases of this confusion. The problem is worse in political science. There, Jeff Gill (1999) found that in four leading journals significance tests were used the wrong way round in 40 to 51 percent of the relevant cases.

The last column of the Table, which deals with congruity adjustment, shows four or at most six papers suffering from this error. But this underestimates—probably very substantially—the actual number of cases because it includes only those in which authors discuss their congruity adjustments. Presumably in many more papers authors tested for serial correlation, etc., and decided not to make the adjustment because it was not required at the five-percent level.

Permissible uses of wrong-way-round significance tests

However, none of the above implies that—even when the sample is not very large—one can never even tentatively reject a hypothesis because of a low t-value. Suppose $p = 0.15$. One can then consider a reasonable minimal value for the coefficient that would support the hypothesis, estimate the p for that, and from that decide on the credibility of the hypothesis (see Berg 2004; Mayo and Spanos 2006). For example, take the hypothesis that illegal immigration has lowered real wages in a certain industry. If this cannot be rejected at the five-percent level, one can test the hypothesis that it has reduced wages by more than a trivial two percent. Another possibility is to rely on a combination of tests. If on many independent tests a coefficient has the right sign but is not significant, one can either formally or informally reject the hypothesis that it is only sampling error that gives the coefficient the right sign (see Pollard and Richardson 1987). It is along these lines that Jean Perrin confirmed Einstein's interpretation of Brownian motion (see Mayo 1996, chapter 7).

The loss function

Although most economists think of significance tests as telling us something about a hypothesis, Z-M view it in a Neyman-Pearson framework, as telling us whether a certain course of action is justified. That requires a loss function.[21] The

[21] Z-M tell us: "[W]ithout a loss function a test of statistical significance is meaningless. … [E]very inference drawn from a test of statistical significance is a 'decision' involving substantive loss… Accepting or rejecting a test of significance without considering the potential losses from the available courses of action…is not ethically or economically defensible." (2008b, 8-9, 15)

main issue here is by whom and at what stage it should be introduced, specifically whether it should be considered as part of the significance test, or instead as part of a larger problem for which we will use the results of the significance test as just one of several considerations. There is also the question of how to interpret a loss function in some of the situations in which we just want to satisfy our curiosity, and not to appraise some policy.

The conventional view is that the econometrician should deal with positivistic issues and turn the results over to the policymaker, who consults a loss function in deciding what action to take.[22] This has two putative advantages. First, it places most value judgments outside of economics. Second, as Hoover and Siegler (2008a) remind us, it avoids the problem that an econometrician's results may be relevant for many different policies, each of which calls for its own value judgments and hence has its own loss function. How is the econometrician to know all these loss functions, particularly when some of the questions which her work can answer will arise only in the future?[23] For example, here are the titles of the first five *AER* papers listed among my references: "Innovations in Large and Small Firms, An Empirical Analysis"; "The Welfare State and Competitiveness"; "Children and their Parents' Labor Supply, Evidence from Exogenous Variations in Family Size"; "Race and Gender Discrimination in Bargaining for a New Car"; "Momma's Got the Pill: How Anthony Compstock and Griswold vs. Connecticut Shaped U.S. Childbearing". How could their authors have determined the appropriate loss function?[24]

Admittedly, this response to Z-M is not entirely satisfactory because in doing the econometrics an econometrician, particularly an LSE econometrician, often has to make decisions based, at least in part, on significance tests, and those de-

[22] It is not clear whether Z-M agree. In many of the instances they give where loss functions are needed one could allow the econometrician to function as the policymaker.

[23] As Hoover and Siegler (2008a, 18) point out, one needs a loss function when deciding how strong to make a bridge, but not to set out the laws used to calculate its strength. Admittedly, postmodernists have criticized the claim that scientific statements can avoid all value judgments, and Rudner (1953) presents a criticism that zeros in on significance tests. However, even if one concedes that science cannot be purged completely of value judgments, one should do so to the extent one can. Even if there is no watertight dichotomy between value judgments and positive judgments, for practical purposes it is often a useful distinction because it is an instance of the division of labor. Policymakers are better at (and have more legitimacy in) making value judgments than econometricians are, and the econometrician's task here is merely to draw their attention to any significant value judgments implicit in the results he presents to them.

[24] Moreover, evaluating policy options correctly requires more than combining econometric results with value judgments. It also requires judgments about the probable gap between the policy as proposed and as it is likely to emerge from the political and administrative caldrons, as well as its unintended effects on factors such as trust in government (see Colander 2001). A policymaker is probably better equipped to deal with such problems than is an econometrician.

cisions cannot be passed on to the policymaker. But since we do not have a relevant loss function for these cases, this qualification has no practical relevance.

The only workable solution is to designate the users of econometric estimates as the ones who should apply the appropriate loss function. Presumably Z-M's objection to this is that policymakers or other users may fail to apply the appropriate loss function and implicitly assume a symmetric one. This concern may well be justified. But the solution is to educate users of significance tests rather than to impose impossible demands on their providers.

Does a loss function have any relevance for deciding what to believe when it is just a matter of knowledge for knowledge's sake? (See Hoover and Siegler 2008a, 18.) The intuitively plausible answer in most cases is no, but how do we know that what on the surface seems like a belief that has no policy implications, will not ultimately have policy implications—perhaps implicitly, by changing the core of our belief set? But as a practical matter, in most cases where we just seek knowledge for knowledge's sake we do not know the loss function, so the symmetrical one implicitly assumed by significance tests is no more arbitrary than any other.

Presenting the results of significance tests

In economics and psychology the four most common ways of presenting the results of significance tests are t-values, *p*'s, F's, and confidence intervals. In response to Z-M's criticism of reporting just t-values, Hoover and Siegler (2008a) and Spanos (2008) point out that if readers are given, as they normally are, the point estimate and either the standard error or the t-value or else the confidence intervals, they can readily calculate the two other measures, so that it does not matter which one they are given. That is so. But it does not address the question of which measure is preferable, given that many readers are time constrained and therefore likely to look only at the measure explicitly provided.

Choosing between the measures

The choice between t-values and *p*'s and F's is inconsequential. What is important is the choice between any of them and confidence intervals. Confidence intervals have substantial advantages, and it is therefore not surprising that they have become more common in the medical literature, and that the American Psychological Association's Board of Scientific Affairs recommended that all estimates of size effects be accompanied by confidence intervals (see Fidler et al. 2004a; Stang, Poole, and Kuss 2010) First, confidence intervals make it much harder to ignore oomph since they are stated in terms of oomph (see McCloskey and Ziliak (2008, 50); Hubbard and Armstrong (2006, 118)). Second, confidence intervals do not lend themselves as readily to wrong-way-round use as do t-values, F's, and *p*'s. While someone might mistakenly treat a hypothesis as disconfirmed because the t-value of the important regressor is, say, only 1.8, she is at least somewhat less likely to do so if told that its upper confidence interval shows it to be important. Confidence intervals thus reduce the hazards highlighted by Z-M's two main criticisms.

Third, when presenting t-values there is a temptation, often not resisted, to mine the data until t ≥ 2 (see Brodeur et al. 2012). Presenting confidence intervals instead is likely to reduce this temptation. Fourth, the use of t-values or *p*'s generates an anomaly that confidence intervals are likely to avoid. It would be almost impossible to publish an applied econometrics paper that does not take account of sampling error. Yet, when an economist uses a coefficient generated by someone in a prior paper she usually employs only the point estimate, thus totally disregarding sampling error. If a paper presents confidence intervals, there is at least a chance that someone using its findings would undertake robustness tests using these confidence intervals.

In some econometric procedures confidence intervals are used frequently. Thus Hoover and Siegler (2008a, 20) point to their use in connection with impulse response functions. Hoover and Siegler note also that confidence intervals are the default setting for VARs in commonly used software packages and are typically reported when using autocorrelation or partial autocorrelation functions, as well as in connection with hazard functions and survivor functions. But in many other situations they are not reported. In my sample of *AER* papers many more papers provided t's than confidence intervals. One reason could be that for many papers confidence intervals would reveal the substantial imprecision of the paper's results, and thus their limited use for policymaking. Congress is not greatly helped if told that a stimulus somewhere between $100 billion and a $1 trillion is needed (cf. Johnson (1999, 769). And even papers that do not aim at direct policy conclusions or at forecasting can face a similar problem. To be told that the 1929 stock market decline accounted for somewhere between two and 80 percent of the subsequent decline in GDP would not satisfy one's curiosity.[25]

Is a standardized acceptance level appropriate?

The prevalence of a standardized acceptance level for *t*'s and *p*'s has the obvious advantage of eliminating the need for readers to make a decision, and it is also

[25] Johnson (1999, 769) provides a whole list of invalid arguments that someone might give for not using confidence intervals.

easier for them to remember that a coefficient is significant than that it is, say, 2.3. But a standardized level also has two disadvantages. One is that an author, fearing that his papers will be rejected unless $t \geq 2$, has a strong incentive to ensure that it does so, even if it means running numerous and quite arbitrarily selected variants of his main regression—including ones that do not conform to the hypothesis as well as his original regression does—until eventually one yields a qualifying t-value.[26] Second, suppose that there are five independent studies of the effect of x on y. All find a positive effect, but their t-values are only 1.8, 1.7, 1.6, 1.5, and 1.4. If they are rejected because of this, the file-drawer problem in any subsequent meta-analysis is exacerbated, and a scientific result may thus be lost.[27] The availability of working-paper versions of articles that have been rejected because of low t-values does not eliminate this problem entirely because some researchers may not proceed to the working-paper stage if their results are not significant, or because the available meta-analyses may not cover working papers.

Conclusion

We should reject Z-M's claim that most of the significance testing done by economists is invalid, but we should also reject the idea that, at least in economics, all is well with significance tests. A basic problem with Z-M's claim is that, at least at times, they fail to realize that the purpose of a significance test is not just to test the maintained hypothesis, but to test whether the researcher's reliance on a sample instead of the entire universe invalidates her results, and that significance tests can therefore be treated as a requirement of data hygiene. Moreover, Spanos (2008) may well be right that statistical misspecification is a more serious problem for econometrics than is the misuse of significance tests. The same may perhaps be true for computational errors (see Dewald, Thursby, and Anderson 1986; McCullough and Vinod 1999) and even for inattention to the meaning and accuracy of the data (see Cargill 2012; Corrado, Hulton, and Sichel 2006; Reinsdorf 2004). But misuse of significance tests is an easier problem to cure.

Nonetheless, 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

[26] Keuzenkamp and Magnus (1995, 18) report that the *Journal of the Royal Statistical Society* (JRSS) has been called the *Journal of Statistically Significant Results*.

[27] For estimates of how seriously the results of meta-analytic studies are distorted by the unwillingness of authors to submit papers with low t values, and the unwillingness of journals to publish such papers, see Sterling, Rosenbaum, and Weinkam (1995) and Brodeur et al. (2012) and the literature cited therein.

is both severe enough and occurs frequently enough to present a serious—and inexcusable—problem. (Perhaps editors should specifically ask referees to check the use of significance tests.) And the error is probably much more widespread when deciding whether to adjust for serial correlation, heteroscedasticity, etc. Someone who wants to argue that the great majority of time-series econometric papers are flawed due to a misuse of significance tests should focus on congruity adjustments. Z-M are also right that an explicit loss function needs to be used when making policy decisions. But the econometrician is generally not able to do so and should leave that up to the policymaker. Also, because many readers are harried, Z-M are correct in their claim that confidence intervals are usually more informative than are t-values or p's.

Moreover, 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. But there is a danger that their reliance on some flawed arguments, as well as their vehemence and overreach, will tempt economists to dismiss their work.

Appendix A: Critique of Z-M's criteria for evaluating significance tests

Note: All quotations are from Ziliak and McCloskey (2008b). In the original the first sentence of each paragraph is in italics. After quoting and sometimes elaborating the criterion, I give my evaluation.

Criterion 1: "Does the article depend on a small number of observations such that statistically 'significant' differences are *not* forced by the large number of observations?" (p. 67) *Evaluation*: This is not an appropriate criterion for capturing a misuse of significance tests. The question that these tests are intended to answer is whether we can reject the possibility that the observed difference can reasonably be attributed merely to sampling error. If it cannot, it does not matter whether that is due to the difference (or the regression coefficient) being large relative to the variance, or to the sample being large. Z-M justify their criterion by saying that we know that with a large enough sample every difference will be significant. But even if that were true, it would be irrelevant, because the function of significance tests is to tell us whether we can claim that the existence of a difference (or the nonzero value of a coefficient) in our sample implies the same for the universe, regardless of whether it does so because the sample size is large, or the difference is large relative to the variance.

Criterion 2: "Are the units and the descriptive statistics for all regression variables included? … No one can exercise judgment as to whether something is importantly large or small when it is reported without units or a scale along which to judge them large or small…" (67) *Evaluation*: What matters is not the comprehensibility of "all" variables, but only of the strategic ones. Hence, this criterion is too tough. Second, authors need not specify the units and descriptive statistics if they are obvious. Apart from that, Z-M have a valid point, but one that has no discernible relation to significance tests per se. If I publish a table for which the units are neither defined nor obvious (say a table in which the units are $10) I am failing to inform readers about my findings, whether I run significance tests or not.

Criterion 3: "Are the coefficients reported in elasticity form, or in some interpretable form relevant for the problem at hand, so that the reader can discern the economic impact? … [O]ften an article will not give the actual magnitude of the elasticity but merely state with satisfaction its statistical significance." (67) *Evaluation*: This is often a valid criterion when applied not to every regression coefficient that is presented but only to the strategic ones. But even then, not always. As discussed in the text, there are cases in which it is the t-value and not the oomph that matters. So this criterion is valid only some of the time.

Criterion 4: "Are the proper null hypotheses specified? Sometimes the economists will test a null of zero when the economic question entails a null quite different from zero…." (68) Z-M's example is testing whether the income elasticity of money is unity. *Evaluation*: This criterion is valid only if the article, after mentioning the (irrelevant) result of testing against zero, does not go on to perform the proper test as well.

Criterion 5: "Are the coefficients carefully interpreted?" (68) Z-M's example is a regression of a person's weight on his height and miles walked per week, where the height variable is statistically significant and the miles-walked variable is not, though its coefficient is large. These results do not imply that if you want to lose weight without exercising just grow taller. *Evaluation*: Yes, this is right, but it is a problem of whether the regression results have been interpreted correctly, and not of significance tests per se. The mistake would be there even if the author had never run a significance test and relied entirely on the large oomph of height.

Criterion 6: "Does the article refrain from reporting *t*- or *F*- statistics or standard errors even when a test of significance is not relevant? A No on this question is another sign of canned regression packages taking over the mind of the scientist." (68-69) Z-M give the example of applying a significance test when the sample consists of the entire universe, a problem I discuss in footnote 5. *Evaluation*: Yes, reporting meaningless measures should be avoided.

Criterion 7: "Is statistical significance at its first use merely one of multiple criteria of 'importance' in sight? Often the first use will be at the crescendo of the article, the place where the author appears to think she is making the crucial factual argument. But statistical significance does not imply substantive significance. … Articles were coded Yes if statistical significance played a second or lower-order role, at any rate below the primary considerations of substantive significance." (69) *Evaluation*: As discussed in the text, there are cases in which statistical significance *should* play a primary role. Second, why is it necessarily wrong to stress statistical significance at the first use or crescendo if the article adequately discusses substantive significance at another point? An author may well want to discuss first whether to believe that the observation, say a positive regression coefficient in her sample, reliably tells us anything about the universe, or could just be dismissed as perhaps due to sampling error, and discuss substantive significance later.

Criterion 8: "Does the article mention the power of the test?" (69) *Evaluation*: If the test rejects the hypothesis there is no reason why its power need be mentioned. Hence it is not relevant in some of the cases.

Criterion 9: "If the article mentions power, does it do anything about it?" (69) *Evaluation*: This criterion partially overlaps with the previous one, and is subject to the same criticism. Treating them as separate criteria gives this criterion sometimes a double weight.

Criterion 10: "Does the article refrain from 'asterisk econometrics,' that is, ranking the coefficients according to the absolute size of their *t*-statistics?" (70) *Evaluation*: Presumably what Z-M mean with "ranking" is the order in which the variables and their coefficients are listed in a table. If so, while such a ranking may enhance a reader's inclination to overvalue statistical significance, it does not itself amount to an incorrect use of significance tests, and is more a matter of style.

Criterion 11: "Does the article refrain from 'sign econometrics,' that is, noting the sign but not the size of coefficients? The distribution-free 'sign test' for matched pairs is on occasion scientifically meaningful. Ordinarily sign alone is not *economically* significant, however, unless the magnitude attached to the sign is large or small enough to matter." (70, italics in original) *Evaluation*: As shown in the text, there is more scope for sign tests in economics than Z-M allow. However, in other cases this is a valid criterion. But it is unlikely that there are many such cases, because it would be strange if the table giving the sign does not also give the coefficient.

Criterion 12: "Does the article discuss the size of the coefficients at all? Once regression results are presented, does the article ask about the *economic* significance of the results?" (70, italics in original) *Evaluation*: As just mentioned there is some scope for sign-only significance tests. However, for many (probably most) papers, Z-M are right; the size of coefficients does matter, and it would often help the reader if it were discussed. But does it *have* to be discussed? It is not clear how much convenience to the reader an author is obligated to provide. If the economic meaning of the coefficient is complex, then an efficient division of labor requires the author to discuss it. However, in some (many?) cases economic significance may be too obvious to require discussion, e.g. an elasticity of hours worked with respect to the wage rate of 0.001. Or the purpose of the article may be to reject previously published papers that do discuss the economic significance of their coefficients, which therefore does not have to be discussed again. Thus it is not clear whether an article should be faulted, and by how much, for presenting oomph only in tables.

Criterion 13: "Does the article discuss the scientific conversation within which a coefficient would be judged 'large' or 'small'?" (71) *Evaluation*: This is not always needed. It may be obvious, or there may not be much of a prior conversation. Moreover, as explained in the text, in some cases only the sign or t-values matter.

Criterion 14: "Does the article refrain from choosing variables for inclusion in its equations solely on the basis of statistical 'significance'? … [T]here is no scientific reason—unless a reason is provided, and it seldom is—to drop an 'insignificant' variable. If the variable is important substantively but is dropped from the regression because it is Fisher-insignificant, the resulting fitted equation will be misspecified…." (71) *Evaluation*: This is discussed in the text.

Criterion 15: "Later, after the crescendo, does the article refrain from using statistical significance as the criterion of scientific importance? Sometimes the referees will have insisted unthinkingly on a significance test, and the appropriate *t*'s and *F*'s have therefore been inserted." (72) *Evaluation*: Without asking the authors, we cannot know whether the inclusion of a significance test after the crescendo was the author's own idea, or was forced on her. But why does the origin of the significance test matter? If it shows that the estimated substantive significance of a coefficient is not just the product of sampling error, it is useful regardless of what prompted it.

Criterion 16: "Is statistical significance portrayed as decisive, a conversation stopper, conveying a sense of an ending?" (72) *Evaluation*: Z-M treat a positive answer as an error. Once again, in some situations it is not. Suppose someone published a paper relating the growth rates of countries to the level of their corporate income tax, and found a negative correlation. If you now write a paper showing that the difference in the growth rates is not statistically significant, and should therefore not be treated as convincing evidence on the appropriate level of corporate income taxes, are you making a mistake?

Criterion 17: "Does the article ever use an independent simulation—as against a use of the regression coefficients as inputs into further calculations—to determine whether the coefficients are reasonable?" (72) *Evaluation*: Such simulations may be useful in some perhaps many cases, but should every failure to use simulations count as a fault? As Wooldridge (2004) points out, useful simulations are sometimes not feasible.

Criterion 18: "In the concluding sections is statistical significance separated from policy, economic, or scientific significance? In medicine and epidemiology and especially psychology the concluding sections are often sizeless summaries of significance tests reported earlier in the article. Significance this, significant that. In economics, too." (72-73) *Evaluation*: I doubt that in economics this is an accurate description of the concluding sections of many articles. And in those cases where significance is the point at issue, it should not count against the article.

Criterion 19: "Does the article use the word *significant* unambiguously?" (73) *Evaluation*: Yes, ambiguity is bad. But that does not mean that the article misuses significance tests.

In summary: Although any attempt to fit the results of this evaluation of Z-M's criteria into a few broad classes requires some judgment calls, I would classify seven of the nineteen criteria (1, 2, 5, 7, 10, 15 and 19) as invalid, ten (3, 4, 8, 9, 11, 12, 13, 16, 17 and 18) as valid in some cases, but not in others, one (14) as debatable, and another (6) as only weakly relevant because little damage results from not satisfying it. However, this judgment is the product of looking at each criterion in isolation and of applying it in a fairly mechanical way to all significance tests. A more nuanced procedure that allows for the fact that not all nineteen criteria are applicable to every significance test, and that different criteria have different weights in particular cases, might result in a much more favorable judgment. But that is similar to looking at the Gestalt of the significance test, as is done in the text.

Appendix B: Reappraising eleven papers that Z-M rank "poor" or "very poor" with respect to their use of significance tests

Ziliak and McCloskey (2008b, 91-92) classify papers published in the *AER* during the 1990s into five categories with respect to their use of significance tests: "exemplary" (6 percent); "good" (14%); "fair" (22%); "poor" (37%); and "very poor" (20%). Eleven of the papers that they rank "poor" or "very poor" are also in my sample, and I discuss them here, classifying them into four categories: "good", "fair", "marginal", and "bad". I start with the ones that Z-M rank lowest, so that the first three are ones that Z-M classify as "very poor", and the others are papers they classify as "poor".

1. S. Lael Brainard (1997), "An Empirical Assessment of the Proximity-Concentration Trade-Off between Multinational Sales and Trade"
Brainard evaluates the proximity-concentration hypothesis that predicts that firms expand across national borders when the benefits of closer access to their customers exceed the benefits obtainable from economies of scale. He builds a model embodying this hypothesis and then runs the required regressions. In the text he only discussed the signs and significance of the variables, but his tables provide the coefficients. And since his regressions are in logs, these coefficients are easy to interpret. All the same, since hasty readers may not bother to look at these tables, it would have been better to discuss the oomph of the strategic variables in the text. But Z-M's grade of "very poor" seems unjustified, and a grade of "good", or at the least "fair", seems more appropriate.

2. Stephen Trejo (1991), "The Effect of Overtime Pay Regulation on Worker Compensation"
To see whether regulations governing overtime pay, such as the time-and-a-half rule, affect total labor compensation, or whether firms offset the requirement to pay more for overtime by lowering regular wage rates, Trejo first compares the extent to which firms comply with overtime-pay regulations for workers at and above the minimum wage since firms are much more likely to lower regular wage rates that are above the minimum wage than those that are at minimum-wage level. Trejo therefore compares compliance rates with the overtime-pay rule for workers at and above the minimum wage. He finds a statistically significant difference, with firms complying less frequently with the time-and-a-half requirement when regular wages are at the minimum level. This is consistent with a model in which

firms—when minimum wage laws do not prevent it—cut regular wages to compensate for having to pay overtime rates. Trejo found that "the estimated effects of this variable are relatively large… Within the covered sector, minimum-wage workers are associated with 3–9 percentage points lower probability of being paid an overtime premium" (729). Moreover, the harder it is for firms to reduce regular wage rates to offset paying time and a half for overtime, the greater is their incentive not to exceed the forty-hour limit. One should therefore find greater bunching of workers at the forty-hour level for firms that pay just the minimum wage than for firms that have more scope to reduce regular wages. And Trejo's regressions confirm that such bunching occurs, with the coefficient of the relevant variable being both "positive and relatively large" (731). He also investigates whether straight-time pay adjusts fully to offset the requirement for overtime pay. It does not. But still, the coefficient that shows the adjustment of straight-time pay is "negative and statistically significant" (735). He leaves it to the reader to obtain its magnitude from the accompanying tables. In a final set of regressions using weekly data, Trejo finds that the coefficients of a variable whose (positive) significance and importance would contradict his hypothesis, do not "achieve statistical significance, are negative in 1974, and in the other years are always less than a third of the value predicted by" the rival theory (737). All in all, the treatment of significance tests in this paper deserves a grade of "good".

3. Randall Kroszner and Raghuram Rajan (1994), "Is the Glass-Steagall Act Justified? A Study of U.S. Experience with Universal Banking"

The Glass-Steagall Act of 1933 prohibited commercial banks from underwriting and trading in corporate securities. A major reason was to avoid potential conflicts of interest, such as a bank taking advantage of its greater information about its borrowers by underwriting securities issued by its weaker borrowers, so that these borrowers can repay their loans to the bank. Kroszner and Rajan investigate whether banks actually succeeded in taking such advantage of inside information by seeing whether securities underwritten by commercial banks or their affiliates performed worse than those issued by investment banks. To do that they constructed 121 matched pairs of security issues underwritten by commercial banks and by investment banks. What they found strongly contradicts the asymmetric-information hypothesis; securities issued by investment banks suffered about 40 percent more frequent defaults than those issued by commercial banks and their affiliates. When measured by the dollar volume of defaults, the difference in favor of commercial banks and their affiliates is even greater. Kroszner and Rajan also show that the difference in default rates is even greater for bonds below investment grade, which is inconsistent with the hypothesis that commercial banks were taking advantage of naïve investors. For some of their regressions they provide both the significance and oomph in their text, while for some others they provide the oomph only in their tables. This is justified because their aim was to challenge the then widely accepted hypothesis that allowing commercial banks and their affiliates to underwrite securities creates a conflict of interest. For that, it suffices to show that the relevant differences have the wrong sign. This paper therefore deserves at least a "good".

4. Robert Feenstra (1994), "New Product Varieties and the Measurement of International Prices"

This is primarily a theoretical paper on how to incorporate new product varieties into demand functions for imports, but it illustrates the procedure by estimating the income elasticity for six U. S. imports. Feenstra cites these elasticity estimates, and thus the oomph, extensively in the text, not just in the tables. He does, however, at one point use a wrong-way-round significance test. But this point is not important for the paper, and I therefore classify it as "marginal".

5. Jeffrey Fuhrer and George Moore (1995), "Monetary Policy Trade-Offs and the Correlation Between Nominal Interest Rates and Real Output"

This paper estimates a small model that explains the observed relation between changes in the short-term interest rate and output, using both VARs and a structural model. It presents its results mainly by charts of autocorrelation functions and autocovariance functions, so that no t-values are mentioned and a reference to "significance" appears only once. That is when Fuhrer and Moore report that they chose the lag lengths for the regressors by reducing "the lag length until the last lag remains statistically significant and the residuals appear to be uncorrelated" (220). Since, as discussed above, that is a questionable use of significance tests, I put this paper into the "marginal" bin, though Fuhrer and Moore should not be castigated for using what is a standard procedure.

6. Ian Ayers and Peter Siegelman (1995), "Race and Gender Discrimination in Bargaining for a New Car"

Ayers and Siegelman sent black and white, and male and female testers to Chicago area car dealers, and compared the prices they were offered. Their regressions show that the race and gender of testers "strongly influence both the initial and final offers made by sellers" (309). Throughout the text they cite numerous oomphs, for example: "For black males, the final markup was 8–9 percentage points higher than for white males; the equivalent figures are 3.5–4 percentage points for black females and about 2 percentage points for white females" (313). Moreover, Ayres and Siegelman sometimes cite oomph even when the t-statistics are far from significant. But at one, minor point, they do use significance tests the wrong way round. Hence, I give their paper a "marginal" grade.

7. Edward Wolff (1991), "Capital Formation and Productivity Convergence Over the Long Run"

Wolff investigates the international convergence of productivity and tests three explanatory hypotheses: the catch-up hypothesis, which implies that the further a country lags technologically, the faster will be its rate of catch-up, an alternative hypothesis that convergence in labor productivities is due to convergence in factor intensities, and a third hypothesis that there exist positive effects of capital accumulation on technological progress. Wolff runs several regressions. While providing the magnitude of regression coefficients in his tables, in his text Wolff discusses primarily their signs and significance. And at one rather peripheral point he excludes two potential regressors because their coefficients are not significant, even though his sample is fairly small. Hence, one might argue that he uses these significance tests the wrong way round. But, given the need to limit somehow the huge number of regressors that one might potentially include (and *perhaps* also the frequency with which insignificant variables are dropped in economics), it seems that "marginal" is a more appropriate grade than the "poor" that Z-M label it.

8. Kenneth Hendricks and Robert Porter (1996), "The Timing and Incidence of Exploratory Drilling on Offshore Wildcat Tracts"

When wildcat drillers bid successfully on drilling leases, they have to decide whether to incur the cost of actually drilling on these leases. In making this decision they look at what owners of other leases in the area are doing, since the productivity of wells within the same area is likely to be correlated. Each leaseholder therefore has an incentive to wait and see how successful others are. How is this problem resolved in practice? After developing the required theory, Hendricks and Porter present tobit regressions of the logs of the discounted annual revenue from drilling tracts. They provide the regression coefficients and t-values in their tables, while in their text they take up the important t-values and some of the regression coefficients. That they discuss only some of the coefficients in the text is not a serious problem because the coefficients as given in the tables are easy to interpret since their variables are measured in logs and have straightforward meanings. Hence, this paper deserves a "good".

9. Albert Alesina and Robert Perotti (1997), "The Welfare State and Competitiveness"

The basic idea of this paper is that a rise in taxes on labor to finance enhanced benefits for pensioners or the unemployed causes unions to press for higher wages, which results in a loss of competitiveness. The distortions thus introduced are greater the stronger are unions, until we reach the point when wage negotiations move to the national level where unions internalize the negative effects of their policies. After developing a model built on these insights, Alesina and Perotti estimate it for a panel of the manufacturing sectors of 14 OECD countries. In doing so they discuss extensively, not just the t-values, but also the regression coefficients. Since these coefficients have clear-cut meanings, the reader is well informed about oomph. And since there are no instances of wrong-way-round significance tests, the paper deserves a "good", not the "poor" that Z-M give it.

10. Jordi Gali (1999), "Technology, Employment and the Business Cycle"
Gali presents a test of real business cycle theory, focusing on the theory's (counterfactual) positive correlation between labor productivity and hours worked, a correlation that can potentially be explained by other shocks. He builds a VAR model embodying both types of shocks. In this model a technological shock must affect labor productivity *permanently*, and Gali uses that as his identifying restriction. In presenting his results he not only gives the regression coefficients in his tables, and presents numerous impulse response functions, but also frequently discusses oomph in his text. There is, however, one place where he uses significance tests wrong way round. This is in deciding whether to adjust for cointegration. When dealing with U.S. data (his main results) he correctly runs his regressions in both ways (and gets similar results), but does not do that when dealing with foreign data. All in all, his use of significance tests deserves a "fair".

11. Robert Mendelsohn, William Nordhaus, and Daigee Shaw (1994), "The Impact of Global Warming on Agriculture: A Ricardian Analysis"
The usual way economists have studied the impact of climate change is to fit production functions containing climate variables for various crops. But as Mendelsohn, Nordhaus, and Shaw point out, such a technological approach overestimates the losses from climate change, because it ignores that farmers can respond by changing both their production technology and their crop mix. Instead, the authors allow for adaptations, such as a shift to entirely new uses for land, by adopting a "Ricardian" approach that looks at how differences in climate affect, not the output of particular crops, but the rent or value of farmland. To do that they regress average land value and farm revenue for all counties in the lower 48 states on climate and non-climate variables. Since in presenting their results they put much greater stress on oomph (that is, on changes in the dollar value of harvests as climate changes) than on t-values, and since they do not use significance tests wrong-way-round, this paper should be graded "good".

Thus in my alternative classification of these 11 papers, six receive a "good", one a "fair", and four a "marginal". It is highly likely that another economist would come up with different grades for some of the papers (and probably so would I if I would repeat the exercise), but he is most unlikely to come up with grades that are anywhere near Z-M's harsh results.

References

Acs, Zoltan, and David Audretsch. 1988. Innovation in Large and Small Firms: An Empirical Analysis. *American Economic Review* 78(Sept.): 678-690.

Alesina, Alberto, and Roberto Perotti. 1997. The Welfare State and Competitiveness. *American Economic Review* 87(Dec.): 921-939.

Altman, Douglas. 1980. Statistics and Ethics in Medical Research. *British Medical Journal* 281(Nov.): 1336-1338.

Altman, Morris. 2004. Introduction. *Journal of Socio-Economics* 33(Nov.): 523-525.

Angrist, Joshua, and William Evans. 1998. Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review* 88(June): 450-477.

Anonymous. 2012. Report on "Ziliak and McCloskey's Criticism of Significance Tests: A Damage Assessment". Referee report on this paper.

Artuç, Erhan, Shubham Chaudhuri, and John McLaren. 2010. Trade Shocks and Labor Adjustment: A Structural Empirical Approach. *American Economic Review* 100(3): 1008-1045.

Ayers, Ian, and Peter Siegelman. 1995. Race and Gender Discrimination in Bargaining for a New Car. *American Economic Review* 85(June): 304-321.

Bailey, Martha. 2010. Momma's Got the Pill: How Anthony Compstock and Griswold v. Connecticut Shaped U. S. Childbearing. *American Economic Review* 100(Mar.): 98-129.

Bardhan, Pranab, and Dilip Mookherjee. 2010. Determinants of Redistributive Policies: An Empirical Analysis of Land Reforms in West Bengal, India. *American Economic Review* 100(Sept.): 1572-1600.

Berg, Nathan. 2004. No Decision Classification: An Alternative to Testing for Statistical Significance. *Journal of Socio-Economics* 33(Nov.): 631-650.

Blanchard, Olivier. 1989. Traditional Interpretations of Macroeconomic Fluctuations. *American Economic Review* 79(Dec.): 1146-1164.

Blaug, Mark. 1980. *The Methodology of Economics*. New York: Cambridge University Press.

Bloom, David, and Christopher Cavanagh. 1986. An Analysis of the Selection of Arbitrators. *American Economic Review* 76(June): 408-422.

Borjas, George. 1987. Self-Selection and the Earnings of Immigrants. *American Economic Review* 77(Sept.): 531-553.

Borjas, George. 1995. Ethnicity, Neighborhoods and Human Capital Externalities. *American Economic Review* 85(June): 365-390.

Brainard, S. Lael. 1997. An Empirical Assessment of the Proximity-Concentration Trade-Off Between Multinational Sales and Trade. *American Economic Review* 87(Sept.): 520-544.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2012. Star Wars: The Empirics Strike Back. *Paris School of Economics Working Paper* No. 2012-29. Paris-Jourdan Sciences Economiques (Paris). **Link**

Cargill, Thomas. 2012. A Critical Assessment of Measures of Central Bank Independence. *Economic Inquiry*, forthcoming. **Link**

Carmichael, Jeffrey, and Peter Stebbing. 1983. Fisher's Paradox and the Theory of Interest. *American Economic Review* 73(Sept.): 619-630.

Chandra, Amitabh, Jonathan Gruber and Robin McKnight. 2010. Patient Cost Sharing and Hospitalization Offsets in the Elderly. *American Economic Review* 100(Mar.): 193-213.

Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens. *American Economic Review* 100(4): 1358-1398.

Cohen, Jacob. 1994. The Earth Is Round (p < 0.05). *American Psychologist* 49(Dec.): 997-1003.

Colander, David. 2001. *The Lost Art of Economics*. Cheltenham, UK: Edward Elgar.

Colander, David. 2011. Creating Humble Economists: A Code of Ethics for Economists. *Middlebury Economics Working Paper* 11-03. Department of Economics, Middlebury College (Middlebury, Vt.). **Link**

Conley, Timothy, and Christopher Udry. 2010. Learning About a New Technology: Pineapples in Ghana. *American Economic Review* 100(Mar.): 35-69.

Corrado, Carol A., Charles R. Hulten, and Daniel E. Sichel. 2006. Intangible Capital and Economic Growth. *NBER Working Paper* No. 11948. National Bureau of Economic Research (Cambridge, Mass.). **Link**

Dafny, Leemore. 2010. Are Health Insurance Markets Competitive? *American Economic Review* 100(Sept.): 1399-1431.

Darby, Michael. 1982. The Price of Oil and World Inflation and Recession. *American Economic Review* 72(Sept.): 738-751.

Dewald, William, Jerry Thursby, and Richard Anderson. 1986. Replication in Economics: The Journal of Money, Credit and Banking Project. *American Economic Review* 76(Sept.): 587-603.

Ellison, Glenn, Edward Glaeser, and William Kerr. 2010. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *American Economic Review* 100(June): 1195-1214.

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

Engsted, Tom. 2009. Statistical vs. Economic Significance in Economics and Econometrics: Further Comments on McCloskey and Ziliak. *Journal of Economic Methodology* 16(Dec.): 393-408.

Evans, David, and James Heckman. 1984. A Test for Subaddivity of the Cost Function with an Application to the Bell System. *American Economic Review* 74(Sept.): 615-623.

Feenstra, Robert. 1994. New Product Varieties and the Measurement of International Prices. *American Economic Review* 84(Mar.): 157-177.

Fidler, Fiona, Neil Thomason, Geoff Cumming, Sue Finch, and Joanna Leeman. 2004a. Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think: Statistical Reform Lessons from Medicine. *Psychological Science* 15(Feb.): 119-126.

Fidler, Fiona, Geoff Cumming, Mark Burgman, and Neil Thomason. 2004b. Statistical Reform in Medicine, Psychology and Ecology. *Journal of Socio-Economics* 33(Nov.): 615-630.

Forsythe, Robert, Forrest D. Nelson, George R. Neumann, and Jack Wright. 1992. Anatomy of an Experimental Political Stock Market. *American Economic Review* 82(5): 1142-1161.

Fowlie, Meredith. 2010. Emissions Trading, Electricity Restructuring, and Investment in Pollution Abatement. *American Economic Review* 100(June): 837-869.

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

Froyen, Richard, and Roger Waud. 1980. Further International Evidence on the Output-inflation Tradeoffs. *American Economic Review* 70(Mar.): 409-421.

Fuhrer, Jeffrey, and George Moore. 1995. Monetary Policy Trade-offs and the Correlation Between Nominal Interest Rates and Real Output. *American Economic Review* 85(Mar.): 219-239.

Gali, Jordi. 1999. Technology, Employment and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations? *American Economic Review* 89(Mar.): 249-271.

Garber, Peter. 1986. Nominal Contracts in a Bimetallic Standard. *American Economic Review* 76(Dec.): 1012-1030.

Gibbard, Allan, and Hal Varian. 1978. Economic Models. *Journal of Philosophy* 75(Nov.): 665-667.

Gigerenzer, Gerd. 2004. Mindless Statistics. *Journal of Socio-Economics* 33(Nov.): 587-606.

Gill, Jeff. 1999. The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* 52(Sept.): 647-674.

Ham, John, Jan Svejnar, and Katherine Terrell. 1998. Unemployment and the Social Safety Net During Transitions to a Market Economy: Evidence from the Czech and Slovak Republics. *American Economic Review* 88(Dec.): 1117-1142.

Haller, Heiko, and Stefan Krauss. 2002. Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online* 7(1).

Harlow, Lisa, Stanley Mulaik, and James Steiger. 1997. *What If There Were No Significance Tests?* Mahwah, N.J.: Lawrence Erlbaum Associates.

Harrison, Ann, and Jason Scorse. 2010. Multinationals and Anti-Sweatshop Activism. *American Economic Review* 100(Mar.): 247-273.

Hendricks, Kenneth, and Robert Porter. 1996. The Timing and Incidence of Exploratory Drilling on Offshore Wildcat Tracts. *American Economic Review* 86(June): 388-407.

Hoover, Kevin. 2011. The Role of Hypothesis Testing in the Molding of Econometric Models. Working paper.

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.

Hoover, Kevin, and Steven Sheffrin. 1992. Causality, Spending and Taxes: Sand in the Sandbox or Tax Collector for the Welfare State? *American Economic Review* 82(Mar.): 225-248.

Horowitz, Joel. 2004. Comments on "Size Matters." Journal of Socio-Economics 33(Nov.): 551-554.

Hubbard, Raymond, and S. Scott Armstrong. 2006. Why We Really Don't Know What Statistical Significance Means: Implications for Educators. *Journal of Marketing Education* 28(Aug.): 114-120.

Johnson, William, and Jonathan Skinner. 1986. Labor Supply and Marital Separation. *American Economic Review* 76(June): 455-469.

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

Joskow, Paul. 1987. Contract Duration and Relationship-Specific Investments: Empirical Evidence from Coal Markets. *American Economic Review* 77(Mar.): 168-185.

Keuzenkamp, Hugo, and Jan Magnus. 1995. On Tests and Significance in Econometrics. *Journal of Econometrics* 67(1): 5-24.

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

Kroszner, Randall, and Raghuram Rajan. 1994. Is the Glass-Steagall Act Justified? A Study of the U.S. Experience with Universal Banking Before 1933. *American Economic Review* 84(Sept.): 810-832.

LaLonde, Robert. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review* 76(Sept.): 604-620.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp. 2010. Is a Donor in Hand Better than Two in the Bush? Evidence from a Natural Field Experiment. *American Economic Review* 100(3): 958-983.

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

Lerner, Josh, and Ulrike Malmendier. 2010. Contractibility and the Design of Research Agreements. *American Economic Review* 100(Mar.): 214-246.

Leth-Petersen, Søren. 2010. Intertemporal Consumption and Credit Constraints: Does Total Expenditure Respond to an Exogenous Shock to Credit? *American Economic Review* 100(June): 1080-1103.

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

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

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

Mayer, Thomas. 2001. Misinterpreting A Failure to Disconfirm as a Confirmation. *University of California, Davis, Department of Economics Working Papers* 01-08. University of California, Davis. Link

Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah, and Aris Spanos. 2006. Severe Testing as a Basic Concept in a Neyman-Pearson's Philosophy of Induction. *British Journal for the Philosophy of Science* 57(2): 323-357.

McCloskey, D. N. 1985. *The Rhetoric of Economics*. Madison: University of Wisconsin Press.

McCloskey, D. N. 2008. Private communication.

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

McCullough, D. B., and H. D. Vinod. 1999. The Numerical Reliability of Econometric Software. *Journal of Economic Literature* 37(June): 633-665.

Mendelsohn, Robert, William Nordhaus, and Daigee Shaw. 1994. The Impact of Global Warming on Agriculture: A Ricardian Analysis. *American Economic Review* 84(Sept.): 753-771.

Mian, Atif, Amir Sufi, and Francesco Trebbi. 2010. The Political Economy of the U.S. Mortgage Default Crisis. *American Economic Review* 100(Dec.): 1967-1998.

Mishkin, Frederic. 1982. Does Anticipated Aggregate Demand Policy Matter: Further Econometric Results. *American Economic Review* 72(Sept.): 788-802.

Morrison, Denton, and Ramon Hankel. 1970. *The Significance Test Controversy: A Reader*. Chicago: Aldine.

Nickerson, Raymond. 2000. Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods* 5(2): 241-301.

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

Pashigian, Peter. 1988. Demand Uncertainty and Sales: A Study of Sales and Markdown Pricing. *American Economic Review* 78(Dec.): 936-953.

Pontiff, Jeffrey. 1997. Excess Volatility and Closed- End Funds. *American Economic Review* 87(Mar.): 155-169.

Pollard, P., and J. T. E. Richardson. 1987. On the Probability of Making Type I Errors. *Psychological Bulletin* 102(1): 159-163.

Reinsdorf, Marshall. 2004. Alternative Measures of Personal Saving. *Survey of Current Business* 84(Sept.): 17-27.

Robinson, Daniel, and Howard Wainer. 2002. On the Past and Future of Null Hypothesis Significance Testing. *Journal of Wildlife Management* 66(2): 263-271.

Romer, Christina. 1986. Is Stabilization of the Postwar Economy a Figment of the Data? Estimates Based on a New Measure of Fiscal Shocks. *American Economic Review* 76(June): 314-334.

Romer, Christina, and David Romer. 2010. The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shock. *American Economic Review* 100(June): 763-801.

Rudner, Richard. 1953. The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science* 20(Jan.): 1-6.

Sachs, Jeffrey. 1980. The Changing Cyclical Behavior of Wages and Prices, 1890-1976. *American Economic Review* 70(Mar.): 78-90.

Sauer, Raymond, and Keith Leffler. 1990. Did the Federal Trade Commission's Advertising Substantiation Program Promote More Credible Advertising? *American Economic Review* 80(Mar.): 191-203.

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**

Stang, Andreas, Charles Poole, and Oliver Kuss. 2010. The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research. *European Journal of Epidemology* 25(Mar.): 225-230.

Sterling, T. D., W. L. Rosenbaum, and J. J. Weinkam. 1995. Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *American Statistician* 49(Feb.): 108-112.

Stekler, H. O. 2007. Significance Tests Harm Progress in Forecasting: Comment. *International Journal of Forecasting* 23: 329-330.

Trejo, Stephen. 1991. The Effects of Overtime Pay Regulation on Worker Compensation. *American Economic Review* 81(Sept.): 719-740.

Wainer, Howard. 1999. One Cheer for Null Hypothesis Significance Testing. *Psychological Methods* 4(2): 212-213.

White, William. 1967. The Trustworthiness of "Reliable" Econometric Evidence. *Zeitschrift für Nationalökonomie* 27(Apr.): 19-38.

Wolff, Edward. 1991. Capital Formation and Productivity Convergence Over the Long Term. *American Economic Review* 81(June): 565-579.

Woodbury, Stephen, and Robert Spiegelman. 1987. Bonuses to Workers and Employers to Reduce Unemployment: A Randomized Trial in Illinois. *American Economic Review* 77(Sept.): 513-530.

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

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

Ziliak, Stephen T., and D. N. McCloskey. 2004. Size Matters: The Standard Error of Regressions in the *American Economic Review*. *Journal of Socio-Economics* 33(Nov.): 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 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. **Link**

Ziliak, Stephen T., 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.

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 e-mail address is tommayer@lmi.net.

Deirdre McCloskey and Stephen Ziliak's reply to this article
Go to Archive of Economics in Practice section
Go to September 2012 issue

Discuss this article at Journaltalk:
http://journaltalk.net/articles/5775