Mayer, Thomas

Working Paper

Ziliak and McClosky's criticisms of significance tests: A damage assessment

Working Paper, No. 12-6

Provided in Cooperation with:
University of California Davis, Department of Economics

Suggested Citation: Mayer, Thomas (2012) : Ziliak and McClosky's criticisms of significance tests: A damage assessment, Working Paper, No. 12-6, University of California, Department of Economics, Davis, CA

This Version is available at:
https://hdl.handle.net/10419/58359

UCDAVIS
DEPARTMENT OF ECONOMICS
Working Paper Series

Ziliak and McClosky's Criticisms of Significance Tests: A Damage Assessment

Thomas Mayer
University of California, Davis - Economics Department

April 20, 2012

Paper # 12-6

D. N. McCloskey and Stephen Ziliak have criticized economists and others for confounding statistical and substantive significance, and for committing the logical error of the transposed conditional. In doing so they sometimes misinterpret the function of significance tests. Nonetheless, economists sometimes make both of these errors – but not nearly as often as Ziliak and McCloskey claim. They also argue –incorrectly – that the existence of an effect, which is what significance tests are about, is not a scientific question. Their complaint that in testing significance economists often do not take the loss function into account is unfounded. But they are right in arguing that confidence intervals should be presented more frequently.

Department of Economics
One Shields Avenue
Davis, CA 95616
(530)752-0741

http://www.econ.ucdavis.edu/working_search.cfm

Ziliak and McClosky's Criticisms of

Significance Tests: A Damage Assessment

Thomas Mayer*

First Draft March 26, 2012

Abstract

D. N. McCloskey and Stephen Ziliak have criticized economists and others for confounding statistical and substantive significance, and for committing the logical error of the transposed conditional. In doing so they sometimes misinterpret the function of significance tests. Nonetheless, economists sometimes make both of these errors – but not nearly as often as Ziliak and McCloskey claim. They also argue –incorrectly – that the existence of an effect, which is what significance tests are about, is not a scientific question. Their complaint that in testing significance economists often do not take the loss function into account is unfounded. But they are right in arguing that confidence intervals should be presented more frequently.

Keywords: Significance tests, t's, confidence intervals, Zilliak, McCloskey, oomph

JEL Classifications: C12 , B4

* University of California, Davis. E-mail: Tommayer@lmi.net. I am greatly indebted to Kevin Hoover and Deidre McCloskey for helpful comments. An earlier version of this paper was presented at the 2010 Conference of Government Economists.

Ziliak and McClosky's Criticisms of Significance Tests: A Damage Assessment

"[I]f economists have natural constants, then the most well known is 0.05." Keuzenkamp and Magnus (1995, p.16)

Significance tests are standard operation 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. But they have many critics, particularly among psychologists who have done much more work on this topic than economists have. 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 these journals moved on their own to discourage significance tests, although with little lasting effect; significance tests still reign in psychology. In medical research, however, the critics appear to have had considerable influence. (See Fidler et al, 2004)[1]

Within economics, although significance tests have been occasionally criticized for many years (see for instance White,1967; Mayer 1980), these criticisms became prominent only in 1985 with the publication of D. N. McCloskely'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*.[2] There they charge that: "Statistical significance is not the same thing as scientific finding. R^2, t-statistics, *p*-values, F-tests, 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", [though they are] "very

[1] Eight years ago Fidler et al (2004, p. 616) reported that ecology has not yet advanced beyond the early stages of reform For a 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 (1993). Nickerson (2000) provides a comprehensive survey of this literature.

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

willing to concede some minor role to even mindless significance testing. Elsewhere they declare: "Significance testing as used has no theoretical justification." (2004, p. 527). So far the main criticisms of their work have come from Kevin Hoover and Mark Siegler (2008a), Tom Engsted (2008), Aris Spanos (2008), and from some papers in a symposium in the *Journal of Social Economics* in 2004 Their book 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 in economics journals. Few books by contemporary economists have stirred interest in so many fields.

I will evaluate the use of these tests only in economics, even though this is unfair to Ziliak and McCloskey 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, some less extreme versions of some of their claims are correct. In doing so I take a pragmatic approach, and look only at those errors in the application of significance tests that are likely to cause readers to draw substantially wrong conclusions, and disregard those that are unlikely to do so. As Gigerenzer (2004) has noted careless statements about significance tests abound. Since this is not a review article of Ziliak's and McCloskey's book I also leave aside some topics that this 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 disussion of their current use. For an evaluation that encompasses this and other items that I omit see Aris Spanos (2008).

Discussions of significance tests are not always clear about what they mean by this 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. Ziliak and McClosky use 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.[3] But, as Tom Engsted (2009) points out, economists generally do not aim at constructing models that are true in the sense that the only errors are sampling errors, but aim at models that are useful. And for that it does not matter if a significance test shows errors larger than what could be explained by the vagaries of sampling. He mentions three specific areas in which such models predominate, stochastic general equilibrium models, linear rational expectations models and models of asset pricing, such as Mehra and Prescott (1985). Ziliak and McClosky ignore most of this literature, perhaps because these models mainly use calibration and simulation tests instead of significance tests. Further, Spanos (2008) and Kramer (2011) argue, Ziliak and McCloskey do not treat the problem created by specification errors adequately.

I. Ziliak and McClosky's Criticisms

Ziliak and McCloskey 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 independent variables have on the dependent variable (which Ziliak and McCloskey call "oomph"), is vastly more important than statistical significance, or what is even worse, they confound the two. Second, they 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.

1. What Matters, Significance or Oomph?

[3] Thus Ziliak and McCloskey 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 it is not likely to be a random sample of the data 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 measurement errors t-values regain their meaning. They now tell us the likelihood that the measured significance of a coefficient is due to the unsystemic part of measurement errors.

There are several issues under this rubric. One is just 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 get it wrong and focus on statistical instead of substantive significance.

(i) What do Significance Tests Tell Us?

At least in some of their writings Ziliak and McCloskey 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, Ziliak and McCloskey imply, but do not state explicitly, we already know. This is illustrated by the following mental experiment they suggest (Ziliak and McCloskey, 2008, p. 23): "Suppose you want to help your mother lose weight and are considering two diet pills with identical prices and side effects." The first one, called "Oomph", will on average take off 20 pounds, but its variance is 10 pounds. The other pill, called "Precision" will on average take off only 5 pounds but has a probable error of only 1 pound. Which pill would you recommend?

This mental experiment is flawed, a point already noted by Hoover and Siegler (2008b, pp. 15-16). 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 random 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? But sample size alone does not tell her how much credence to give your means; variance also matter. 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 Ziliak and McCloskey's unwarranted assumption that we know the mean of the universe, rather than just a sample mean, 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 Ziliak and McCloskey'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." (Ziliak and McCloskey, 2008, p. 86, italics in original.) Yes, statistical significance is not

needed for a coefficient to have 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 Ziliak and McClosky'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 with the right sign, because the researchers mistakenly believe that what makes a variable important is its significance along with the right sign, and not its oomph.[4] 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. This becomes apparent in light of Deborah Mayo's (1996) plausible 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.

(ii) Is Existence a Scientific Question?

There is also a flaw in Ziliak and McCloskely's claim that significance tests only tell us whether an effect exists, and that this is a philosophical and not a scientific question.[5] But existence *is* a scientific question because it makes little sense for scientists to be concerned about the size of something that does not exist, and whose seeming appearance is merely the result of say, sampling error. It would be hard to obtain an NSF grant to measure the time phlogiston requires to achieve combustion. And we do observe natural scientists asking about existence.[6] Hoover and Siegler

[4] They 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, p. 70, first italics added, second in original.)

[5] Here Ziliak and McCloskey may be implicitly acknowledging that significance tests conern the difference between a coefficient obtained from the sample, and one (hypothetically) obtained from the universe.

[6] Thus Hoover and Siegler (2008a, p. 27) show that, contrary to Ziliak and McCloskey's claim, physical scientists also use significance tests. In their reply McCloskey and Ziliak (2008, pp. 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.

(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.[7] Currently (2012) there is 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 no oomph is needed: (a) the speed of 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.

Within economics Giffen goods provide a telling example. Economists are interested in knowing that 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 even when we are interested in oomph it does not always matters more than existence. Consider the hiring policy of a firm. Suppose the education variable shows a much greater oomph in a regression explaining the firm's employment decisions than the racial variable does. If the racial variable has the expected sign and is significant, you have confirmed the claim of racial discrimination. By contrast, suppose you find a substantial oomph for the racial variable, but its t is only 1.2. Then you do not have strong a case to take to court. Ziliak and McCloskey 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) call 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 a only a small oomph in a regression equation.

[7] Ziliak and McCloskey (2008, pp. 48-49), however, reject this interpretation of that test because statistical significance tests did not play a role in it. But isn't the issue here whether existence matters for science, and not whether significance tests are used to establish existence?

For causally oriented hypotheses, regression tests can be further divided into two classes. One, direct tests, are tests that ask about whether the variable has a large oomph, or if it explains much of the behavior of the dependent variable. If neither condition holds 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 lack adequate data for a direct test.

For instance, 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. Surely, few readers of his *A Theory of the Consumption Function* (1957) found the relative size of these propensities to consume interesting for their own sake, and therefore had 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.[8] Standing by itself each of these seven tests is only a soft test, because even if the permanent income theory is wrong there is presumably a 50 probability that farm families will have a lower marginal propensity to consume. But if all of these seven tests yield results in the direction predicted by the permanent income theory, then one to 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 larger than can reasonably be attributed 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.

[8] 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)

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 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 those who wonder why the standard theory predicts badly, regardless of the oomph of that variable. Similarly, it would be useful to have a variant of the Fisher relation that predicts movements of exchange rates better, even if 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 refute 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 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 induced 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 5 percent significance level the opposite was the case, Randall Kroszner and Raghuram Rajan (1994) refuted this hypothesis without having to discuss oomph.

All in all, none of the above denies that in many cases oomph is central. But it does mean that Ziliak and McCloskey (2008, p. 50) went much too far when they wrote that: "Existence is seldom if ever the issue." As the first column of Table 1 (see p. 32) based on a sample of 50 papers in the *American Economic Review* (AER) shows, oomph was neither required or very important in at least 8 (16 percent) and arguably in as many as 24 percent of the papers. While this result rejects Ziliak and McCloskey's strong claim, it still means that, at least in economics, the oomph of strategic coefficients usually deserves substantial emphasis.

(iii) How Often do Economists Confuse Statistical Significance with Substantive Significance?

Very often , say Ziliak and McCloskey. Having surveyed 369 full length AER articles from January 1980 to December 1999 that contain significance tests, they claim that: "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." (Ziliak and McCloskey, 2008, pp. 74, 80). Even though, as Engsted (2009) points out, there are many cases, such as reduced form usage of VAR's, where the magnitude of a particular coefficient is of little interest, Ziliak and McCloskey's results are still surprising, particularly since, plotting confidence intervals is the default setting of frequently used econometric software packages. (Hoover and Siegler, 2008a, p. 20.) How then did Ziliak and McCloskey obtain their dramatic results? They did so by giving each paper a numerical score depending on its performance on a set of nineteen 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. (Ziliak and McCloskey, 2008, pp. 72-73). Such a grading requires much subjectivity. What is "ambiguous" to one reader may be unambiguous to another. Moreover, suppose a paper uses "significant" in an ambiguous way in the" Introduction", but then in the "Conclusion" 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 in the "Conclusion", but does so at length elsewhere? It is therefore not surprising that Hoover and Siegler (2008a, p. 5) have criticized Ziliak and McCloskey's questions for frequently requiring subjective judgments, as well as for being a hodge-podge, containing some questions that are redundant, and some that indicate good practice, while others indicate bad practice. Moreover, as they pointed out, some questions duplicate others, which results in double counting, and hence an arbitrary weighting.[9] And they found the case studies of five AER papers that Ziliak and McCloskey provided entirely unconvincing. Similarly, Jeffrey Woodridge wrote: "I think … [Ziliak and McCloskey] 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 seems to violate Ziliak and McCloskey's standards, but which makes perfect sense in the broader context of the paper." (Woodridge, 2004, p. 578) Appendix A lists and evaluates each of Ziliak and McCloskey's questions.

[9] For Ziliak's and McCloskey's reply and Hoover and Siegler's rejoinder see McCloskey and Ziliak (2008) and Hoover and Ziliak (2008b).

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 to ask whether that is polluted by failing to distinguish statistical from substantive significance. Patrick 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 the papers' conclusions were affected by an inappropriate use of significance tests. In 23 out of the 118 papers (19 percent) in his sample significance tests were used inappropriately, but in only 8 of them (7 percent) did "it matter to the paper's main conclusion." (O'Brien, 2004, p. 568). This relatively low percentage, he suggested, may explain why Ziliak and McCloskey have had so little success in moving economists away from significance tests.

In my own attempted replication of Ziliak and McCloskey's results I followed O'Brien, as do Hoover and Siegler, by looking not at the specific wording of particular sentences but at a paper's overall Gestalt for a sample of fifty papers, thirty-five of them taken from Ziliak and McCloskey's sample (17 from the 1980's and 18 from the 1990's), and to update the sample, 15 papers from the 1991-2000 period. Specifically, I asked whether, a harried reader who is not watching for the particulars of significance testing would obtain the correct take-away-point with respect to significance and oomph.[10] Admittedly, this procedure requires subjective judgment, and different economists may evaluate some papers differently from the way I do.[11] But Ziliak and McCloskey's criteria are also subjective. So that readers can readily judge for themselves which procedure is preferable Appendix B provides summaries of the eleven papers in my sample that performed worst on Ziliak and McCloskey's criteria. They perform much better on mine.

The second column of Table 1 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 their discussion

dashes and footnotes indicate in-between cases. Table 1 treats as a "yes" cases where the authors do not discuss oomph in the text, but give the relevant coefficient 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

[10] I assume a harried reader because, given the great volume of reading material than descends upon us, it seems unlikely that most papers receive a painstaking reading. Further, I focus on the 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.

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

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 whether 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. And even for the same issue it may vary from time to time. When the bill rate is 5 percent it does not hinder the Fed as much if it takes 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.[12] The 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.

The results shown in the second column of the Table are in sharp contrast to Ziliak and McCloskey'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 often. Further (though this is not shown in the Table) there is no evidence anywhere of a confusion of significance and

magnitude. As column (3) shows, for none of the fifty papers is their take-away 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 take-away points *arguably* wrong.

Walter Kramer (2011, p. 462) 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 ('significance' used for both?) [or] [m]uch ado about statistically, but economically small coefficients or effects" In addition 28 percent of the papers discarded (wrongly, he believes) "economically significant

[12] 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 rarely is.

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 penalize a paper for discarding an economically large effect if it is statistically insignificant.

Neither O'Brien's results nor mine (and certainly not Kramer's) should be read as a wholesale rejection of Ziliak and McCloskey's contention. One reason is that in papers in some lesser-ranking journals – as well as in journals outside of economics -- the confusion of statistical with substantive significance could well be more common.[13] 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 means many errors. Hoover and Siegler (2008a) criticize Ziliak and McCloskey for making an already well-know point. 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 point.

II. 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? If so, Ziliak and McCloskey tell us, editors and referees are not doing their job. (See also Cohen 1994; Kramer, 2011; Mayer, 1980; 1993, 2001; Mayo, 1996.)

I. The Transposed Conditional.

Except in the below-discussed case of very large samples, treating a low t-value as evidence that some effect does *not* exist is to commit an error of logic, the error of the transposed conditional.[14] Jacob Cohen (1994, p. 24 italics added) illustrates this error by contrasting the following two syllogisms:

[13] Mayo and Spanos (2006, p. 341) call the confusion between statistical and substantive significance the "[p]erhaps most often heard and best known fallacy" in using significance tests.
[14] For a succinct discussion see Cohen (1994); for a comprehsive discussion of this problem in the psychological literature see Raymond Nickerson (2000) .

"If the null hypothesis is correct, then the datum (D) cannot occur.

It has, however, occurred

Therefore the null hypothesis is false."

And

"If H_0 is true this result (statistical significance) would *probably* not have occurred.

This result has occurred.

Then H_0 is probably not true"

The second syllogism, unlike the first, is invalid. 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 is true the test would not have disconfirmed it. (See Mayo and Spanos, 2006.) But tests of power are scarce in economics.

An intuitively simple way of making this point is to say that the results obtained from testing a hypothesis fall into one of three bins: a "confirmed" bin, a "disconfirmed" bin, and a "cannot tell" bin. If the coefficient of x has a t-value of, say 1.2, all this means is that the hypothesis that x aids in predicting y (or is causal to y) cannot be placed in the "confirmed" bin, but that does not mean that it belongs into the "disconfirmed" bin. If it did, it would be easy to disconfirm any hypothesis – just use a small enough sample.[15] A more sophisticated way is to follow Mayo and treat t-values as telling us not the probability that the hypothesis is correct, but the severity of the test that the hypothesis has passed. And if hypothesis H has failed a severe test, that does not imply that ~ H has passed such a test.

[15] 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 although in his data $x = 5$, and $y = 6$, he cannot reject at the 5 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?

Sample size plays a fundamental role here: 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? And herein lies the kernel of validity in the use of reverse significance tests. Suppose working with a sample of 100,000 we find that a variable that the hypothesis predicts to be positive is actually positive but it is 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.

2. Two Further Arguments against Wrong-Way-Round Significance Tests.

This 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 hard - to - pass tests fail to disconfirm it, that justifies tentatively accepting it, at least as a working hypothesis.[16] This may seem to imply that when a series of independent hard tests all fail to confirm a hypothesis because the coefficients have the right sign, they also have low t-values, we can treat the hypothesis as disconfirmed. But this principle is not applicable when in the various tests the relevant coefficient has the right sign. When philosophers speak of "failure to disconfirm" they mean failure to provide *any* evidence against the hypothesis. But 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. In Mayo's formulation it has failed a severe test, but it *has* passed a less severe test.

Moreover, treating a failure to confirm the hypothesis at the 5 percent level as though it were equivalent to disconfirming it is in sharp contrast to Ronald Fisher's intention. He intended the 5 percent significance level to be a substantial hurdle to scientific acceptance. But accepting the proposition that the hypothesis has been disconfirmed

[16] Fot this the tests need to be hard, and not in Mark Blaugh's 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 hypothesies is thererfore often only a poor, though very convenient, foil to the maintained hypothesis.

merely because the hypothesis cannot be confirmed at the 5 percent level, gives this alleged disconfirmation a much too easy pass into the corpus of science.[17]

Yet the distinction between having failed to confirm a hypothesis and having succeeded in disconfirming it is ignored in many cases. Ziliak and McCloskey (2008) cite numerous instances.[18]

3. Congruity Adjustments

My own survey in Table 1 distinguishes between significance tests that deal directly with a maintained substantive hypothesis and those that deal with whether a hypothesis needs 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 distribution when the hypothesis was initially formulated in terms of natural numbers. Other examples include breaks in values of coefficients, unit roots, serial correlation and heteroscedasticity.[19] The 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 adjustment only if one can reject at the 5 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 Occum's razor will not do because not adjusting for, say serial correlation, is only computationally and not philosophically simpler than not adjusting. What we need, but do not have, is an explicit loss function. In testing the maintained hypothesis the usual justification for the 5 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.

4. Frequency of Wrong-Way-Round Tests.

[17] Economists are not the only ones who use wrong-way-round significance tests. Nickerson (2000, p. 261) reports that this is frequently done in clinical studies in psychology.
[18] Sedimeier and Gigerenzer (1989) attribute the continuation of this error to a clumsy attempt to combine Fisher's approach to significance testing with Neyman-Pearsons'. Ziliak and McCloskey discuscuss at length the difference between the two, and acclaim the superiority of the latter.
[19] The same problem arises when deciding on the appropriate lag length by truncating when lagged coefficient become insignificant.

As column (4) of Table 1 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 5 percent level is equivalent to treating its negation as confirmed. And, as discussed in the notes to Table 1, there are five additional papers that do so if one applies a stricter standard than I did, giving a potential total of 25 percent. Previously (Mayer 2001), I looked at papers in 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 even worse in political science. There Jeff Gill (1999) found that in four leading journals significance tests were used wrong-way round in 40 to 51 percent of the relevant cases.

The last column of Table 1, which deals with congruity adjustment, shows 4 or at most 6 papers suffering from this error. But this underestimates - probably very substantially -- the actual number of cases because it includes only those in which authors mention their congruity adjustments. Presumably in many more papers authors tested for serial correlation etc., and decided not to make the adjustment because the adjustment was not required at the 5 percent level.

5. 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.) 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. It is along these lines that Jean Perrin confirmed Einstein's interpretation of Brownian motion. (See Mayo, 1996.)

III. The Loss Function

Although most economists think of significance tests as telling us something about an hypothesis, Ziliak and McCloskey view it in a Neyman - Pearsons framework, as telling us whether a certain course of action is justified. That requires a

loss function.[20] The main issues here are by whom and at what stage should the loss function be introduced, and also whether a loss function is relevant where we want to know what the data show just 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 policy-maker, who consults a loss function in deciding what action to take. This has several advantages. First, it places most (ideally all) value judgments outside of economics. Second, as Hoover and Siegler (2008a) point out, it avoids the problem that the 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 the econometrician's work can answer will arise only in the future?[21] For example, here are the titles of the first five papers listed in the reference section: "Innovations in Large and Small Firms, An Empirical Analysis"; "The Welfare State and Competitiveness"; "Children and their Parent's 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 Griswald vs. Connecticut Shaped U.S. Childbearing". How could their authors have determined the appropriate loss function? Third, effective policy making requires more than combining econometric results with value judgments, it also requires judgments about the probable gap between a proposed policy and a policy that is likely to emerge from the political process, as well as unintended effects that a policy may have on such factors as the public's trust in government -- the type of problems that David Colander (2001)

[20] Thus spake Ziliak and McCloskey: "[W]ithout a loss function a test of statistical significance is meaningless. … But every 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." (Ziliak and McCloskey, 2008, pp. 8-9, 15.)

[21] As Hoover and Siegler (2008, p. 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, the claim that scientific statements can avoid all value judgments has been criticized by post-modernists, and Rudner (1953) presents a criticism that focuses 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. Similarly, even if there is no watertight dichotomy between value judgments and positive judgments, for practical purposes it is often a useful distinction.

discusses under the rubric of "the art of political economy." A policymaker is probably better equipped to deal with such problems than is an econometrician.

The only workable solution is to designate the users of econometric estimates as the ones who should apply the appropriate loss function. Presumably Ziliak and McCloskey's objection to this conventional solution is that policy-makers or other users may fail to apply the appropriate loss function, and implicitly assume a symmetric one. This worry may well be justified. But the solution is to educate users of significance tests rather than imposing 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, p. 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 implications for action by perhaps changing the metaphysical 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 what loss function should apply, so an (implicit) symmetrical one is no more arbitary than any other.

IV. 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 Ziliak and McCloskey's criticism of reporting just t-values Hoover and Siegler (2008a) 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 confidence intervals, they can readily calculate any of the two other measures, so that it does not matter which one they are given. That is clearly correct. But it does not address the question of which measure is preferable. That readers can calculate, say confidence intervals, does not imply that they will do so. Harried readers, and perhaps even many careful readers, are likely to settle for the measure that is explicitly presented to them. On a pragmatic level It therefore does matter which one is.

1. 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 the American Psychological Association's Board of Scientific Affairs recommends that all estimates of size effects be accompanied by confidence intervals. (See Stang, Poole and Kuss, 2010, p. 229.) First, confidence intervals make it much harder to ignore oomph since they are stated in terms of oomph. (See McCloskey and Ziliak, 2008, p. 50; Anker, 2009, p. 135; Hubbard and Armstrong, 2002, p. 118.) Second, confidence intervals do not lend themselves as readily to wrong-way-round use as do t-values 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 import of the two main criticisms that Ziliak and McCloskey level against significance tests.

Third, 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 a paper based on a regression that does not take account of sampling error by presenting either t-values, *p*'s, F's or confidence intervals. 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, p. 20) point to their use in connection with impulse response functions. Moreover, As they point out, they are the default setting for VAR's 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 may be that for many papers confidence intervals would reveal the substantial imprecision of the paper's results, and thus their limited use for policy-making. -- Congress is not greatly helped if it is told that a stimulus somewhere between $100 billion and a $1 trillion is

needed. (Cf. Johnson, 1999, p. 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 2 percent and 80 percent of the subsequent decline in GDP would not satisfy our curiosity.

2. 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 easier for them to remember that a coefficient is significant than that it is, say 2.3. (See Berg, 2004.) But it also has some 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, 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 an appropriate t-value.[22] Moreover, excluding from science research in which the relevant t-value is less than 1.96 has bad consequences when combined with the reluctance of journals to publish "unscientific" papers. Suppose that there are 5 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. 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.

V. Conclusion

Both Ziliak and McCloskey's claim that most of the significance testing done by economists is invalid, as well as the counter-claim that, at least in economics, all is well with significance tests, should be rejected. A basic problem with Ziliak and McCloskey's claim is that, at least at times, they fail to realize that the purpose of a significance tests is not just to test the maintained hypothesis, but to test whether the researcher's reliance on a sample in place of the entire universe invalidates her results, and that significance tests can therefore be treated as a requirement of data hygiene.

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

Nonetheless, Ziliak and McCloskey are right in saying that one must guard against substituting statistical for substantive significance and that economists should pay more attention to substantive significance. They are also right in criticizing the wrong-way round use of significance tests. In the testing of maintained hypotheses this error is both severe enough and occurs frequently enough to present a serious problem. And it seems almost universal when deciding whether to adjust for serial correlation, heteroscedasticity, etc. Ziliak and McCloskey are also right that an explicit loss function needs to be used in deciding on the appropriate significance level when making policy decisions. But the econometrician is generally not able to do so, and must leave that up to the policy-maker. Their claim that confidence intervals are usually more informative than t-values or p's is also correct.

Moreover, in countering the mechanistic way in which significance tests are often used, and in introducing economists to the significance-test literature in other fields, they have rendered valuable services. Once the sharp edges of their extreme claims and of their scathing criticisms of their opponents have worn off, their work can be seen as a useful contribution, not only to applied economics, but also to applied work in many other fields. But there is a danger that their vehemence and overreach will tempt economists to dismiss their work.

Appendix A
Critique of Ziliak and McCloskey's Criteria for Evaluating Significance Tests

Note: All citations are from Ziliak and McCloskey (2008). In the original the first sentence of each paragraph is in italics.

1. *Criterion*: "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 hypothesis that the observed difference can reasonably be attributed merely to sampling error. If it cannot, it does not matter whether this is due to the difference (or the regression coefficient) being large relative to the variance, or to the sample being large. Ziliak and McCloskey justify their criterion by saying that we know that with a large enough sample every difference will be significant. But even if true, it would be irrelevant, because the function of significance test is to tell 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.

2. *Criterion:* "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." (p. 67) *Evaluation:* What matters is not the comprehensibility of "all" variables, but only the strategic one(s). Hence, this criterion is too tough. Second, authors need not specify the units and descriptive statistics if they are obvious. If not, Ziliak and McCloskey have a valid point, but one that has no discernable 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.

3. *Criterion:* "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." (p. 67) *Evaluation*: This is often a valid criterion when applied, not to every regression coefficient that is presented, but only to the strategic one(s). 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.

4. *Criterion:* "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." (p.68) Ziliak and McCloskey'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.

5. *Criterion:* "Are the coefficients carefully interpreted?" (p. 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, and never mind 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 careless author had never run a significance test and relied entirely on the large oomph of height.

6. *Criterion:* "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." (pp. 68-69) Ziliak and McCloskey give the example of reporting the results

of a significance test when the sample is the entire universe –a topic discussed in the text. *Evaluation:* Yes, reporting meaningless measures should be avoided, But what harm does it do, except to expose the author to well-deserved ridicule?

7. Criterion: "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 consideration of substantive significance." (p. 69) E*valuation:* 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.

8. *Criterion: "*Does the article mention the power of the test?" (p. 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.

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

10. *Criterion:* "Does the article refrain from 'asterisk econometrics', that is, ranking the coefficients according to the absolute size of their t-statistics?" (p.70) *Evaluation:* Presumably what Ziliak and McCloskey 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.

11. *Criterion:* "Does the article refrain from 'sign econometrics', that is noting the sign but not the size of the 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." (p. 69 italics in original) *Evaluation:* As shown in the text there is more scope for sign tests in economics than Ziliak and McCloskey admit. 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.

12. *Criterion*: "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?" (p. 70) *Evaluation:* As just mentioned there is some scope for "signs-only" significance tests. However, for many (probably most) papers Ziliak and McCloskey 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 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 rates 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 magnitudes only in a table.

13. *Criterion*: "Does the article discuss the scientific conversation within which a coefficient would be judged 'large' or'small' ?" (p. 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.

14. *Criterion:* 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." (p. 71) *Evaluation:* This is discussed in the text.

15. *Criterion:* "Later after the crescendo, does the article refrain from using statistical significance as the criterion for 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." (p. 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 sample error it is useful regardless of what prompted it.

16. *Criterion:* "[I]s statistical significance portrayed as decisive, a conversation stopper, conveying a sense of ending?" (p. 72) *Evaluation:* Ziliak and McCloskey treat a positive answer as an error. Once again, in some situations it is not. Or consider the following: Someone wrote an article relating the growth rates of countries to the level of their corporate income rates, 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 conclusive evidence when discussing the appropriate level of corporate income taxes, are you making a mistake?

17. *Criterion*: "Does the article ever use an independent simulation – as against a use of the regression coefficient as inputs into further calculations – to determine whether the coefficients are reasonable?" (p. 72) E*valuation:* Such simulations may be useful in some, perhaps many, cases, but should every failure to do so count as a fault?

18. *Criterion:* "In the concluding section is statistical significance separated from policy, economic or scientific significance? In medicine and epidemiologist and especially in psychology the conclusions are often sizeless summaries of significance tests reported earlier in the article. Significance this, significance that. In economics too." (p. 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.

19. *Criterion:* "Does the article use the word *significant* unambiguously?" (p.73) *Evaluation:* Yes, ambiguity is bad. But that does not mean that it uses significance tests in a wrong way.

In summary: Although any attempt to fit the results of this evaluation o Ziliak and McCloskey'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 irrelevant because little damage results from not meeting it. However this unfavorable 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 allows different criteria to 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 done in the text.

Appendix B
Reappraising Eleven Papers that Ziliak and McCloskey Rank "Poor" or "Very Poor" with respect to their use of Significance Tests.

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

three are ones that Ziliak and McCloskey classify as "very poor", and the others are papers they classify as "poor".

1. Brainard (1997) "An Empirical Assessment of the Proximity-Concentration Trade-Off between Multinational Sales and Trade."

Lael Brainard evaluates the proximity – concentration hypothesis which 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 Ziliak and McCloskey's grade of "very poor" seems unjustified, and a grade of "fair" seems more appropriate.

2. 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, Stephen 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 a 3-9 percentage points lower probability of being paid an overtime premium" (Trejo, 1991, p 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" (Trejo, 1991, p. 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" (Trejo, 1991, p. 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 other

years are always less than a third of the value predicted by" the rival theory. (Trejo, 1991, p. 737.) All in all, the treatment of significance tests in this paper deserves a grade of "good".

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

The Glass-Steagal Act (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. Krozner 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. And when measured by the dollar volume of defaults, rather than by frequency of default, the difference in favor of commercial banks and their affiliates is even greater. Krozner 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 is to challenge the then widely accepted hypothesis that allowing commercial banks and their affiliates to underwrite securities creates a conflict of interest. And for that it suffices to show that the relevant differences have the wrong sign. This paper therefore deserves 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 an wrong-way-round significance test. But this point is not important for the paper, and I therefore classify it as "fair".

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 VAR's and a structural model. It presents its results mainly by charts of autocorrelation functions and autocovarance 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 remained statistically significant and the residuals appear to be uncorrelated." (Fuhrer and Moore (1995, p. 221) 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 Peterman 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," (p. 309.) Throughout the text they cite numerous oomphs, for example, "[f]or black males the final mark-up 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." (1995, p.313.) Moreover, Ayres and Siegelman sometimes cite oomph even when the t statistics are far from significant. But at one, not very important point, they do use significance tests wrong-way-round, writing: " None of these coefficients … indicating that the seller's race did not influence the bargaining outcome." (p.315). For this reason 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 behind, the faster will be its rate of catch-up. An alternative hypothesis is that convergence in labor productivities is due to convergence in factor intensities, and a third hypothesis is 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 significance tests the wrong-way-round, and that Ziliak and McCloskey are therefore justified in grading the paper as "poor". But, given the need to limit somehow the huge number of variables that one might potentially include (and *perhaps* the frequency with which insignificant variables are dropped in economics), it seems that "marginal" is a more appropriate grade than the "poor" that Ziliak and McCloskey give it.

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

When wildcat drillers have successfully bid 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. So each leaseholder 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" and not the "poor" that Ziliak and McCloskey give it.

10. Jordi Gali (1999) "Technology, Employment and the Business Cycle."

Gali presents here 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 it requires a technological shock to 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 upon 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 at least a "fair".

11. Robert Mendelson, 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 Mendelson, 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 papers should be graded “good

Summary.

Thus in my alternative classification of these 11 papers 5 receive a “good”, 3 a “fair”, and 3 a “marginal”. It is highly likely that someone else who also classifies them by their Gestalt would come up with a somewhat different result, but he is most unlikely to come up with one that resembles Ziliak and McCloskey’s.

References

Acs, Zoltan and Audretsch, David (1988) "Innovation in Large and Small Firms: An Empirical Analysis," *American Economic Review*, September, 78, 678-90.

Alesina, Alberto and Perotti, Roberto (1997) "The Welfare State and Competitiveness", *American Economic Review* December, 87, pp.921-39.

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

Ayers, Ian and Siegelman, Peter (1995) "Race and Gender Discrimination in Bargaining for a New Car," *American Economic Review* June, 85, 304-21.

Bailey, Martha, (2010) "Momma's got the Pill: How Anthony Comstock and *Griswald v. Connecticut* Shaped U. S. Childbearing," *American Economic Review*, March 100, 98-129.

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

Benhabib, Jess and Jovanovic, Boyan (1991) "Externalities and Growth Accounting", *American Economic Review*, March, 81, 82-113.

Berg, Nathan (2004) "No-decision Classification: an Alternativeto testing for Statistical Significance," Journal of Socio-Economics, November, 33, 631-50.

Blanchard, Olivier (1989) "Traditional Interpretations of Macroeconomic Fluctuations," *American Economic Review*, December, 79, 1146-64.

Bloom, David and Cavanagh, Christopher (1986) "An Analysis of the Selection of Arbitrators*," American Economic Review*, June, 76, 408-22.

Borjas, George (1987) "Self-Selection and the Earnings of Immigrants," *American Economic Review*, September, 77, 531-53.

Borjas, George (1995) "Ethneticity, Neighborhoods and Human Capital Externalities," *American Economic Review*, June, 85, 365-90.

Brainard, S. Lael (1997) "An Assessment of the Proximity-Concentration Trade-Off between Multinational Sales and Trade," *American Economic Revie*w, September, 87, pp. 520-544.

Carmichael, Jeffrey and Stebbing, Peter (1983) "Fisher's Paradox and the Theory of Interest," *American Economic Review*, September, 73, 619-30.

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

Chen, Yan, et al (2010) "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens," *American Economic Review*, September, 100, 1358-98.

Cohen, Jacob (1994) "The Earth is Round ($p < 0.05$)" *American Psychologist*, December, 49, 993-1003.

Colander, David (2001) *The Lost Art of Economics*, Cheltenham, England, Edward Elgar

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

Dafney, Leemore (2010) "Are Health Insurance Markets Competitive?" *American Economic Review*, September, 100, 1399-1431.

Darby, Michael (1982) "The Price of Oil and World inflation and Recession," *American Economic Review*, September, 72, 738-51.

Ehrarn, Artuc, et al (2010) "Trade Shocks and Labor Adjustment," *American Economic Review*, June, 100, 1008-45

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

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

Evans, David and Heckman, James (1984) A Test for Subaddivity of the Cost Function with an Application to the Bell System*," American Economic Review*. September, 74, 6!5-23.

Falk, Ruma and Greenbaum, Charles (1995) "Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception," *Theory and Psychology*, 5, 1, 993-1003.

Feenstra, Robert (1994) "New Product Varieties and the Measurement of International Prices," (1994) *American Economic Review*, March, 84, 157- 78.

Fidler, Fiona, et al (2004) "Editors can Lead Researchers to Confidence Intervals, but Can't Make them Think: Statistical Reform Lessons from Medicine," *Psychological Science*, Feb, 15, 119-26.

Forsythe, Robert, et al. (1992) "Anatomy of an Experimental Political Stock Market," *American Economic Review*, December, 82, pp. 1142-61.

Fowley, Meredith (2010) "Emission Trading, Electricity Restructuring, and Investment in Pollution Abatement," *American Economic Review*, June, 100 837-69.

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

Froyen, Richard, and Waud, Roger (1980) " Further International Evidence on the Output-inflation Tradeoffs*," American Economic Review*, March, 70, 409-21.

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

Gali, Jordi (1999) "Technology, Employment and the Business Cycle: Do Technology Shocks explain Aggregate Fluctuations?" *American Economic Review*, March, 89, pp. 249-71.

Garber, Peter (1986) "Nominal Contracts in a Bimetallic Standard," *American Economic Review*. December, 75, 1012-30.

Gibbard, Allan and Varian, Hal (1978) "Economic Models," Journal of Philosophy, 75, November, 665-67.

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

Gill, Jeff (1999) "The Insignificance of Null Hypothesis Significance Testing," *Political Research Quarterly*, September, 52, 647-74.

Ham, John, Svejnar, John and Terrell, Katherine (1998) "Unemployment and the Social Safety Net during Transitions to a Market Economy: Evidence from the Czech and Slovak Republics," *American Economic Review,* December, 88, pp. 1117-41.

Harlow, Lisa, Mulaik, Stanley and Steiger, James (1997) *What if there were no Significance Tests?* Mahwah, N.J., Lawrence Erlbaum Associates.

Harrison, Ann and Scorse, Jason (2010) "Multinationals and Anti-Sweatshop Activism," *American Economic Review,* March 100, 247-73.

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

Hoover, Kevin (2011) "The Role of Hypothesis Testing in the Molding of Econometric Models," unpublished manuscript.

Hoover, Kevin and Siegler, Mark (2008a) "Sound and Fury: McCloskey and Significance Testing in Economics," *Journal of Economic Methodology*, March, 15, pp.39-56.

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

Hoover, Kevin and Sheffrin, Steven (1992) "Causality, Spending and Taxes: Sand in the Sandbox or Tax Collector for the Welfare State?" *American Economic Review***,** 82, March, 225-48.

Hubbard, Raymond and Armstrong, S. Scott (2006) "Why We Really Don't Know what Statistical Significance Means: implications for Educators," *Journal of Marketing Education*, 28 114- 20.

Johnson, William and Skinner, Jonathan (1986) "Labor Supply and Marital Separation," *American Economic Review,* June, 76, 455-69.

Johnson, Douglas (1999) "The Insignificance of Statistical Significance Testing" (1999) *Journal of Wildlife Management*, July, 63, 763-72.

Joskow, Paul (1987) " Contract Duration and Relationship-Specific Investments: Empirical Evidence from Coal Markets," *American Economic Review* March, 77, 168-85.

Keuzenkamp, Hugo and Magnus, Jan (1995) "On Tests and Significance in Econometrics, *Journal of Econometrics*,, 67:1, 103-28.

Kramer, Walter (2011) "The Cult of Statistical Significance,"RatSWD Working Paper 176.

Kroszner, Randall and Rajan, Raghuram (1994) "Is the Glass-Steagall Act Justified? A Study of the U.S. Experience with Universal Banking before 1933" *American Economic Review*, September 84, pp. 810-32.

LaLonde, Robert (1986) "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, September 76, 604-20.

Landrey, Craig, et al (2010) "Is a Donor in Hand Better than Two in the Bush?" Evidence from a Natural Field Experiment", *American Economic Review*, June, 100, 958-984.

Lerner, Josh and Malmendier, Ulrike (2010) "Contractibility and Design of Research Agreements, *American Economic Review,* March, 100, 214-46.

Leth-Petersen, Soren (2010) "Intertemporal Consumption and Credit Constraints: Does Total Expenditure Respond to an Exogenous Shock to Credit?" *American Economic Rev*iew, June 100, 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*, April, 18, 165-78.

Mayer, Thomas (1993) *Truth versus Precision* Aldershot, England, Edward Elgar

Mayer, Thomas (2001) "Misinterpreting A Failure to Disconfirm as a Confirmation," Economics Department, University of California, Davis, working paper 01-08.

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

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

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

McCloskey, D. N. (2008) Private Communication.

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

Mehra, Rajnish and Prescott, Edward (1985) "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, March, 15, 145-61

Mendelsohn, Robert, Nordhaus, William and Shaw, Daigee (1994) "The Impact of Global Warming on Agriculture: A Ricardian Analysis," *American Economic Review*, September, 84, 753-72.

Mian, Atif, Sufi, Amir and Trebbi, Francesco (2010) "The Political Economy of the U.S. Mortgage Default Crisis," *American Economic Review*, December, 100, 1967-98.

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

Morrison Denton and Hankel, Ramon (1970) *The Significance Test Controversy: A Read*er, Chicago, Aldine

.

Nickerson ,Raymond (2000) " Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy", *Psychological Methods*, 5:2, 21-301

O'Brien, Anthony (2004) "Why is the Standard Error of Regressions so Low Using Historical Data?" *Journal of Socio-Economics*, November, 33, 565-70.

Pashigian, Peter (1988) "Demand Uncertainty and Sales: A Study of Sales and Markdown Pricing*"* *American Economic Review*, December, 78, 936-53.

Pontiff, Jeffrey (1997), "Excess Volatility and Closed- End Funds," *American Economic Review*, March, 87, 155-69.

Porter, Theodore (2008) "Signifying Little," *Science*, June, 320, 1292.

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,* June, 76, 314-334.

Romer Christina and Romer, David (2010) "The Macroeconomic Effects of Tax Changes*," American Economic Review*. June, 100, 763-801.
Rudner, Richard (1953) "The Scientist Qua Scientist Makes Value Judgments," *Philosophy of Science*, January, 20, 1-6.

Sachs, Jeffrey (1980) "The Changing Cyclical Behavior of Wages and Prices, 1890-1976," *American Economic Review*, March, 70, 78-89.

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

Sedimeier, Peter and Gigerenzer, Gerd (1989) "Do Studies of Statistical Power have an Effect on the Power of Studies?" (1989) *Psychological Bulletin*, 105:2, 309-16

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 Live*s." *Erasmus Journal for Philosophy. and Economics,* Autumn 1, 154-64.

Stang, Andreas, Poole, Charles and Kuss, Oliver (2010) "The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research," *European Journal of Epidemology*," March, 25, 225-30.

Trejo, Stephen (1991) "The Effects of Overtime Pay Regulation on Worker Compensation," *American Economic Review*, September, 81, pp. 719-40.

Wainer, Howard (1999) "One Cheer for Null Hypothesis Significance Testing," *Psychological Methods, 4:2*, 212-23.

White, William (1967) "The Trustworthiness of 'Reliable' Econometric Evidence," *Zeitschrift fur Nationaleconomie,* April, 27, 19-38.

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

Woodbury, Stephen and Spiegelman, Robert (1987) "Bonuses to Workers and Employers to Reduce Unemployment: A Randomized Trial in Illinois" *American Economic Review*, September, 77, 513-30.

Woodridge, Jeffrey (2004) "Statistical Significance is Okay, Too: "Comments on Size Matters," *Journal of Socio-Economics*, November 33, 577-79.

Ziliak, Stephen (2011) http://blogs.wsj.com/numbersguy/a-statistical-test-gets-its-closeup-1050/?blog_id=168&post_id=1050.

Ziliak, Stephen and McCloskey D. N. 2004) Size Matters: The Standard Error of Regressions in the *American Economic Review*," *Journal of Socio-Economics*, November, 33, 527-46.

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

TABLE 1. USES AND MISUSES OF SIGNIFICANCE TESTS IN 50 A.E.R. PAPERS

	(1)	(2)	(3)	(4)	(5)
	Oomph required or very important for topic	Paper oomph	provides: correct take-away point with respect to oomph	Significance tests used wrong-way round for: testing maintained hypothesis	congruity adjust-ments[a]
Papers:					
1980's:					
Acs & Audretsch (1986)	Yes	Yes	Yes	Yes	No
Blanchard (1989)	Yes	Yes	Yes	No	--[b]
Bloom & Cavanagh (1989)	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 (1977)	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
Hendircks & Porter (1996)	Yes	Yes	Yes	No	No
Hoover & Sheffrin (1992)	No	Yes	Irrelevant	Yes	No
Krozner & 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
Woolf (1991)	Yes	Yes	--[n]	No	No
2010:					
Bailey (2010)	Yes	Yes	Yes	Yes	No
Bardhan & Mookerjee (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
Dafney (2010)	No[o]	Yes	Yes	No	No
Ehrarn (2010)	Yes	Yes	Yes	No	No
Ellison et al (2010)	Yes	Yes	Yes	Yes	No
Fowley (2010)	Yes	--[p]	Yes	No	No
Harrison & Scores (2010)	Yes	Yes	Yes	No	No
Landrey et al (2010)	Yes	Yes	Yes	No	No
Lerner & Malmendier (2010)	Yes	Yes	Yes	--[q]	No
Leth-Peterson (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. … [The results] … must be seen as dependent on an a priori assumption on the time series properties of the series." (Blanchard, 1989, p 1151) I believe that this absolved 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” since the paper 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.