

# **Sound and Fury: McCloskey and Significance Testing in Economics**

Kevin D. Hoover  
Department of Economics  
University of California  
1 Shields Avenue  
Davis, California 95616

Tel. (530) 752-2129  
E-mail [kdhoover@ucdavis.edu](mailto:kdhoover@ucdavis.edu)

Mark V. Siegler  
Department of Economics  
California State University, Sacramento  
6000 J. Street  
Sacramento, California 95819-6082

Tel. (916) 278-7079  
E-mail [msiegler@csus.edu](mailto:msiegler@csus.edu)

First Draft, 31 August 2005

We thank Deirdre McCloskey and Stephen Ziliak for providing us with the individual scores from their 1980s survey and for the individual scores broken down by question for the 1990s survey. We thank Ryan Brady for research assistance and Paul Teller and the participants in the UC Davis Macroeconomics Workshop for valuable comments.

**Abstract**  
**of**  
**Sound and Fury:**  
**McCloskey and Significance Testing in Economics**

For about twenty years, Deidre McCloskey has campaigned to convince the economics profession that it is hopelessly confused about statistical significance. The argument has two elements: 1) many practices associated with significance testing are bad science; 2) most economists routinely employ these bad practices. McCloskey's charges are analyzed and rejected. That statistical significance is not economic significance is a banal and uncontroversial claim, and there is no convincing evidence that economists *systematically* mistake the two. Other elements of McCloskey's analysis of statistical significance are shown to be ill-founded, and her criticisms of practices of economists are found to be based in inaccurate readings and tendentious interpretations of their work.

JEL Codes: C10, C12, B41

## 1. The Sin and the Sinners

For twenty years, since the publication of the first edition of *The Rhetoric of Economics* (1985a), Deirdre (*né* Donald) N. McCloskey has campaigned tirelessly to convince the economics profession that it is deeply confused about statistical significance.<sup>1</sup> With the zeal of a radio preacher, McCloskey (2002) declares the current practice of significance testing to be one of the two deadly "secret sins of economics":

The progress of science has been seriously damaged. You can't believe anything that comes out of the Two Sins [tests of statistical significance and qualitative theorems]. Not a word. It is all nonsense, which future generations of economists are going to have to do all over again. Most of what appears in the best journals of economics is unscientific rubbish. [p. 55]

...

Until economics stops believing . . . that an intellectual free lunch is to be gotten from . . . statistical significance . . . our understanding of the economic world will continue to be crippled by the spreading, ramifying, hideous sins. [p. 57]

As well as contributing to a debate internal to the profession, McCloskey engaged a wider audience. She has decried the sins of economics in the pages of *Scientific American* (McCloskey 1995a, b) and in a contribution to a series of tracts aimed at "anthropology . . . other academic disciplines, the arts, and the contemporary world" (McCloskey 2002, endpaper). Her attacks on applied economists have been widely reported *inter alia* in the *Economist* (2004).

In perhaps the most influential of McCloskey's many tracts, she and her coauthor characterize the main point as "a difference can be permanent . . . without being 'significant' in other senses . . . [a]nd . . . significant for science or policy and yet be

---

<sup>1</sup> Other writings on this topic include the second edition of *The Rhetoric* (McCloskey 1998) as well as McCloskey (1985b, 1992, 1994, 1995a, b, 1997, 1999, 2002, 2005), McCloskey and J. Richard Zecher (1984), McCloskey and Stephen T. Ziliak (1996), and Ziliak and McCloskey (2004a, b).

insignificant statistically . . . [McCloskey and Ziliak 1996, p. 97]. To avoid any misapprehension, let us declare at the outset that we accept the main point without qualification: a parameter or other estimated quantity may be statistically significant and, yet, economically unimportant or it may be economically important and statistically insignificant. Our point is the simple one that, while the economic significance of the coefficient does not depend on the statistical significance, our certainty about the accuracy of the measurement surely does.

But McCloskey's charges go beyond this truism. On the one hand, she charges that significance testing is mired in sin: many practices associated with significance testing are bad science. On the other hand, she charges that economists are, by and large, sinners – not only mistaking statistical significance for economic significance, but routinely committing the full range of sins associated with significance testing. The observation that statistical significance is not economic significance is banal and uncontroversial. We have been unable to locate anywhere in McCloskey's voluminous writings on this point a citation to even a single economist who defends the contrary notion that statistical significance demonstrates economic importance. The assertion that the mistake is commonplace is used to add buoyancy to an otherwise unsustainable bill of particulars against common statistical methods and the economists who use them.

McCloskey offers three principal charges against significance testing. First, as the title of Ziliak and McCloskey (2004a) puts it, "size matters." A coefficient that is estimated to have economically large size, *even if it is statistically insignificant*, cannot properly be neglected.

Second, McCloskey adopts a Neyman-Pearson statistical framework without qualification. Applied economists, she argues, are failures as scientists since they do not specify precisely the hypotheses that they regard as alternative to their null hypothesis and that they do not specify a loss function:

No test of significance that does not examine the loss function is useful as science . . . Thus unit root tests that rely on statistical significance are not science. Neither are tests of the efficiency of financial markets that rely on statistical instead of financial significance. Though to a child they look like science, with all that really hard math, no science is being done in these and 96 percent of the best empirical economics. [McCloskey 1999, p. 361]

It is not just that economists do not examine loss functions, McCloskey charges that they generally ignore type II error and the power of tests. Furthermore, the fixation on 5 percent (or other conventional test sizes) is a sign of not taking the trade off between size and power seriously.

Third, McCloskey argues that even the Neyman-Pearson framework has only a limited application in economics. It is, she believes, appropriate only when "sampling error is the scientific issue (which it is commonly not . . .)" (McCloskey 1999, p. 361). In general, she leaves the strong impression that tests of statistical significance have next-to-no place in economics for a variety of reasons. Tests are appropriate only when the data are a proper sample and not when they constitute the whole population. Yet, in most cases, especially when time series are involved, McCloskey maintains that the economist deals with a population or, worse, a sample of convenience (McCloskey 1985a, pp. 161, 167; 1985b, p. 203; McCloskey and Ziliak 1996, p. 112). Acknowledging McCloskey as the source of their maintained hypothesis, Hugo A. Keuzenkamp and Jan R. Magnus (1995, pp. 20-21) famously challenged economists to produce a clear-cut example of a case in which a significance test has ever been decisive with respect to an important

economic hypothesis (McCloskey and Ziliak 1996, pp. 111-112; and Ziliak and McCloskey 2004a, p. 543, cf. Lawrence H. Summers 1991).

McCloskey believes that, because of its use of significance tests, economics has become, to use Richard P. Feynman's (1985, pp. 308-317) analogy, a "cargo cult science." Anthropologists are said to have observed after World War II that certain Pacific islanders built straw radio huts and other replicas of military facilities in the hope that, by mimicking the forms of military activity, the aircraft would return with food, drink, and other modern goods as they had during the war. McCloskey (2002, pp. 55-56) means to echo Feynman's criticism of pseudo-sciences, implying that in using significance tests economists mimic the outward forms, rather than the substance, of science.

To demonstrate not only that significance testing is a methodological sin but also that economists are willing sinners, McCloskey charges that the training of economists in statistics and econometrics is negligent for not stressing the distinction between economic and statistical significance sufficiently. The principal evidence offered with respect to training is an analysis of statistics and econometrics textbooks. To demonstrate that the practices of applied economists betray a deep confusion about statistical significance, McCloskey relies on two surveys of articles from the *American Economic Review* – one for the 1980s (McCloskey and Ziliak 1996) and one for the 1990s (Ziliak and McCloskey 2004a). The surveys consist of nineteen questions (see Table 1) scored so that "yes" represents good practice and "no" bad practice.

We reiterate that we accept that statistical significance is not economic significance. No doubt there are cases of people mistaking one for the other. Yet, we

doubt that the problem is anywhere near as widespread as McCloskey asserts. We shall demonstrate, first, that McCloskey's broader analysis of significance testing cannot withstand scrutiny. Significance tests, properly used, are a tool for the assessment of signal strength and not measures of economic significance. By and large, economists use them this way. So, second, McCloskey's evidence that economists routinely confuse economic and statistical significance or typically engage in unsupportable statistical practices is unconvincing.

## **2. Statistical Practice: Good or Bad?**

### 2.1 THE LOGIC OF SIGNIFICANCE TESTS

At the risk of laboring the familiar, we set the stage with a review the logic of significance tests. The test of statistical significance has a venerable history (see Stephen M. Stigler 1986, 1999). Most applications fall under two related types. The first type asks whether two sample moments could have been drawn from populations with the same distribution. Francis Ysidro Edgeworth (1885, pp. 187-188) provides an early example:

In order to detect whether the difference between two proposed Means is or is not accidental, form the probability-curve under which the said difference, supposing it were accidental, would range. Consider whether the difference between the observed Means exceeds two or three times the *modulus* of that curve. If it does, the difference is not accidental. For example, in order to determine whether the observed difference between the mean stature of 2,315 criminals and the mean stature of 8,585 British adult males belonging to the general population is significant, we form the curve according to which the difference between the mean of a random selection of 2,315 and 8,585 mean fluctuates. And we shall find that the observed difference between the proposed Means, namely about 2 (inches) far exceeds thrice the modulus of that curve, namely 0.2. The difference therefore "comes by cause."

Edgeworth's strategy is exactly the same as used with a modern significance test. His "modulus" is just a rescaling of the standard deviation:  $\sqrt{2} \times$  the standard deviation (see Stigler (1986, pp. 310-311). Ziliak and McCloskey advise "the profession [to] adopt the standards set forth 120 years ago by Edgeworth . . .," but those standards are simply the ones in common use today, only more stringent (cf. Joel L. Horowitz 2004, p. 552). Stigler remarks that twice the modulus, Edgeworth's threshold for statistical significance corresponds to a two-side test with a size of 0.005, "a rather exacting test."<sup>2</sup>

Edgeworth's testing strategy is the same one used today when the question is whether two distributions are the same: on the assumption that the data conform to a particular probability distribution (such as the normal) or can through some transformation be made to do so, compute the distribution under the null hypothesis that the moments are the same and reject the null if the actual difference falls in the tail of the distribution as determined by a critical value. The critical value, typically, but not

---

<sup>2</sup> Ziliak and McCloskey (2004a, pp. 530-531) completely misread Edgeworth, when they attempt to enlist him as an ally. It is Jevons who takes the position that "size matters," and Edgeworth who argues that we must attend to statistical significance. Yes, Edgeworth distinguishes between economic and statistical significance, but Ziliak and McCloskey miss his point when they assert that he corrects Jevons for ignoring an economically, as opposed to a statistically, significant difference of 3 or 4 percent in the volume of commercial bills in different quarters. Edgeworth (1885, p. 208) writes: "Professor Jevons must be understood to mean that such a difference may for practical purposes be neglected. But for the purposes of science, the discovery of a difference in condition, a difference of 3 per cent. and much less may well be important." Edgeworth did not dispute Jevons's judgment that 3 to 4 percent is economically small or criticize him for it. Rather he argued that science may care about differences that are not of practical (i.e., economic) importance. To underwrite that view, he conducted a test of *statistical* significance: after correcting for a secular increase in the volume of bills, he found that the means of the first and second quarters differed by an amount equal to about 0.8 times the modulus ( $0.8 = 2.5 / \sqrt{10.5}$ ) or, in modern terminology, by 1.1 standard deviations. Edgeworth concluded: "There is, therefore, 'no great difference,' as Professor Jevons says; still a slight indication of a real law – enough to require the continuation of the inquiry, if the subject repaid the trouble." Far from contradicting Jevons on the matter of economic importance, Edgeworth found that the differences were not statistically significant by his usual standard of two to three times the modulus; nevertheless, the result, he believed, was significant enough that it might encourage economic scientists to investigate further, even if it would not repay the trouble of a practical man. His caveat about a slight indication of a real law reflects the intuition that, as we might now put it, a *p*-value of 0.21 could fall into the acceptance region for a researcher who placed a high value on detecting faint signals. (Edgeworth (1885, p. 201) provides another example of an economically small difference that should not be neglected scientifically because it is statistically significant.)



always, chosen as 5 percent, determines the probability of type I error under the null hypothesis (i.e., the size of the test): if the null hypothesis were true what is the probability that we would find an absolute value larger than the critical value? A small size (high critical value) reduces the probability that we will identify sampled populations as possessing truly different moments.

Of course, there is another question: what is the probability that we would wrongly identify the moments as equal when they are truly different? What is the probability of type II error? (Equivalently, what is the power of the test?) The question is not specific enough to admit of an answer. There is a tradeoff between size and power. In the extreme, we can avoid type I error by accepting all null hypotheses, and we can avoid type II error by rejecting all null hypotheses (cf. William H. Kruskal 1968a, p. 245). The choice of an intermediate size, such as 5 percent, is frequently conventional and pragmatic (as with Edgeworth's two to three times the modulus rule), but aims to make sure that type I error is tightly limited at the cost of having power only against alternatives that are sufficiently far from the null.<sup>3</sup> The power of a test can be computed only for specific alternative hypotheses. Still, absent a well formulated alternative, we know that, for any given size, type II error will explode (power approach zero) if the true difference in moments is small enough.

---

<sup>3</sup> Ronald Aylmer Fisher, who is an object of McCloskey's (1998, p. 112; see also Ziliak and McCloskey 2004a, pp. 530-531, 542-544) special scorn clearly understands the conventional nature of the customary 5 percent size: "it is convenient to take [a size of 5 percent or critical value of 1.96] as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant" (Fisher 1946, p. 44). Ziliak and McCloskey (2004a, p. 531) imply that Fisher has contradicted himself or, at least, shifted ground in moving from *convenient* in one sentence to *formally regarded* in the next. We see Fisher and declaring a pragmatic convention, using *formally* in the sense of "[A]s a rule"; under normal circumstances" (*Oxford English Dictionary*, definition 4.b). Our view is reinforced by the fact that Fisher discusses the implications of a number of different test sizes in the same passage.

The second type of test asks not whether sample moments are the same, but whether an estimated parameter is consistent with a population in which that parameter takes a definite value or range. Student's  $t$ -test is, perhaps, the most familiar example of the type. If  $\hat{\beta}$  is an estimated regression coefficient,  $\beta_{null}$  the value to be tested, and  $\hat{\sigma}$  the estimated standard error, then  $t = \frac{\hat{\beta} - \beta_{null}}{\hat{\sigma}}$  has a known distribution under the null hypothesis, conditional on the underlying normality of the errors. Thus, the probability of finding  $|t|$  greater than a particular critical value is given by the size of the test corresponding to that critical value (e.g., 5 percent corresponds to 1.96). The second type of test is, therefore, a special case of the first type.

It is perhaps not emphasized frequently enough that reasoning from significance tests proceeds from a statistical model, not directly from the raw data. As Edgeworth (1885, pp. 186-187, 208) already understood data may have to be transformed to account for the underlying processes that generate them before they will conform to a distribution supporting statistical tests. Reasoning from (raw or transformed) data is only as good as the conformity of the data to the supposed probability distribution. Specification tests (e.g., tests of homoscedasticity, serial correlation, or normality), which are significance tests of the first type, provide evidence that supports the conjecture that the statistical model is a good one. We take up McCloskey's failure to address this vital use of significance tests in section 2.5.

In a Neyman-Pearson framework in which the investigator is able to consider a well-defined alternative hypothesis explicitly, acceptance and rejection of a null hypothesis are symmetrical concerns, and the choice is how to strike a balance between type I and type II error. When alternative hypotheses are not explicit, acceptance and

rejection are asymmetrical. A value greater than the critical value rejects a hypothesis, but a value less than the critical value does not imply acceptance, but failure to reject it. The point is not a misplaced fastidiousness about language or mere "wordplay" (Kruskal 1968a, p. 245), even in those cases in which a failure to reject leads to ignoring the hypothesized effect.

Rather significance tests are a tool for the assessment of signal strength. Rejection indicates a clear signal. Failure to reject offers no evidence for choosing between two possibilities: there is no signal to detect or noise overwhelms the signal.

Fisher (1946, p. 44) acknowledges this asymmetry when he states that, given a 5 percent (or other conventional size), "[s]mall will escape notice if the data are insufficiently numerous to bring them out . . ." In pointing out the asymmetry of the significance test, we are not asserting that statistical significance is either necessary or sufficient for economic significance. A noisily measured effect may be economically important; a well measured effect may be economically trivial.

## 2.2 DOES SIZE MATTER INDEPENDENTLY OF STATISTICAL SIGNIFICANCE?

In one sense, the answer to this question is obvious: size (here the magnitude of influence or what McCloskey refers to as "oomph" (1992, p. 360; cf. Ziliak and McCloskey 2004a, p. 527) clearly matters. But who ever said otherwise? A well-measured but trivial economic effect may be neglected. But how should we regard an effect that is economically large but poorly measured in the sense that it is statistically insignificant? McCloskey (2002, p. 50) answers this way:

The effect is empirically there, whatever the noise is. If someone called 'Help, help!' in a faint voice, in the midst of midst of lots of noise, so that at the 1% level

of significance (the satisfactorily low probability that you will be embarrassed by a false alarm) it could be that she's saying "Kelp, kelp!" (which arose perhaps because she was in a heated argument about a word proposed in the game of Scrabble), *you would not go to her rescue?* [McCloskey 2002, p. 50; cf. 1998, p. 117]

The principal claim – "the effect is empirically there, whatever the noise is" – is extraordinary. Noise masks the signal. It may be there or it may not be there. The point is that we do not know.

Clearly, if the costs and benefits are sufficiently skewed, we may seek more data in order to reduce the noise. If the apparent faint cries for help come from a rubble heap after an earthquake (a situation in which we expect there to be victims), we literally dig further to get more data. Our immediate conclusion is not that there is someone alive down there, but that there could be. Yet, if the signal does not improve, we may reasonably give up looking; for, after all, the cost of looking in one pile, may mean that we are not finding victims in another. The point is that whether we respond to a signal depends on an interaction between the value of what is signaled (here a life) and the background noise. If the potential cost of type II error is large, we may choose a large test size (here we dig in response to a faint signal).

Notice, however, it is the *potential* size of the payoff that matters (i.e., the value of what is signaled), not the size of the effect as estimated from the faint signal (i.e., the value of the signal). The point is clearer in a different sort of example: in a clinical trial a single subject with a migraine headache is given a licorice jelly bean, and the migraine quickly subsides. Should we conclude that licorice jelly beans are effective against migraines? Clearly not; the noise overwhelms the signal. Yet, the principle is no different if we have five subjects, three of whom, experience subsidence of migraine

symptoms after eating a licorice jelly bean. A rate of 60 percent is large, but we can have no confidence that it will stay large as more observations accrue. A sixth observation, for example, will either raise the estimate to 67 percent or reduce it to 50 percent. Which way it goes depends, in part, on whether the true systematic effect lies above or below the 60 percent estimate and, in part, on the size and variability of other influences – the noise. The key point, well known to statisticians and econometricians, is that the presence of the noise implies that the measurement is not reliable when there are five or six or some other small number of observations. The systematic effect is so badly measured that we can have no confidence that it is not truly some other quite distant number, including zero. The function of the significance test is to convey the quality of the measurement, to give us an idea of the strength of the signal. The principle involved when  $N = 1$  or 5 is no different than when  $N = 10,000$ .

One might well doubt that McCloskey would advocate using so silly an example as the licorice jelly beans, but in fact she does:

*It is not the case that statistically insignificant coefficients are in effect zero. The experiments on aspirin and heart disease were halted short of statistical significance (at the level the medical researchers wanted to have) because the effect was so large in life-saving terms that it was immoral to go on with the double-blind experiment in which some people did not get their daily dose of aspirin. [McCloskey 1998, p. 118]*

McCloskey cites the aspirin-heart attack story frequently (additional references:

McCloskey 1995, p. 33 and 2002, p. 52; 2005, p. 22). If it were true, it would illustrate the reasoning that she means to endorse. But, alas, the story is false: the study was

halted early because the effects of aspirin were *both* large and statistically significant.<sup>4</sup>

To have halted it early simply because the results were large, though badly measured, would be to make the licorice-jelly-bean mistake.

The message matters, but the message must rise above the noise before we can know what it is. The Second Coming of Christ would surely be a defining event in human history, but the most devout Christian (perhaps especially the most devout Christian) should hesitate from declaring a vision of Jesus when it could equally well be the play of light on a snowdrift.

McCloskey writes as if statistical significance or insignificance were a mere artifact of sample size (McCloskey 1998, pp. 117-119; 123-124; McCloskey and Ziliak 1996, pp. 98-99, 101; Ziliak and McCloskey 2004a, p. 535). As Ziliak and McCloskey (2004a, pp. 540-541) put it, "all hypotheses are rejected, and in mathematical fact, without having to look at the data, you know that they will be rejected at any pre-assigned level of significance."<sup>5</sup>

McCloskey frequently writes as if the fact that estimated standard error falls as sample size rises implies that an estimate on a small data set will necessarily yield essentially the same mean value as one on a large data set, though the former will be

---

<sup>4</sup> According to the study in question, "the external Data Monitoring Board of the Physician's Health Study took the unusual step of recommending the early termination of the randomized aspirin component of the trial, primarily because of a *statistically* extreme beneficial effect on nonfatal and fatal myocardial infarction had been found." The "difference between total myocardial infarction between aspirin and the placebo group has a  $P < 0.00001$ " (Charles H. Hennekens 1988, p. 262). McCloskey based her interpretation on secondary sources rather than the original study. McCloskey was apparently misled by the statements in a letter to the editor of the *New England Journal of Medicine* stating that "researchers found the results so positive that they ethically did not feel they should withhold aspirin's benefits from the control placebo group" and, in a news report (*FDA Consumer* January-February 1994), that "[t]here was, however, no significant difference between the aspirin and placebo groups in number of strokes . . . or in overall deaths from cardiovascular disease" (both cited in e-mail McCloskey to Kevin D. Hoover 31 July 2003). But "positive" need not mean, as McCloskey takes it, large and statistically insignificant nor are heart attacks the same thing as strokes or overall deaths from cardiovascular disease.

<sup>5</sup> McCloskey (1985, p. 202) was more moderate and more correct when she noted that this claim does not hold when the true hypothesis is "literally zero" – see also section 2.5 below.

insignificant and the latter significant (McCloskey 2002, pp. 50-52; Ziliak and McCloskey 2004b, p. 674). This is the basis for Ziliak and McCloskey's (1996, p. 108) claim that policy ought to be guided by *insignificant* mean estimates – ignoring the noise and calculating costs and benefits as if the mean were known: "You have to go with what God has provided" (McCloskey 2002, p. 53; cf. sections 3 and 4.2 below).

The argument is fallacious. Yes, the estimated standard error falls as the sample size rises, but the noise does not necessarily shrink around a fixed mean (cf. Jeffrey M. Wooldridge 2004, pp. 578-579). The point is clearly expressed by Rebecca Blank (1991, p. 1053) in a paper on the effects of blind refereeing in the *American Economic Review* that showed that women's acceptance rates were lower, though statistically insignificant, in blind rather than non-blind samples:

It is true that these differences would be significant with a larger sample of papers. The lack of statistical significance, however, indicates that one cannot assume that a larger sample of papers would produce the same result.

Surely, Blank's study addresses an important question for the economics profession. The message of the significance tests is not that women *are* treated fairly in the refereeing process, but that we do not have *clear* evidence that they are not.

To make the point clear, consider a world in which investment is generated according to the following equation:

$$(1) \quad \textit{investment} = \textit{constant} + \alpha \times \textit{interest rate} + \beta \times \textit{tax rate} + \textit{error}.$$

If tax rates and  $\beta$  are big, then tax rates are economically important. But let us suppose that for the entire sample period, the tax rate is constant. This assumption in no way

undercuts the economic importance of tax rates. Yet, if we tried to estimate equation (1) as a regression, the package would stall: the tax rate and the constant are perfectly collinear and the product matrix of the independent variables is singular. This is equivalent to an infinitely large standard error on  $\beta$ . In this case, no measurement is possible unless we eliminate the tax rate from the regression.

Now add some noise. Construct the tax variable as a constant plus a very small random term – just big enough to allow the econometric package successfully to invert the product matrix. The standard error on  $\beta$  will be huge (the  $t$ -statistic near zero), and the coefficient estimate could essentially fall anywhere from negative to positive infinity: the true  $\beta$  might equal 1 (and that might well be economically important), and the estimated  $\hat{\beta}$  equal 2,000. (If  $\hat{\beta}$  does not differ much from  $\beta$ , redraw the error terms and try again. A small number of tries should produce a large difference.)

Obviously, we should not regard this estimate as the true one nor use it for policy analysis. The message of the standard error is that the signal is swamped by the noise, so that we cannot detect the true value. Were the number of observations or the variability of taxes much larger, then we might get a decent estimate. But, in general, it will be one that converges on the true value, not – as McCloskey's practice implies – one that approximates the initial estimate with any degree of reliability.

McCloskey's unsupportable notion that an estimate can be reliably used independent of the accompanying noise leads to a perverse conclusion. As written, question 1 (Table 1) proposes that few observations are to be preferred to many. Of course, that is nonsense; more information is better than less. What is true, as McCloskey and Ziliak (1996, p. 102) hint, is that the power of tests rises with the number of



observations, so that there may be a more favorable tradeoff between type I and type II error in large samples that can be captured by using a test with a smaller size. Failing to reduce the test size with an increasing number of observations may express an idiosyncratic preference for power over size, but it is not a mistake in logic.

### 2.3 THE NEYMAN-PEARSON FRAMEWORK

The only situations in which McCloskey appears to accept the usefulness of statistical significance is when it is cast into a strict, decision-theoretic Neyman-Pearson framework, the marker of which is the existence of an explicit loss function: "You can't run science without a loss function" (McCloskey 1998, p. 118; cf. 1985b; 1992, p. 359; 1999, p. 361; 2002, p. 58; Ziliak and McCloskey 2004a, p. 543). Questions 8 and 9 (Table 1) concerning statistical power indirectly address the issue of loss functions.

McCloskey confuses science and engineering. Broadly speaking, the goal of science is to understand the way the world works; the goal of engineering is to achieve particular outcomes. Science and engineering frequently stand in a relationship of mutual aid. Still, they are conceptually distinct. The Neyman-Pearson framework is most at home in applications such as production control in which it is possible to formulate a tractable probability model in which a null hypothesis with a crisp alternative can be related through a clearly defined loss function. Such engineering problems may provide illuminating metaphors for science, but science cannot broadly advance by applying such a framework.

Because tractability is central, the Neyman-Pearson framework can be applied only to "small worlds" (to use Leonard Savage's 1954 [1972], pp. 82-91, evocative

term).<sup>6</sup> The way of truth is one; the ways of error many. In some tightly defined engineering problems (some of which may be embedded in a scientific investigation), the detailed Neyman-Pearson approach may have considerable purchase. But how could it be applied to a larger scientific inquiry, in which it would be impossible to articulate, much less to incorporate into a loss function, every alternative to the null hypothesis? And whose loss function should we choose?

Consider that landmark in the history of physics, Isaac Newton's mechanics. Newton, of course, did not appeal to statistics in his *Principia*.<sup>7</sup> Types I and II error are not unique to statistical inference, however, and the case for or against an optimal balance between them based on a loss function applies in a wider domain.<sup>8</sup> In formulating his mechanics, none of Newton's choices were governed by a balancing of type I and type II error. Around which of the uncountable (and to him often unforeseeable) practical applications of his mechanics should Newton have structured his loss function? An engineer designing a bridge may have a loss function; he may find Newtonian mechanics useful in articulating that loss function; but the truth of Newton's mechanics in no way depends on the benefits of the traffic flowing over the bridge nor yet on the costs of its falling down.

---

<sup>6</sup> Savage used this term to explain the limits on his own preferred sort of personalist Bayesian statistics, which he regarded as "a natural late development of the Neyman-Pearson ideas" (see Keuzenkamp 2000, pp. 84-86 for the citation and discussion).

<sup>7</sup> Newton did use statistical ideas in assessing historical evidence, but seems to have foregone applying them to a problem of practical personal gain in assessing the quality of the coinage while Master of the Mint (Stigler 1999, pp. 394-397).

<sup>8</sup> Graham Elliott and Clive W.J. Granger (2004, p. 549) make the point that the degree of bending of starlight in Arthur Eddington's famous observations of the eclipse in 1919 was too small to matter practically, yet nevertheless served to distinguish Einstein's from Newton's mechanics. Ziliak and McCloskey's (2004b, pp. 668-669) riposte that Eddington did not use statistics misses the point: Elliott and Granger never said that he did, nor need they even have believed it implicitly, since the argument about loss functions is more general than statistics. Kruskal (1968b, p. 218) points out that many objections to *statistical* inference apply "equally to any mode of analysis – formal, informal, or intuitive."

If the science of physics need not depend on its practical import, why must the science of economics? McCloskey may wish to maintain that economics, by its nature, should be more worldly than physics. Perhaps. But that is quite different than asserting that, to be a science, it must be connected to worldly concerns through a loss function.<sup>9</sup>

McCloskey is engaged in a sleight of hand, equivocating on the words "science" and "economics" and their derivatives. The *Oxford English Dictionary* gives two pertinent definitions of "economic": definition 2.a: "[r]elating to the science of economics; relating to the development and regulation of the material resources of a community or nation"; and 2.c: "[p]ractical or utilitarian in application or use." To argue that "economic significance" is a parallel notion to "scientific significance," McCloskey implicitly appeals to the first definition, while to sustain the centrality of the loss function to economics and science, she must appeal implicitly to the second.

An example clarifies the illusion. In discussing an article by Gary Solon (1992) in which the correlation between the incomes of fathers and sons is at issue, Ziliak and McCloskey (2004a, p. 539) assert that "a tightly fit correlation of 0.2000000001\*\*\*, would say nothing of *economic* significance." We take the example to presuppose that "tightly fit" implies that the last digit is statistically significant. We readily grant that nothing of practical importance could hang on whether the true value were the one estimated or exactly 2. Yet, if an economist could in fact discriminate an economic magnitude in the tenth decimal place (a degree of accuracy that is sometimes found in physics) it would be quite remarkable and of great *scientific* interest to economists – not for its practical import but for the degree of refinement that it would imply about our

---

<sup>9</sup> McCloskey (2002, pp. 37-38) certainly claims that economics is more worldly than mathematics, which she regards not as a science, but as "a kind of abstract art," though not to be disdained for that.

empirical tools.<sup>10</sup> In general, as Edgeworth observed (see footnote 2), the purpose of science need not be practical. Some scientists have a practical turn of mind, and would never work on a problem that did not have foreseeable practical consequences – though even in that case, it is not obvious that a loss function can *generally* be successfully articulated. Yet, economics, like other sciences, distinguishes between what is true about the world and the practical import of a truth.

We do not deny that practical considerations of probable gain and cost play a part in the design and management of research. The notion of the Economy of Research was initiated by the American philosopher Charles S. Peirce (1885), who was also an important contributor to statistics and its applications to psychology and other fields.<sup>11</sup> Loss functions useful in the economic analysis of research may be difficult to formulate. But, even when they can be formulated, the relevant measures of costs and, especially, benefits are generally defined in terms of their import to the enterprise of pure science and not in terms of their practical import for the wider world. Indeed, the scientific enterprise itself need not have any practical ambitions. For example, Newton believed that gravity followed an exact inverse-square law. If, in fact, the relevant power were shown to differ from 2 in the twenty-ninth decimal place, there might be no practical implications: rockets would still fly to the moon, despite using the old formula (see section 2.6 below). Yet, the result would be highly important, as the *exactness* of the inverse-square law was an essential element in the theory.

---

<sup>10</sup> Feynman (1985, p. 7) cites, as one of many examples, measurements of Dirac's number, which are accurate to the 11<sup>th</sup> decimal place (more precisely,  $4 \times 10^{-9}$  percent), the equivalent of measuring the distance between New York and Los Angeles to the width of a human hair. Feynman cites the accuracy of this result, about five times more accurate than the predictions of the relevant theory, not for any gain or loss that it implies, but for the beauty of the conformity of theory and observation.

<sup>11</sup> On the economy of research, see James R. Wible 1994, 1998; on Peirce as a statistician, see Stigler 1999, ch. 10.

Practical implications matter, but they are not the first-order concerns of science. Rather than seeing science concerned with how things are, McCloskey sees it as *primarily* a guide to practical decisions. But that makes a nonsense of the history of science – not only of economics – as well as current practice.

We also do not deny the usefulness of the Neyman-Pearson approach as a tool of scientific investigation where the alternatives are clearly articulable, the probability models tractable, and the loss functions relevant to the scientific problem. (But as Kruskal 1968a, p. 240, points out, most significance testing cannot be cast into a crisp decision-theoretic mode; cf. Gerd Gigerenzer 2004, p. 591). The philosopher Deborah Mayo (1996), for instance, defends the Neyman-Pearson approach, particularly against Bayesianism taken as an all encompassing philosophy of science. She finds the Neyman-Pearson approach important in a program of strict local testing aimed at minimizing error – that is, at efficiently extracting signal from noise. Statistical testing, in this view, is not about directly confirming or disconfirming a high-level theory, but about obtaining reliable observations. Unlike McCloskey, however, she does not condemn Fisher for not adopting the Neyman-Pearson approach. Like Edgeworth before him, Fisher rightly saw the statistical problem as signal extraction. A reliable observation is not generally the final step in scientific reasoning, but a piece of evidence that will be assessed in a wider context. Microscopes and telescopes provide reliable observations only when in focus. Similarly, significance tests are among the focusing criteria for statistical data (cf. Hoover 1994).

Advocates of the Neyman-Pearson approach frequently favor reporting confidence intervals. McCloskey advocates reporting confidence intervals and complains

that "... economists seldom report confidence intervals" (Ziliak and McCloskey 2004a, p. 534). The claim is false. The standard practice of the large number of empirical macroeconomists who report impulse-response functions is to graph the confidence intervals. More broadly, a search of economics journals archived in JSTOR over 1980-1999, a period that covers both of McCloskey and Ziliak's surveys, finds 131 entries in articles in the *American Economic Review* for phrases indicative of reporting or using confidence intervals.<sup>12</sup> A search of all thirty-nine economics journals archived in JSTOR over the same period finds 1,788 entries.

Whatever the truth of McCloskey's view on how frequently economists report them, her favorable regard for confidence intervals is puzzling in light of her disdain for significance tests. The conventional significance test for normal errors rejects when the estimate is more than two standard errors away from the null. The confidence interval is constructed, in this case, as  $\pm$  two standard errors from the estimate. To say that zero or some other hypothesized value lies outside the confidence interval is isomorphic to computing the significance test taking the hypothesized value as the null. Even those researchers who do not report confidence intervals frequently use confidence-interval reasoning and, in most cases, report sufficient information for a reader to compute the confidence intervals themselves (coefficient estimates plus standard errors, *t*-tests – or even *p*-values – are enough to construct confidence intervals). The complaint, then, is one of taste rather than principle (see Elliott and Granger 2004, p. 548). And it is

---

<sup>12</sup> The search terms were: "confidence interval(s)," "error band(s)," "2 (two) standard error(s)," "2 (two) standard deviation(s)." The numbers for the *American Economic Review* are not strictly comparable to McCloskey and Ziliak's surveys which exclude articles from the *Papers and Proceedings* (May) numbers and shorter articles.

inconsistent: if significance tests are scientifically suspect, so must confidence intervals also be.

## 2.4 THE PROBABILITY APPROACH TO ECONOMETRICS

In many situations, McCloskey believes that significance tests, even in a proper Neyman-Pearson framework, have no place, because economists possess the entire population:

For one thing, as we have said, the errors that tests of significance deal with are errors of sampling, but in many cases there is no sample involved: we have the entire universe of observations of the general price level in the United States and the United Kingdom 1880-1940. [McCloskey 1984, p. 134; cf. McCloskey 1985b, p. 204]

McCloskey misses a vital point. If we calculate the mean weight of Mrs. Clary's sixth-grade, we may regard the class as the entire population, and the mean as a descriptive statistic. But if we use the mean of Mrs. Clary's class as an indication of the likely mean of Mrs. Coyle's sixth-grade class, then we are regarding Ms. Clary's class as a sample of wider population (e.g., of the sixth-grade class in Graham Road Elementary School). Any time we project a statistic out of sample or consider its projectability to be the vital question, we regard the observed data as a sample. Of course, questions of the representativeness of the sample, of the homogeneity of the population, and so forth arise immediately, but they are not the essence of McCloskey's point.

Consider a real-world example familiar to many universities: a study that compares salaries of female and male faculty. The key question in such a study is not the raw fact of the difference in means between men and women. Instead, the objects should be, first, to determine whether there are differences that, controlling for other factors not indicative of discrimination, can be attributed to sex and, second, to shed light on the

mechanisms behind any discriminatory difference (e.g., does it occur because one sex is hired at lower rates of pay or because promotion is slower?).

If we construct a dataset that covers the entire faculty of the University of California at Davis we do not have – as McCloskey insists – the whole population, since the questions that we want to answer most involve a population wider than the current faculty in a variety of senses. To take one example: if the current mechanisms remain in place, would the next cohort of women hired be paid the same as men with similar relevant characteristics? Here the mechanism is regarded as open-ended, and we simply wait for more realizations. A second example: are the differences between the pay of men and women the result of differences in relevant characteristics or sex discrimination? Here the population is a hypothetical one: we use the actual faculty cohort (a sample) to infer what outcomes would be for a cohort that differed from the actual one only in the sex assignment of its members (unobserved members of the population).

Ziliak and McCloskey (2004a, p. 543) dismiss appeals to hypothetical populations as “metaphysical” – with the implication that the term is synonymous with “nonsense” (also McCloskey 1985a, pp. 161-162)<sup>13</sup> But the scientific question is not generally, “what happened?”, but “what lies behind whatever happened?” The policy question is generally, “what will happen if . . .?” Answers to either sort of question contemplate mechanisms that are capable of realizing outcomes other than, and additional to, those already observed. The observed are, then, part of a sample, and the measures of the precision of estimates of key features of the underlying mechanism – expressed for

---

<sup>13</sup> She is inconsistent. In other moods, McCloskey (2002, pp. 30, 32) berates economists for dismissing metaphysics.



example – in standard errors, confidence intervals, and significance tests, are of great utility.

McCloskey seems to imagine that such measures of precision are at home only in the narrow case of situations similar to production control. If you want to produce ball bearings within a certain tolerance, but do not wish to test each one, test a sample (with a Neyman-Pearson loss function guiding acceptance or rejection), and project the outcome to the whole population. But science is not an assembly line. Many experiments are run and *all* the results are recorded. By McCloskey's reasoning, the data from such experiments do not supply samples, but are in fact the whole population. But, of course, the point of the experiment is discover what would happen were the experiment to be run again and to develop evidence bearing on the mechanisms that lie behind the experimental realizations.

Economic data are rarely experimental. McCloskey (1985b, p. 204; 1998, pp. 138) particularly sees time series data as never meeting the conditions of being a sample of a population: history has only one run. Yet, experiments are similar in this respect. An experiment can never be repeated exactly. When we say that are repeating the "same" experiment, we are making a judgment that we have repeated the *relevant* aspects of the set up. We ignore many others: the chemist does "not note that this phenomenon was produced on such a day of the week, the planets presenting a certain configuration, his daughter having on a blue dress, he having dreamed of a white horse the night before, the milkman having been late that morning, and so on" (cf. Peirce 1934, vol. 5, para. 591). It is not that these could not under any conceivable circumstances be relevant;

rather, we make fallible judgments that they are not in these circumstances in fact relevant.

Trygve Haavelmo's (1944) great contribution to econometrics was to show that non-experimental data, which did not automatically conform to a statistical model such as contemplated by Fisher, could nevertheless be modeled as an economic structure plus errors, where the errors conform to a probability model (see Mary S. Morgan 1990, p. ch. 8, and Aris Spanos 1995). A time series in this conception is, in part, built up of realizations of a probability distribution, so that the variables are a sample of population that will 1) provide further realizations if allowed to run; and 2) could have provided different paths had the errors been realized differently. There is only one actual path of a time series, but there could have been others, and it makes sense to ask how far could a path that was not realized depart from the one that was actually realized without there being any change in the underlying structure. Making such assessments is, of course, one of the uses of standard errors, confidence intervals, and significance tests.

Time series and experimental data are different; yet, in respect, of whether we do or do not have the whole population, it is not a difference in kind. There is a special uncertainty, which is not captured in sampling error, about homogeneity. In the case of time series, radically different behavior may result because the deep structure generating the data changes (a structural break). Such a concern is particularly pertinent for time series, but again it has analogues in other kinds of cases. Significance tests of the first type ("is the moment of this sample distinguishable from the moment of that sample?") are useful in detecting such departures from homogeneity, since they address the question of whether the signal rises above the noise. Would, say, the difference of the means be

larger than what would normally follow from random variation in a homogeneous structure?

It is true that we have all the observations of the general price level in the United States and the United Kingdom 1880-1940 that were actually realized. But with respect to the question of purchasing-power parity, in which context, McCloskey originally made her claim, so what? As scientists, we are interested in whether the economic mechanism displays purchasing-power parity. The 1880-1940 period can be seen as a sample of a population that extends into earlier and later times. And, just as Mrs. Clary's sixth-grade class can be seen as a sample of the sixth-grade classes in Virginia in 1967, as well as a sample of the sixth-grade classes at Graham Road School, the time series of prices can be seen as a sample of a population that extends across other countries as well as other times and over realizations that never occurred. We do not wish to minimize the problems of ensuring that such populations conform to an adequate probability model and are, therefore, sufficiently homogeneous to sustain useful inference. But again, significance tests are among the tools that we use to provide evidence on that point.

It is puzzling that McCloskey, an advocate of the cliometric approach to economic history, should fail to grasp the point that we must situate observed data in the context of population data that could have been, but was not in fact, observed. The great achievement of the cliometric school is to understand that historical data need not be viewed as "one damned thing after another" but can instead be viewed as the outcome of an enduring economic mechanism that conforms reasonably to economic theory. The counterfactual or hypothetical nature of such scientific knowledge is acknowledged for example in Robert W. Fogel's (1964) famous counterfactual analysis of the development

of American railroads and the literature it spawned. Fogel essentially compares one run of the American economy as it was actually realized to one run in which there were no railroads. Such a comparison relies on the idea that what was actually observed was but one possibility – a sample from a hypothetical population (cf. James Woodward 2003).

Hoover (2001, ch. 4) argues that all causal explanation of particular events (the crash of an airplane or a stock market) relies on subsuming the contributory elements to particular generic causal mechanisms. If these elements cannot be viewed in such a way that (in stochastic cases) the observations form a sample of a larger, but unrealized, population, then no causal and, therefore, no historical explanation is possible.

It is equally puzzling that McCloskey, who strongly advocates simulations as an alternative to significance testing, should fail to see that simulations are necessarily counterfactuals – what could have been but was not or what could be but is not yet (Table 1, question 17; McCloskey and Ziliak 1996, pp. 104, 112). The meaningfulness of any such simulation rests on exactly the same grounds as the meaningfulness of treating realized data as one possible draw from an unobserved population.

## 2.5 SPECIFICATION SEARCH

McCloskey condemns the practice of using significance tests to eliminate variables from regression equations and the related practice of specification search (Table 1, question 14; McCloskey 1992, p. 361; McCloskey and Ziliak 1996, pp. 104, 106; Ziliak and McCloskey 2004a, p. 534). In emphasizing the size of the estimated conditional means of regression equations (the “oomph”), she avoids one of the most central issues in statistics and a vital use of significance tests. The approach is question-begging,

assuming the very thing that needs to be demonstrated: "the accuracy of [the] estimated mean [of a regression coefficient] depends on the properties of the error term, the specification of the model, and so forth. *But to fix ideas suppose that all the usual econometric problems have been solved*" (McCloskey and Ziliak 1996, p. 98, emphasis added). But how would we justify such a supposition?

As we stressed in the last section, statistical measurements of economic magnitudes are meaningful only in the context of a statistical model that adequately characterizes the data. To establish the congruence of the model to the data requires testing (see Spanos 1995 and Søren Johansen 2005). Significance tests are used to provide evidence relevant to the existence of serial correlation, heteroscedasticity, departures from normality, structural breaks, and other departures from the statistical model that justifies the assumption that a particular coefficient estimate is a good one. Such tests are typically applied in the asymmetrical manner of Fisher (reject/unable-to-reject instead of reject/accept) rather than in the context of a particular loss function. But, even if such a loss function were available, the connection between the critical value in, say, a test of serial correlation and the ultimate practical economic gains or losses from an error in assessing a regression coefficient are so complex and indirect that they would defy calculation. The power of such tests is not typically ignored; it is a major consideration among professional econometricians, even if it is less frequently discussed by applied economists.

Aside from some desultory condemnations of the use of such specification tests (e.g., McCloskey 1999, p. 361 on unit roots tests), McCloskey has little systematic to say about error specification, taking it to be unproblematic. In contrast, she clearly attacks

the elimination of variables from regression specifications on the grounds of their statistical insignificance. She appears to have three (closely related) objections. The first is that a statistical criterion may result in the elimination of economically significant variables. The second is closely related: a large sample size generally produces statistically significant estimates even when they are not significant economically, which suggests that a statistical criterion is irrelevant. Finally, elimination of coefficients smacks of data mining, which is widely regarded as an unacceptable practice.

As we saw in section 1, typical significance tests relate the magnitude of a measured coefficient to the sample variability. But sample variability is itself a fortuitous product of the sample size. The economic meaning of a coefficient estimate should not then, in McCloskey's view, depend on its statistical significance. But again, statistical significance conveys important information about the accuracy of the measurement, even though it does not determine the economic significance of an estimated coefficient. As the example of the investment equation in section 2.2 above makes clear, eliminating a variable from a regression may be a reasonable thing, despite its economic significance. The role of the standard error and the significance test is to tell us where we find ourselves along the continuum from the impossibility of measurement in the case of an absolutely constant variable to the perfect accuracy of an infinite sample.

Elimination of variables that are badly measured is not a statement about their ontological status – in the investment equation, taxes are economically important *ex hypothesi*. Rather, it is a measurement strategy. McCloskey's general claim that any coefficient is significant if the sample size is large enough is false (see section 2.2 above). As Kruskal (1968a, p. 246) reminds us, some nulls may be exactly zero. And if a

coefficient is exactly zero, then a larger and larger sample simply shrinks the standard error around that value.

There are two cases in which a regression coefficient should naturally return a value of exactly zero (Hoover 2001, ch. 2, sec. 2.4). First, a coefficient may have a range of possible values that includes zero. Policy variables, for instance, may include zero as a focal point when the policymakers intend to use their instruments to eliminate the influence of a particular factor. For example, suppose that the money were causally linked to real GDP and the interest rate and that the Federal Reserve adjusted the money supply in such a way as to hit a target interest rate precisely. If they are successful a regression of the interest rate on money and income should find zero coefficients, except for sampling error (see Hoover 2001, pp. 47-49).

Second, an estimated coefficient may converge on zero because the variable in question is not causally connected to the dependent variable. In the first case, a variable can be conceived of as a set of possible values ( $\Omega$ ) which takes a particular value ( $\beta = 0 \in \Omega$ ). In the second case, the set of possible values is empty ( $\Omega = \emptyset$ ).

If neither of these cases apply (the true  $\beta \neq 0$ ,  $\Omega \neq \emptyset$ ), then for a sufficiently large number of observations, provided that the variable itself is not constant, the estimate of  $\beta$  will be statistically significant at whatever size one likes: more data leads to more precise measurement. Additional data are an unqualified good thing, which never mislead. A very precisely measured but substantively small coefficient may not be practically significant. Being precisely measured, one can be confident in that judgment. A badly measured but large coefficient may or may not be practically significant: the

message of the significance test is that one cannot be confident one way or the other; more data are needed.

A variable that happens to be set to zero remains economically significant – it could have been set to some other value. A variable that is not causally connected to the dependent variable is economically insignificant. Both should show up as indistinguishable from zero in a regression equation. Non-statistical information is vital to distinguishing these two cases (Hoover 2001, ch. 8). Evidence that a coefficient cannot be distinguished from zero, means either that one of these cases holds or that the magnitude of the coefficient is close enough to zero that the test lacks power. We have no choice in such a case but to admit that our measurements are imprecise, to keep an open mind, and to seek more evidence. But if the coefficient is statistically significant, then we can legitimately address its economic magnitude.

Variable elimination on the basis of significance tests is a form of data mining. McCloskey (1985a, p. 170; 1985b, p. 201; 1992, p. 361; 1997, p. 242) accepts the conventional condemnation of data mining without much further analysis.<sup>14</sup> One class of objections is that data mining generates size distortions that make statistically insignificant results appear to be statistically significant. For instance, Michael Lovell (1983, pp. 2-4, cited by McCloskey 1985b, p. 201, and McCloskey and Ziliak 1996, p.

---

<sup>14</sup> Data mining is discussed in detail in Hoover (1995) and Hoover and Stephen J. Perez (2000). On the one hand, McCloskey condemns data mining; on the other hand, McCloskey (1985a, pp. 139-140; 1985b, p. 162) cites favorably Edward Leamer's (1978) analysis of specification search and his (1983) "extreme-bounds analysis." Both involve data mining. Oddly, Ziliak and McCloskey (score sheets for 2004a, personal communication) omit both Leamer's (1983) article and Michael McAleer, Adrian Pagan, and Paul A. Volcker's (1985) rebuttal, both of which appeared in the *American Economic Review* and meet the criteria for their survey of articles from the 1990s. Similarly, Thomas F. Cooley and Stephen F. LeRoy's application of extreme-bounds analysis to money demand, which is itself favorably cited by McCloskey (1985a, p. 140), is omitted from the survey of the 1980s. More oddly still, in light of the favorable evaluation of extreme-bounds analysis, Ross Levine and David Renelt's (1992) article, which is a straightforward application of extreme-bounds to cross-country growth regressions, scores only 3 out of 19 in the 1990s survey – the third worst performance on their survey in the 1990s.



112) demonstrates that, when an independent variable is truly random and all potential regressors are orthogonal, the probability of finding significant regressors from a fixed set is higher than the nominal test size. How much higher depends on how many regressors are considered. Lovell's conclusions are closely related to the characteristics of optional stopping rules, which have been carefully analyzed by statisticians (see Mayo 1996, chapter 10), and to much discussed, but less well analyzed, "file-drawer problem" – that is, to the publication only of significant results (see, for example, James Kahn, Steven Landsburg, and Alan Stockman 1996; Kruskal 1968a, p. 245; Ziliak and McCloskey 2004a, p. 534).

It is common, but not warranted to generalize from these objections to a general condemnation of specification search. Lovell (1983) shows that with a fixed set of regressors and a fixed sample size that three search algorithms (stepwise regression, max  $\bar{R}^2$ , and max-min  $|t|$ ) show substantial size distortions. However, these particular search algorithms pay no attention to the congruence of the data with the hypothesized statistical model of the error, which turns out to be crucial.

A theorem due to Halbert White (1990, pp. 379-380) states that, for a fixed set of specifications and a battery of specification tests, as the sample size grows toward infinity and increasingly smaller test sizes are employed, the test battery will, with a probability approaching unity, select the correct specification from the set. Both type I and type II error fall asymptotically to zero. The theorem says that, given enough data, only the true specification will survive a stringent enough set of tests. The theorem provides a deep justification for search methodologies, such as the "LSE approach" (Grayham Mizon 1995), that emphasize rigorous testing of the statistical properties of the error terms.

McCloskey and Ziliak's (2004a, p. 530) jabs at David F. Hendry's (1980, pp. 27-28) mantra, "the golden rule of econometrics is test, test, test," miss the mark: Hendry and other LSE econometricians do not mistake statistical for economic significance, rather than insist that a coefficient can be well measured only in the context of a statistical model that meets stringent criteria that would support accurate measurement. Specification testing provides evidence of the congruence between the supposed model and the data.

In Monte Carlo experiments, Hoover and Perez (1999, 2004) and Hendry and Hans-Martin Krolzig (1999) and Krolzig and Hendry (2001) demonstrate that LSE search algorithms have, in contrast to the algorithms criticized by Lovell, size near the nominal size of the underlying tests and power near the best possible for given signal strengths. These results depend crucially on specification search (data mining). As Hendry and Krolzig (2004) put it the costs of search (when the search is conducted sensibly) are small relative to the costs of inference from using an incorrectly specified model.

In supposing that "all the usual econometric problems have been solved," McCloskey herself avoids the real work of econometrics. Kruskal (1968b, p. 218) – one of McCloskey's (1998, p. 116; Ziliak and McCloskey 2004a, p. 530) most revered authorities – says, "it would be ridiculously rigid to refuse to use inferential tools [hypothesis tests] in the exploration of data." In opposing significance tests for specification adequacy, McCloskey would deny economists the tools to do their work.

## 2.6 REAL SCIENTISTS DO USE SIGNIFICANCE TESTS

Real scientists, McCloskey believes, have little use for significance tests. She hopes that economics will "grow up and start focusing its energies on doing proper science (the way physics or geology or anthropology or certain parts of literary criticism do it) . . ."

(McCloskey 2002, p. 1). Ziliak and McCloskey (2004a, p. 533) state that they have found "by examining *The Physical Review* that physicists approximately never use tests of statistical significance; so too in the magazine *Science*, the chemists and geologists; many biologists reporting their results in *Science* are less clear-minded on the matter . . ." (cf. McCloskey 1999, p. 357-358; 2002, pp. 46 and 54; Ziliak and McCloskey 2004b, p. 669).

McCloskey is wrong; significance tests are a standard tool in the physical sciences. A skewed pattern of tool use reflects, not a difference between real science and cargo-cult science, but differences in the observational problems faced in different fields. Each chooses tools appropriate to the problem to hand.

To gather some evidence on this point we performed a search of the JSTOR Archive of academic journals for the period 1980 through 1999 (the period of the two surveys of the *American Economic Review* reported in McCloskey and Ziliak 1996 and Ziliak and McCloskey 2004a) using a set of keywords that would indicate the use of tests of statistical significance.<sup>15</sup> The *Physical Review* is not archived by JSTOR, so we searched its online version on the website for the American Physical Society's journals. The search provides counts of the number of articles (not the number of instances of a use) in which any one of the keywords is found. For instance, Table 2 shows that for a

---

<sup>15</sup> The key words were "statistically significant," "statistical significance," "significance test," "test of significance," "significance tests," "tests of significance," "t-test," "F-test," and "chi squared."

search of seven General Science journals (a category that includes the journal *Science* cited by Ziliak and McCloskey), 23 percent of articles used one of the keywords.

To get a sense of the relative differences in use of significance tests between fields, we performed searches of the seven General Science journals in which the words "astronomy," "biology," "chemistry," "cosmology," "geology," "medicine," and "physics" appear. An inspection of the actual papers found in the chemistry search suggest that many of them are biological or medical in nature. Since McCloskey (1999, p. 32; 2002, pp. 1, 49; 1998, p. 112; Ziliak and McCloskey 2004a, p. 533) has tarred biology and medicine with the same brush as economics, we also performed searches in which "chemistry" and at least one of the keywords, but not "biology," "biological," "medicine," or "medical," should appear. In addition, we performed keyword searches for 19 Anthropology journals, 57 History journals, 63 Language and Literature journals, and, for comparison, 39 Economics journals and, separately, the *American Economic Review*.

The search filter is obviously a crude one, but it does give some idea of the relative use of statistical significance in various disciplines. The lowest rates of use are found in Language and Literature and in History. Since most of the work published in these fields is not scientific in the way that term is typically understood, the fact that the hit rates are as high as 1 percent and 4 percent is, perhaps surprising. The highest rate is found in Economics, with a hit rate of 53 percent. Economics is a heavier user of this statistical apparatus than the other sciences examined. McCloskey's general impression that medicine and biology are relatively heavy users of statistical significance tests is also confirmed.

But what about the "hard" sciences? They range from 8 percent for physics to 16 percent for cosmology. Physics, when measured over the general science journals, uses keywords related to statistical significance at 8 percent measured over the general journals, but at 14 percent, the same rate as chemistry, when measured over the *Physical Review*.

In reply to Ziliak and McCloskey, Horowitz (2004, p. 552) points out that physicists use the concept of statistical significance even if they do not use the name. But the evidence shows that they *do* use the name. Still, Horowitz's observation underlines that our estimates are lower bounds on the real use of significance tests. If we include "confidence level," since confidence intervals, as we saw in section 2.3, are isomorphic to significance tests, the number of hits for the *Physical Review* rises to 3,820 or 26 percent.

McCloskey (1999, p. 357) offers a specific challenge:

I invite you to look at a copy of the *Physical Review* (version C, say, the one about Nuclear Physics, the issue of November 1990, just as an example) and confirm that its 2,263 pages *contain not a single use of . . . statistical significance . . .* [emphasis added]

Visual inspection of that issue of the *Physical Review* (which in fact runs from page R1791 through 2270 – only 480 pages) shows that about 9 percent of the articles use keywords from the significance family. One example is Southon *et al.* (1990, p. R1900): "It should be noted, however, that attempts 5-8 to duplicate earlier d-d experiments revealed no *statistically significant* neutron fluxes above background" (emphasis added). While substantially smaller than the rates for economics, the data flatly contradict the conclusion that such scientists "approximately never use tests of statistical significance."

Are these examples of "real science"? The titles displayed in Table 3 are an unsystematic sample that highlights the range of applications of statistical significance in the hard sciences.

How do physicists, chemists, and geologists use statistical significance? Just like economists, if the papers cited in Table 3 are typical. Significance tests are found most frequently in non-experimental contexts in which signals must be identified against background noise or in experimental contexts in which multiple observations of what are hypothesized to be the same phenomenon must be compared. A particularly nice example is "An Experimental Test of the Inverse Square Law of Gravitation at Range of 0.1 m," by Y.T. Chen, Alan H. Cook, and A.J.F. Metherell (1984, pp. 64). The authors use tests of significance to evaluate a delicate laboratory experiment (an update of Cavendish's famous torsion-pendulum experiment of 1798) aimed at testing whether the power on distance in the empirical law of gravitation is exactly 2 as predicted by Newton's inverse-square law:

The  $t$ -test has been applied to establish the significance of the difference between the observed and calculated values of  $\Delta F/F$  in two experiments.

Further details are given in a table of  $t$ -tests (Table 7 of their paper), and these are used to assess the accuracy of the observed measurements.

The example is interesting, in part, because it illustrates the importance to physics of discrepancies too small to affect most practical applications. The experiment aims to discriminate between theories in a domain in which their predictions are not practically different, but are nonetheless scientifically different. How should these physicists define a loss function?

In this case, as in a number of other cases that we examined in detail, the researchers do not mistake the statistical significance or accuracy of the measurement for the significance of the physical measurement. (As we argue in section 3, there is little evidence that economists systematically commit this error either.) But they do use tests of statistical significance as a measure of the strength of a signal relative to the noise.<sup>16</sup>

Real science uses significance tests, but clearly there is a hierarchy in which physics and chemistry use them less frequently than cosmology, and much less frequently than biology, medicine, and economics. The difference lies not in the legitimacy of significance tests as scientific tools, but in material differences among the disciplines. Significance testing is used most in situations in which signal extraction is critical. Such situations arise most frequently (as in cosmology or economics) when experimentation is not possible or in experimental situations (as with Chen *et al.*'s torsion-pendulum experiments) where effects are small and perfect shielding from disturbing influences impossible.

When significance tests are seen appropriately as measures of signal strength, Keuzenkamp and Magnus's (1995, pp. 20-21) challenge to give an example of "a paper that contains significance tests which significantly changed the way that economists think about some proposition" appears to be rather beside the point (cf. McCloskey and Ziliak 1996, pp. 111-112). The statistical tests in Chen *et al.*'s torsion-pendulum experiments are used not to assess the theory of gravitation *per se*, but to assess the quality of the observations relative to the assumed statistical model. The measurements contribute to the weight of evidence about the inverse-square law, but no single measurement or

---

<sup>16</sup> Additional evidence is found in Staley (2004), who discusses the use of significance tests in high-energy physics in considerable detail.

experiment – and, therefore, no one test of statistical significance – is likely to be decisive. The role of significance tests is a modest one. Stigler (1999, pp. 364-365) reports that individual measures of the speed of light have not only varied through time but their the confidence intervals around the different measures do not always overlap, much less converge. He attributes this phenomenon (of which there are other examples) to the fact that simplified error models based on in-sample variability tell only part of the story, where often other sources of error prove to be more important (cf. McCloskey and Ziliak 1996, p. 112). That other sources of error sometimes dominate statistical noise does not, however, warrant ignoring the noise nor does it militate against specification tests, which are tests of statistical significance used to detect errors – both statistical (such as departures from normality) and non-statistical (such as structural breaks).

To understand that a measure of statistical significance does not of its nature decide for or against a substantive hypothesis does not mean either that the tests are useless or that empirical evidence in which they play a key role is not influential. Dominant professional opinion has shifted with respect to many economic phenomena based on accumulated empirical evidence: for example, with respect to the interest elasticity of the demand for money, the nature of the Phillips curve, or the stickiness of aggregate prices. Papers that used tests of statistical significance as tools to assess the accuracy of their measurements *relative to the assumed model of the errors*, contributed to the weight of the evidence in these cases. Economics is no different from physics in this regard.

### **3. Do Econometrics Textbooks Mislead Students?**



Up to now we have concentrated on McCloskey's wider criticisms of tests of statistical significance. Yet, we do not doubt that there are instances – both of students and professional economists – failing to distinguish between statistical and economic significance. The important questions are: How frequently do economists make this error? And do they make it out of negligence or through misapprehension? McCloskey and Ziliak (1996; also Ziliak and McCloskey 2004a) address these questions along two paths. First, they argue that econometrics textbooks provides evidence of a pedagogical failure. Second, they use their two reviews of the articles on applied economics in the *American Economic Review* to quantify economists' errors with respect to significance testing. We begin with the question of pedagogy and will take up the surveys in the next section.

In McCloskey and Ziliak's (1996, pp. 99-101) view, students of econometrics have been systematically misled by its textbooks. With few exceptions, the distinction between economic and statistical significance is never drawn. Statistics textbooks that are typically used in courses prerequisite to econometrics also fail to draw the distinction.

We share the view that econometrics instruction, at least as represented in mainstream textbooks, fails to address or, relative to some earlier textbooks, has abandoned discussions of many practical issues that arise in applied economics – discussion of the distinction between economic and statistical significance is just one among many. Econometrics textbooks have come over time to place more emphasis on estimators, their properties, and statistical theory (see Hoover 2004). One may well question the wisdom of this turn to formalism. But is it a sin? Or is it rather – at least in part – an example of the division of labor? Neither statistics nor econometrics textbooks

can be held responsible for the complete education of the applied econometrician. For the most part, McCloskey and Ziliak attack the textbooks for sins of omission rather than commission. It is rather like attacking a grocery for not selling cars.

Nowhere does McCloskey provide evidence of an economist or an author of an econometrics textbook who provides a principled argument in favor of the equivalence of economic and statistical significance. To conjure a sin of commission, McCloskey is forced to argue that the textbooks provide deleterious examples. The single case of Jack Johnston's *Econometric Methods* (1972) provides the principal evidence (McCloskey and Ziliak 1996, p. 100, McCloskey 1998, p. 117). Johnston uses a set of data on casualties in road accidents in the United Kingdom and vehicle registrations.<sup>17</sup> McCloskey's analysis is a study in careless reading. McCloskey reads Johnston as if he were offering a serious analysis of a policy, when he is clearly illustrating how to make certain statistical calculations.

The data are an 11-period time series taken from an econometrics paper from the Oxford University undergraduate examination in Philosophy, Politics, and Economics in 1959. Such data are hardly meant to be taken seriously for economic analysis, and Johnston uses them only to illustrate how to calculate regression coefficients, standard errors, and confidence intervals, as well as how to perform the analysis of variance (all a reminder of how recently the calculation of statistics by hand was the norm).

---

<sup>17</sup> McCloskey and Ziliak (1996, p. 100) report this example as "spanning pages 17 to 43" and McCloskey (1998, p. 117) as "extending over twenty-six pages [*sic*]." The first use of the data does occur on page 17 and the last on page 43. The pages in between are mostly given over to a theoretical exposition of regression, confidence intervals, correlation, and hypothesis testing, in which the data are used to illustrate calculations on pp. 17-18, 37-38, and 42-43, taking up just about three pages of equivalent text out of twenty-seven.

McCloskey's attack is directed at an illustration of how to calculate a  $t$ -test for the hypothesis that an observed observation differs more than randomly from a predicted observation. Having estimated a simple linear relationship between traffic casualties and vehicle registrations, Johnston first shows how to compute a forecast and its confidence interval (*pace* McCloskey, who claims that economists seldom compute confidence intervals – see section 2.3).<sup>18</sup> In a passage that McCloskey chose not to quote, Johnston is careful to indicate the limitations of a simple illustration:

It should perhaps be emphasized that in a real situation one would not make a simple prediction from the statistical relationship unless a thorough examination of relevant factors such as road and travel developments, the nature and type of vehicles, and the uses to which were to be put seemed to indicate the complex of accident-producing forces to be substantially unchanged. [Johnston 1972, p. 42]

The hypothesis test that draws McCloskey's fire is suggested by the question implied in this passage: has the relationship changed? Johnston's example supposes that both sterner penalties for traffic violations (not a factor in his regression equation) have been introduced and that the number of vehicle registrations is 8 million, while actual casualties are 270,000. Johnston asks: "Does this indicate a significant change in the relationship or might these figures easily come from the same relationship as before?" (Johnston 1972, p. 42)<sup>19</sup> This is clearly a sensible question. An estimated relationship can supply an economically relevant prediction only if the data in the prediction period can be characterized by the same statistical model as those in the estimation period. His  $t$ -test rejects constancy at a critical value of 5 percent, but not at 1 percent. His conclusion that it "is suggestive of a reduction, being significant" (p. 43) implies in

---

<sup>18</sup> Johnston's estimated equation is  $casualties = 55.85 + 0.321registrations + error$ , with an estimated standard error of regression of 11.18.

<sup>19</sup> Although Johnston omits the adjective "statistical" before "significant," there is no confusion about his meaning. (McCloskey (1998, p. 117) herself glosses Johnston's answer to this question as referring to *statistical* significance).

context not a policy conclusion, and particularly not one that sterner penalties are ineffective, but a framing of the evidence: to believe that the actual data are compatible with the original model, one must believe that the last data point is one of those rarer than a 1/20 but not so rare as 1/100 draws from the original model.<sup>20</sup>

McCloskey and Ziliak (1996, p. 100) are quite right that, in this example, Johnston does not consider the human cost of ignoring sterner penalties. Why should he? The example is a cooked one with no particular relevance to the world, and it is offered principally to illustrate calculations that would in fact be informative in a real case. Yet, McCloskey and Ziliak (1996, p. 100) tax Johnston with failing to acknowledge that "the 100,000 lives that were saved in the reduction as measured . . . as 'significant.'"

Beyond the fact that this magnitude is beside Johnston's point, it is based on an apples-and-oranges comparison. Johnston's data ends in 1957. In one example, going beyond the data, he supposes that vehicle registrations rise to 10 million (from 7.43 million in 1957); and, in the example in question, he considers them to be 8 million (with no indication of whether this is an increase or decrease). If the two examples are distinct, both refer to a hypothetical 1958, and actual casualties (not *deaths* as McCloskey and Ziliak imply) have fallen by 4,000 in the second case (since they had been 274,000 in 1957). If the second case is meant to follow directly from the first (so that it refers to a hypothetical 1959), then we do not know how much actual casualties have fallen, because Johnston does not supply the actual casualties for the first case.

The only way that we can see to get McCloskey and Ziliak's 100,000 is to compare the *predicted* value when vehicle registrations are 10 million to the actual value

---

<sup>20</sup> With reference to Johnston's conclusions, McCloskey and Ziliak (1996, p. 100) note that "[s]tatistically [his] usage is unobjectionable," an honest and charitable caveat that is completely omitted in McCloskey's (1998, p. 117) later paraphrase of the same discussion.

when they are 8 million ( $367,850 - 270,000 = 97,850 \approx 100,000$ ). But that number has no bearing whatsoever on how much sterner penalties might reduce casualties. The fall in vehicle registrations of 2 million itself accounts for 62,400 ( $= 0.321 \times 2$  million) of the decrease in casualties, and these cannot be attributed in any way to sterner penalties.

The correct comparison would be between the number of casualties conditional on 8 million vehicles when penalties were lighter and the number when they were sterner – that is, between the prediction of the original model and those of a model characterizing the new regime.<sup>21</sup> Unfortunately, with only one data point, we cannot estimate the second model. The next best thing is to ask whether the actual data could have been generated under the first regime. The relevant comparison is then between the predicted number of casualties when penalties were lighter and the actual number when they are sterner, both conditional on 8 million vehicles. Johnston himself reports this difference as a reduction of 35,450. His *t*-test addresses the question of whether it is likely that the original model could throw up a value that large through random variation. Such a test does not provide definitive evidence – nor does Johnston suggest that it does – but it is relevant in that it highlights that one must maintain a very strict control against type I error to reject the view that the difference is a genuine effect and not merely a random fluctuation.

---

<sup>21</sup> McCloskey and Ziliak's (1996, p. 100) remark that Johnston uses the "universe of road casualties in the United Kingdom 1947-1957 as though it were a sample of size 11 from some universe." But, as we saw in section 2.4, unless the data can be seen as, at least potentially, a sample from a larger population (here particularly the years after 1957), such hypothetical scenarios (the type that McCloskey advocates with her remarks in favor of simulations) would be pointless, because it is only if there is a continuity of mechanism that implies some homogeneity, so that the original sample and the hypothetical predictions belong to a common universe that anything, including an estimate of casualties prevented can be legitimately gleaned from regression. If the 1947-1957 time series are *sui generis*, then McCloskey's claim that the data support an estimate of many lives saved is utterly without basis.

McCloskey is passionate in his denunciation of Johnston's (and by extension most econometricians) humanity:

He turns away from the human significance of what he has imagined and concludes from his table of Student's-*t* that thousands of causalities avoided after all might not be "significant" if you, the researcher, sitting in your warm room with a nice cup of tea, wish strongly enough not to *over* estimate the effect, considering the small sample.

. . .

It will strike outsiders to the sandbox game of statistical significance as incredible that Johnston should leave such an impression, that thousands of casualties are as nothing beside the embarrassment of a researcher who might overestimate the effect of sterner penalties. But that is how the procedures lead the econometrician to act. [McCloskey 1998, p. 117]

Yet, no fair-minded or charitable reading of Johnston could reach the conclusion that he advocates ignoring thousands of casualties. The issue is simply not germane to his narrow purpose. Still, if anything, the statistics he presents would most often be interpreted as favoring some effect of sterner penalties, since a 5-percent size is more commonly used than a 1 percent size with so small sample. But how large an effect? There is no way to get a good answer. The best answer would require estimation under the alternative hypothesis, which would be possible only with more data. But the actual data do convey a useful message: it seems likely that sterner penalties have a measurable effect; so wait for enough information to become available (maintaining the sterner policy) before reaching an economic or policy conclusion about its efficacy.

#### **4. Are Economists Confused About Statistical Significance?**

##### **4.1 THE DESIGN AND IMPLEMENTATION OF THE TWO SURVEYS**

If the evidence that econometrics textbooks actually teach error is thin, are economists nonetheless confused or negligent with respect to statistical significance in their practice?

McCloskey and Ziliak address this question with their two surveys of papers in the *American Economic Review* in the 1980s and the 1990s. The questions in Table 1 are a hodge-podge and far broader than the original, narrow one: do economists mistake economic and statistical significance? Although a case might be made for some of the broader practices surveyed contributing to the primary mistake of confusing economic and statistical significance, the connection is often tangential.

McCloskey and Ziliak report the results of the survey of the 1980s as percentages answering "yes" to each question. That is unexceptionable. But the 1990s survey is more ambitious. Each author is scored by the number of affirmative answers to the survey questions, and author's grouped into ranges from best to worst practice (Ziliak and McCloskey 2004a, Table 4).<sup>22</sup> Scoring in itself implies interpersonal comparisons among authors, and Ziliak and McCloskey (2004a) do not flinch from intertemporal comparisons as well: "significance testing is getting worse" (pp. 527, 529). Such comparisons immediately raise methodological questions.

What is the metric on which authors are compared? Even to add up a score is to imply that each question bears an equal weight. Yet as we have seen in earlier sections, the questions themselves reflect a hodge-podge of practices: some indicate truly good practices (e.g., questions 4, 13, 19), although not ones that can be judged without interpretation; others reflect personal preferences that may be desirable but do not have any *logically necessary* connection to the confusion of economic and statistical

---

<sup>22</sup> It is unclear why Ziliak and McCloskey chose to group authors in ranges rather than report individual scores.

significance (e.g., questions 2, 3, 5); others describe practices that are desirable in some contexts but not others (e.g., questions 6, 8, 10, 11, 17); while others – depending on how they are interpreted in practice – may not reflect good practice at all (e.g., questions 1, discussed in section 2.2 above, and 14, discussed in section 2.5 above). In any case, they cannot be reasonably equally weighted.

And some questions are redundant, which permits doubling counting and, therefore, unequal weighting of an arbitrary kind. To ask whether an author confuses economic and statistical significance at the “crescendo” of the article (question 7), after the “crescendo” (question 15), and in the conclusion (question 18) is akin to asking whether he makes the error on p. 1, p. 2, p. 3. . . p.  $n$  and then treating each as an independent error, turning one mistake into several. The error may be a particularly important one, but the implicit weighting method bears no stable relation to its importance.

McCloskey and Ziliak conclude from their surveys not just that the authors of the articles in the *American Economic Review*, but that the “supermajority of economists” (McCloskey and Ziliak 2004b, p. 668) misunderstand the difference between statistical and economic significance (cf. McCloskey 1999, p. 361; 2002, p. 52; McCloskey and Ziliak 1996, p. 111; Ziliak and McCloskey 2004a, p. 530). And that the problem is getting worse. Given McCloskey’s objections to “samples of convenience” and skepticism about hypothetical populations, such comparisons and extrapolations are illegitimate by McCloskey’s own lights. If tests of statistical significance cannot be applied to 19<sup>th</sup>-century data on purchasing power parity, because the 19<sup>th</sup> century comprises the whole population, while the 20<sup>th</sup> century comprises a different population,



then meaningful comparisons between the practices of economists in the 1980s and 1990s are equally out of court. The point does not hinge on statistical significance itself, but on whether two groups of data can be situated as parts of larger populations – ones that may have as yet unobserved members. McCloskey denies the validity of the comparison abstractly, and yet in practice she contradicts her professions and draws inferences from the comparison. She can hardly help doing so, since the point is supposed to be about the underlying behavior of the economics profession, the mechanism – for which the raw facts are merely evidence not an end in themselves.

If such comparisons are legitimate, then questions arise: is an increase in the number of authors assigned “no” for question 16 from 70 percent in the 1980s to 82 percent in the 1990s a genuine increase in the underlying probabilities of committing an error or simply random variation attributable to unstable and uncontrollable factors? This is, of course, exactly the sort of question that motivated Edgeworth and others to develop tests of statistical significance in the first place.

The questions in the two surveys for the most part involve a subjective element. What is the “crescendo” of an article (question 15)? What constitutes a “conversation stopper” (question 16)? What is a “proper” null (question 4)? When are coefficients interpreted “carefully” (question 5)? Subjectivity alone does not rule out scientific or reproducible procedures. Psychologists and sociologists have considered the problem of how to “objectify” subjective observations (Eye and Mun 2004) of, for example, conversations between husbands and wives. Generally, these procedures require a period of training and a calibration between different observers or for the same observer operating at different times. Clearly, consistency in reading and scoring different articles

would be critical in any such survey – especially if the data from a later survey are to be compared sensibly to those of an earlier one.

The group of questions that deal directly with the confusion of economic with statistical significance is particularly problematic. Is an author scored as making an error if, in the same sentence (or paragraph) that he indicates statistical significance, he does not also refer to economic significance? Is he automatically scored as making an error if he omits the adjectives statistical or economic in front of significance? The only way to answer such questions would be to know what protocols were followed in scoring the papers. And the best way to know such protocols would be have the records of the mapping that indicate precisely which passages in the text warrant particular judgments with respect to each question.

Unfortunately, Ziliak has informed us that such records do not exist. And McCloskey and Ziliak declined our requests to reconstruct these mappings retrospectively for a random selection of the articles (e-mail McCloskey to Hoover 11 February 2005). Absent such information, including any description of procedures for calibrating and maintaining consistency of scoring between the two surveys, we cannot assess the quality of the scoring or the comparability between the surveys.

McCloskey's reason for not sharing the mappings appears to be, first, that they are they are utterly transparent and, second, that the relevant information is contained in the scores themselves:

Think of astronomers disagreeing. We have supplied you with the photographic plates with which we arrived at our conclusions [i.e., the question-by-question scores for the 1990 survey]. . . The stars [i.e., the articles in the *American Economic Review*] are still there, too, for you to observe independently. [e-mail McCloskey to Hoover 19 February 2005]

It is surprising that a professor of English (as McCloskey is in addition to being a professor of economics), who should be sensitive to the wiles of language, could believe that the interpretation of texts, even of economics texts, is mechanical and not subject to variations in judgment and perspective that would require regulation and calibration to be consistent across different observers and that would remain debatable in all cases.

Take, for example, the apparently straightforward question 8: "Does the paper mention the power of the tests?" McCloskey comments, "[i]t means . . . uh . . . mentioning power. Type I, Type II, you know, that sort of stuff" [e-mail McCloskey to Hoover 19 February 2005]. Is it enough that the word "power" appear? Presumably, only if it refers to statistics. But what about variants such as "forecasting power" that usually do not refer directly to statistical tests, but are closely allied to them, such that they could be recast as references to type II error? What about circumlocutions that address the issues of type II error without actually mentioning it or power directly? These are questions of interpretation that are best revealed from the detailed mappings and that are not revealed at all from the scores.

The necessity of interpretation shows that the analogy with astronomy is utterly false. The scores are the equivalent of an astronomer making rather refined calculations taking photographic plates of the stars as the raw data. These plates, naturally, require interpretation. The astronomer then publishes the calculations in a table, reaching a conclusion important for astronomy. Another astronomer, who wants to understand more fully the published results, asks the first astronomer to see the plates. To which the first astronomer replies, "Why, I have shown you my tables of data – they are in the published paper – and the heavens themselves are in full view. What more could you need?" But,

of course, that misses the point. The plates, not the heavens, are the astronomer's data. And while there could be interest in another astronomer doing it all from scratch, there is a perfectly reasonable scientific interest, and a much honored scientific tradition, in seeing the data that was actually generated – that is, the plates (analogously the mappings from texts to scores) themselves.

#### 4.2 CASE STUDIES

The substantive difficulties of interpretation can be made more vivid through in depth case studies. We have examined every article that McCloskey and Ziliak (1996) and Ziliak and McCloskey (2004a) discuss in the body of their texts over above scoring for their surveys. In the interest of space we comment on five – two from the survey of the 1980s and three from the 1990s. We focus on the main issue: does the author confuse statistical and economic significance? Our choices are not random; they reflect articles that we believe illustrate clearly the flaws in McCloskey and Ziliak's analysis. But neither are they intentionally biased. McCloskey and Ziliak single each one out for special criticism, and each one scores low in their surveys.<sup>23</sup> Our approach is holistic in the sense that we concentrate on what a reasonable reader would take away from the paper at the end of the day and not, for example, on whether or not the adjectives "statistical" and "economic" qualify "significance" in particular instances.

---

<sup>23</sup> For the 1980s, there are 182 articles. Woodbury and Spiegelman (1987) scored 7 out of 19 and ranked the 41<sup>st</sup> percentile of papers in the 1980s (i.e., 41 percent scored 7 or less). Darby (1984) scored either 2 or 3 and is in the 8<sup>th</sup> or 4<sup>th</sup> percentile (the latter a seven-way tie for last place). The ambiguity arises because McCloskey and Ziliak's score sheet reports 3 yes and 17 no for Darby's article, but there are only 19 questions. For the 1990s, there are 137 articles. Bernanke and Blinder (1992) scores 8 (58<sup>th</sup> percentile); Becker, Grossman, and Murphy (1994) scores 6 (30<sup>th</sup> percentile); Bernheim and Wantz (1995) scores 1 (last place and the 1<sup>st</sup> percentile).

Darby (1984): "The U.S. Productivity Slowdown: A Case of Statistical Myopia"

Of Michael Darby's article, McCloskey and Ziliak (1996, p. 104) write:

The misuse in Michael Darby (1984) is balder [than in Romer and Sachs 1980]: his only argument for a coefficient when he runs a regression is its statistical significance (pp. 311, 315), but on the other hand his findings do not turn on the regression results.

Darby (1984, p. 301) hypothesizes that the sharp slowdown in measured productivity in the 1970s is the result of demographic changes that, when accounted for appropriately, eliminate any secular decline in the growth rate of technical progress. Darby comments on the (statistical) significance of regressors in his equation (11) and performs  $F$ -tests of various sets of zero restrictions (Table 6, p. 311). This specification exercise suggests his final specification (equation (12), p. 311). Darby uses statistical significance as a measure of the precision of his estimates (for example, as indicating that "no direct effect [of oil-price increases on productivity] is directly detectable" (p. 311), but nowhere does he confuse the measures of statistical significance with the economic meaning of the equation. For example, immediately following equation (12) he considers

the implications of this equation for the level of productivity in the year 1973. The average value of  $CD_t$  [a price-control variable] in 1973 is 0.7857 which . . . implies that the logarithm of labor productivity in 1973 was overstated by 0.0369. This means that the 1965-73 growth rates of private labor productivity are *overstated* by  $3.69/8 = 0.46$  percent per annum and correspondingly that the 1973-79 growth rates are understated by  $3.69/6 = 0.61$  per annum. . . applying this adjustment to the quality-adjusted growth rates . . . eliminates any evidence of a major 1973-79 productivity slowdown. [p. 311]

Whether the argument is right or wrong, the reasoning is about the economic size of the coefficients as they relate to the question of interest, the productivity slowdown; and, far from not turning on the regression results, they depend in detail on the coefficient

estimates. Statistical significance is not invoked as settling the question of economic interest.<sup>24</sup>

Woodbury and Spiegelman: "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois"

Quoting the following passage, McCloskey and Ziliak (1996, pp. 107; cf. McCloskey 1998, pp. 132-134) accuse Stephen A. Woodbury and Robert G. Spiegelman (1987) of "not thinking about the economic meaning of a coefficient":

The fifth panel also shows that the overall benefit-cost ratio for the Employer Experiment is 4.29, but it is not statistically different from zero. The benefit-cost ratio for white women in the Employer Experiment, however, is 7.07, and is statistically different from zero. Hence, a program modeled on the Employer Experiment also might be attractive from the state's point of view if the program did not increase unemployment among nonparticipants. Since, however, the Employer Experiment *affected* only white women, it would be essential to understand the reasons for the uneven effects of the treatment on different groups of workers before drawing conclusions about the efficacy of such a program. [Woodbury and Spiegelman 1987, p. 527; emphasis added]

McCloskey and Ziliak (1996, p. 108) gloss "affected" as

the estimated coefficient is statistically significantly different from a value the authors believe to be the relevant one. The 4.29 benefit-cost ratio for the whole Employment Experiment is, according to the authors, *not useful or important for public policy*. The 7.07 ratio for white women is said to "affect" – to be important – because it passed an arbitrary significance test. . . The argument that the 4.29 figure does not "affect" is unsound, and could be costly in employment forgone.

While we would agree that "affect" is a poorly chosen word – at least for those who are intent on finding a confusion between economic and statistical significance – we find no evidence in this passage of a mistake in logic. Much of the conclusion of Woodbury and Spiegelman's (1987, pp. 528-529) article is devoted to the *economic*

---

<sup>24</sup> An additional reference to statistical significance (Darby 1984, p. 315) essentially states that a change in specification did not result in an estimate of the coefficient on *CD*, outside its confidence interval in equation (12), therefore not triggering any reassessment of the *economic* interpretation of that equation.

significance of their coefficient estimates. For example: "The results of the Claimant Experiment are unequivocal and strong. The incentive created by the \$500 bonus . . . reduced state regular benefits . . . by an average of \$158, and reduced . . . insured unemployment by more than one week . . ." (Woodbury and Spiegelman 1987, p. 528). The gravaman of McCloskey and Ziliak's charge is that Woodbury and Spiegelman dismiss the economic importance of the overall estimate. But that conclusion relies on an uncharitable, decontextualized, and wooden reading of the article.

Woodbury and Spiegelman present estimates of both the overall ratio and that for four component groups (white women, white men, black women, and black men). The difference between the insignificant overall ratio and the significant ratio for white women suggests that either the number of experimental subjects in three of the categories is too small to get a precise estimate or that they belong to distinct populations. Would McCloskey and Ziliak have them apply the overall ratio to white women, because they are part of the whole, even though the specific ratio for white women is much more precisely estimated and economically quite different from the overall ratio? Or would McCloskey and Ziliak divide the population into its component groups and rely on the individual group estimates of benefit-cost ratios? The first choice simply ignores the evidence that the data are not drawn from a homogeneous population; the second commits the licorice-jelly-bean mistake of section 2.2.

Woodbury and Spiegelman instead recognize that a) the data are not homogeneous and b) some of the components are too noisily measured to draw firm conclusions. They do not (*pace* McCloskey and Ziliak) use statistical insignificance to advocate ignoring these groups for policy purposes. Instead, they note the need for

further study: "it would be essential to understand the reasons for the uneven effects of the treatment on different groups of workers [estimated benefit cost-ratios that differ by factors as large as 44 as well as different levels of statistical significance] before drawing conclusions about the efficacy of such a program" (Woodbury and Spiegelman's 1987, p. 527) While they do not provide additional analysis of this issue, earlier they provided a careful *economic* – not statistical – analysis of the difference of the differences by race and sex in the receipt of, qualification for, and take-up rates of bonuses, which illustrates the type of analysis that they probably have in mind (Woodbury and Spiegelman 1987, pp. 523-527). The authors provide abundant evidence of sensitivity to the distinction between economic and statistical significance, and do not make the specific mistake that McCloskey and Ziliak attribute to them.

Bernanke and Blinder (1992): "The Federal Funds Rate and the Channels of Monetary Transmission"

By "asterisk econometrics" (question 10), McCloskey and Ziliak (1996, p. 103) mean "ranking the coefficients according to the absolute size of  $t$ -statistics." Ziliak and McCloskey (2004a, p. 534) interpret the question more broadly as covering any hierarchy or grouping of  $p$ -,  $F$ -, and  $t$ -statistics. Surely, the question is not how this information is conveyed, but what information is supplied and what is needed. The mere fact of presenting a ranking does not imply a confusion of economic and statistical significance. Frequently, the information included in asterisked tables is rich enough to permit readers to examine the data more fully. For example, a common form presents a coefficient estimate and either a  $t$ -statistic (against the null of zero) or a standard error. From this information, one can construct confidence intervals (favored by McCloskey) or tests with



other nulls (see section 2.3 above). If an author also indicates significance through asterisks,  $p$ -values, or other means, the additional information cannot be bad in itself, but only if it is misused.

Is it misused? Ziliak and McCloskey (2004a, p. 234) cite Ben S. Bernanke and Alan S. Blinder (1992, pp. 905 and 909) as clear practitioners of "asterisk econometrics." But what we see is something different. Bernanke and Blinder conduct Granger-causality tests, which ask whether the lagged values of a variable contain incremental information about another variable. The natural test is  $F$ -test that measures whether the exclusion of the lags of the first variable results in a statistically significant increase in the standard error of regression. The statistical evidence that one variable Granger-causes another does not imply that it is an important determinant, but only that its signal rises measurably above the noise according to a conventional threshold. The question that Granger-causality addresses, whether a conditional distribution is different from a marginal distribution cannot even be addressed through coefficient estimates without standard errors; for it is a question about the statistical distribution, not about the mean. The tables that Ziliak and McCloskey single out do not present  $F$ -statistics in a hierarchy (they are not rank ordered, and there are no asterisks), but they do present evidence relevant to Granger-causality. What is more, Bernanke and Blinder do not confuse economic and statistical significance in interpreting this evidence. Their table is part of an analysis that is ultimately evaluated in terms of an economically meaningful standard: the fraction of the variance of various real variables that is predicted by various currently observed variables (see their Tables 2 and 4).

Bernheim and Wantz: "A Tax-Based Test of the Dividend Signaling Hypothesis"

Ziliak and McCloskey (2004a, p. 535) also accuse B. Douglas Bernheim and Adam Wantz (1995) of practicing asterisk econometrics. Yet the table that they cite of Bernheim and Wantz (1995, p. 547) does not rank its test statistics, and the only asterisk indicates an alternative scenario and not a significance level. The question that Bernheim and Wantz address is again not one of coefficient values, but of the timing of regime changes, for which a test of whether data from possibly different regimes could have been generated by the same model is appropriate. The interest in the timing of regime changes is perfectly consistent with a concern for *economic* magnitudes, which they address with the "bang-for-buck" ratios calculated later in the paper.

In addition to asterisk econometrics, Ziliak and McCloskey (2004a, p.534) also accuse Bernheim and Wantz (1995) of practicing "sign econometrics" (question 11); that is "remarking on the sign but not the size of the coefficient. . ." (Ziliak and McCloskey 2004a, p. 534; cf. McCloskey and Ziliak 1996, p. 103). Contrary to McCloskey and Ziliak, "sign econometrics" is not an error in every case. If a coefficient is badly measured, then its sign is also badly determined.<sup>25</sup> But if it is well measured, the sign sometimes matters as well as, or more than, the absolute magnitude. Whether I am a minute off the departure time for my train is much more costly if I am late than if my error takes the opposite sign and I am early. An error occurs only if an inference is made that depends on the absolute magnitude instead of, or in addition to, the sign. The

---

<sup>25</sup> Ziliak and McCloskey's position is not fully consistent. We agree that "a poorly fit correlation with the expected sign would say nothing" (Ziliak and McCloskey 2004a, p. 539). Yet they intended the inaccurately reported aspirin example to show that the size of the measured effect matters even when estimates are statistically insignificant. And if the "oomph" matters, so does its direction. McCloskey would surely advocate a different policy if the correlation between aspirin and heart attacks were positive, though statistically insignificant.

objection to “sign econometrics” stands in an uneasy relationship to question 5 (Table 1), which demands theoretical interpretation of coefficients. Theory often has something definite to say about direction, while agnostic on absolute value.

To illustrate “sign econometrics,” Ziliak and McCloskey (2004a, p. 534) quote Bernheim and Wantz (1995, p. 543) who “report that ‘the coefficients [in four regressions on their crucial variable, high-rated bonds] {Ziliak and McCloskey’s interpolation} are all negative . . . However, the estimated values of these coefficients,’ they remark, ‘are not statistically significant at conventional levels of confidence.’” The quotation appears to support Ziliak and McCloskey’s point only because it is selective and taken out of context. Bernheim and Wantz (1995, p. 543) make only a passing remark on the signs of the coefficients from which no inferences are drawn; and, in the sentence immediately following the quoted one, they say, “More significant effects – both *economically* and statistically – are found in Table 2” (emphasis added). Two paragraphs later they discuss the *economic* magnitudes of the precisely estimated bang-for-buck ratios derived from their estimates.

Becker, Grossman, and Murphy (1994): “An Empirical Analysis of Cigarette Addiction”

Ziliak and McCloskey (2004a, p. 540) are particularly proud to criticize Gary S. Becker, Michael Grossman, and Kevin M. Murphy (1994) :

you can see that we are anxious not to be accused of making our lives easy by picking on the less eminent economic scientists. . . Sign econometrics and asterisk econometrics decide nearly everything in the paper, but most importantly the “existence” of addiction.

Once again, the charges are unsupported by the evidence. The central analysis is about the *economic*, not the statistical, significance of the data. Right in the introduction, Becker *et al.* (1994, pp. 396-397) summarize their findings:

a 10-percent increase in the price of cigarettes reduces current consumption by 4 percent in the short run and by 7.5 percent in the long run. In contrast, a 10-percent increase in price for only one period decreases consumption by only 3 percent. These estimates illustrate the importance of the intertemporal linkages in cigarette demand implied by addictive behavior.

Not a word about statistical significance. The only way in which this passage falls short of what McCloskey advocates is that the authors do not add, "And, wow, that is a big effect!"

Becker *et al.*'s primary goal is to distinguish between two hypotheses: myopic addiction and rational addiction. Their claim is that the key empirical distinction is between a positive and a zero coefficient on the future consumption term. Here sign, rather than size, is of the essence. Statistical significance matters because the authors will place faith only in evidence garnered from measurements that rise above the statistical noise. Their purpose is a scientific one (is one theory more likely to be correct or not?) rather than an engineering one (what can I do with this measurement?). But (*pace* McCloskey and Ziliak) they indicate clearly the relevance of their results for government regulation and revenues (p. 397), leaving it to the policymakers to work out the details in relationship to their own ends. What is big depends on the purposes, but the facts about the world are what they are whatever one's purposes. The scientist tries to point to the facts; the policymaker must use them.

Ziliak and McCloskey's (2004a, p. 541) central claim that Becker *et al.* confuse economic and statistical significance is simply false, although they admit that

“[e]ventually [Becker *et al.*] report (though they never interpret) the estimated magnitudes of the price elasticities of demand for cigarettes” (p. 541).<sup>26</sup> Yes, Becker *et al.* use tests of statistical significance to measure the precision of estimates. But with respect to their main empirical results (reported in their Table 3), they immediately follow it with the calculation of elasticities and a careful discussion of what they imply for the effects of changes in cigarette prices (Becker *et al.* 1994, p. 407). This discussion is entirely *economic* with no further references to statistical significance.<sup>27</sup> It is this economic evidence that is used to assess the competing theories:

Clearly, the estimates indicate that cigarettes are addictive, that past and future changes significantly impact current consumption. This evidence is inconsistent with the hypothesis that cigarette consumers are myopic. Still, the estimates are not fully consistent with rational addiction, because the point estimates of the discount factor ( $\beta$ ) are implausibly low . . . correspond[ing] to interest rates ranging from 56.3 percent to 226.6 percent. [Becker *et al.* 1994, p. 407]

Since the estimates referred to here are those of the immediately preceding paragraph, which do not refer to tests of statistical significance, “significantly” is naturally interpreted here as referring to the magnitude of the effect and not its relationship to its standard error. That Becker *et al.* take these *economic* magnitudes to be the relevant ones is reinforced by the implications for interest rates, the calculations of which rely on point estimates and which are judged by their economic plausibility. In what sense, then, can Becker *et al.* be said to “never interpret” their estimates?

One hint of an answer is Ziliak and McCloskey’s (2004a, p. 541) assertion,

“[c]igarette smoking may be addictive. But Becker, Grossman, and Murphy have not

---

<sup>26</sup> Ziliak and McCloskey (2004a, p. 541) continue: “But their way of finding the elasticities is erroneous.” The charge is never substantiated by pointing out the particular errors. Be that as it may, whether Becker *et al.* are right or wrong on this point has no bearing on the central question of whether they confuse economic and statistical significance.

<sup>27</sup> A similar discussion of a different specification, couched in economic terms, is found on p. 408. And, despite Ziliak and McCloskey’s having scored Becker *et al.*, as not conducting any simulations, a small simulation is reported on p. 409.

shown why, or how much." If "why" is taken to mean what mechanism causes smokers to be addicted, of course they have not shown why. That is a physiological question, and not one that they sought to address. Their article addresses a different question: can the behavior of smokers be seen as consistent with a myopic or a rational model of addiction? Whether or not their conclusions are correct, the evidence that they muster does in fact bear on this question. If "how much" is taken to mean some measure of the intensity of the addiction, they leave that measurement implicit in their estimates. For that too does not have direct bearing on the question of interest. They do draw out the implications of their estimates for the discrimination between competing models and for elasticities relevant to public policy. They clearly provide an economic interpretation of the significance of their estimates.

### **5. Apocalypse Now?**

McCloskey's analysis of the state of significance testing in economics is apocalyptic: "Econometric testing, as against estimation, is not worth anything at all. Its marginal product is zero" (McCloskey 1997, p. 242). "Almost all econometric fittings have to be done over again. All of them. . . all the work of the elders has been wasted. . . "Eminent statisticians and many econometricians declare statistical significance to be bankrupt. Yet scientific practice does not change at all." (McCloskey 1997, p. 361). And her advice has the direct simplicity of the Old Testament prophet: "You go figure. But when figuring don't use statistical significance" (McCloskey 1997, p. 361).

Our message is this: Take comfort, things are not so dark as all that.

McCloskey's analysis makes two claims: first, the practice of significance testing is unsound; second, the economics profession buys into those practices and is, as a result, hopelessly confused. Her evidence on both counts is utterly unpersuasive.

We agree that economic significance is quite distinct from statistical significance. But this is a banality, and nowhere does McCloskey point to a single economist who offers a principled, reflective defense of the contrary proposition. The other myriad errors that McCloskey attributes to statistical significance testing remain unproven.

Statistical significance is a measure of the strength of a signal relative to background noise. McCloskey's slogan "size matters" refers to the fallacious conclusion that we ought to be guided by noisy measurements as if there were no noise at all. But experience shows, for example, that, when pilots act as decisively on noisy signals as on clear ones, airplanes sometimes crash with spectacular loss of life.

McCloskey believes that statistical significance has limited applicability only in a Neyman-Pearson framework. We do not object to the Neyman-Pearson framework for those problems to which it can be effectively applied. But "no science without a loss function" is a slogan, not an argument. In fact, McCloskey offers no argument, beyond fallacious appeals to authority, for the Neyman-Pearson framework as a *generally applicable* methodology of science, nor does she provide an example of how any large-scale scientific advance was achieved through an appeal to loss functions or other formal decision-theoretic devices. Yes, scientists must make decisions. Sometimes small problems can be cast in a tight Neyman-Pearson decision-theoretic framework. But at all times, scientists require observations that rise above the noise, and statistical significance tests are a tool for assessing such observations.

Economics is mainly an observational, as opposed to an experimental, science. In arguing against viewing economic data as samples from larger (perhaps hypothetical) populations, McCloskey would deny the tools of probability to economists. Yet, the great achievement of econometrics since Haavelmo is precisely to understand that, with appropriate transformations, non-experimental data can be cast into a framework to which tractable probability models apply. Specification testing is essential to establishing the success of the statistical model. McCloskey prefers estimation to testing. But all estimators work within a statistical model. Without specification tests, the statistical model is left implicit and not subject to investigation, validation, or debate. As Kruskal put it, to call on economists to abjure statistical significance testing in this case is "ridiculously rigid."

Surprising for a self-professed humanist, McCloskey appears to suffer from physics envy. She is deeply worried that an economics that uses significance tests is not a true science. Of course, that other sciences do not use significance tests is not a compelling argument, even if it were true – different tools for different problems. But it is simply false: physical sciences, as well as biological, medical, and other social sciences, all use significance tests to a greater or lesser degree.

Setting aside all the broader arguments about significance testing, the second element of McCloskey's analysis is the most prominent. Individual economists – and especially students – may sometimes through carelessness or misunderstanding mistake statistical for economic significance. But again, we deny that there is any persuasive evidence that economists on the whole are confused about the distinction.



The charge that econometric pedagogy inculcates and perpetuates this mistake is mostly based on a highly selective reading of econometrics and statistical texts, focusing on what they do not say. None advocates the mistake as sound practice. The only evidence that McCloskey offers for positive miseducation is based on three pages of Johnston's once-popular econometrics text. We have shown that McCloskey charges of error in Johnston's text are utterly false, relying not only on selective reading but on blatant misinterpretation.

"And the economists, oh, the economists, how they counted, and still count" (McCloskey 2002, p. 6). McCloskey is herself no exception. Yet, the two surveys of the *American Economic Review* fail to convince. They are poorly designed. We were unable to obtain the details of their implementation – particularly with regard to how interpretive decisions were made and what protocols (if any) were used to calibrate them and maintain consistency of scoring within each survey and between the two. Detailed comparison of McCloskey and Ziliak's analysis of selected individual articles to the original texts shows selective and highly inaccurate readings that undermine any confidence that the scoring of the surveys themselves is at all reliable.

In reaching these judgments, we are not saying that applied econometrics is always done correctly nor do we defend the particular articles against all criticism. There may be problems. Some work is more persuasive than others. In some cases, widely used methodologies may be suspect. Yet, McCloskey's blanket condemnation of statistical significance testing misses the mark and contributes next to nothing to constructively addressing the shortcomings of applied econometrics.

## References

- Becker, Gary S., Michael Grossman, and Kevin M. Murphy. (1994) "An Empirical Analysis of Cigarette Addiction," *American Economic Review* 84(3), 396-418.
- Bernheim, B. Douglas and Adam Wantz. (1995) "A Tax-Based Test of the Dividend Signaling Hypothesis," *American Economic Review* 85(3), 532-551.
- Bernanke, Ben S., and Alan S. Blinder. (1992) "The Federal Funds Rate and the Channels of Monetary Policy," *American Economic Review* 82(4), 901-921.
- Blank, Rebecca M. (1991) "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence in the *American Economic Review* 81(5), 1041-1067.
- Chen, Y.T., Alan H. Cook, and A.J.F. Metherell. (1984) "An Experimental Test of the Inverse Square Law of Gravitation at Range of 0.1 m," *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, 394(1806), pp. 47-68.
- Cooley, Thomas F. and Stephen F. LeRoy. (1981) "Identification and Estimation of the Money Demand," *American Economic Review* 71(5), 825-844.
- Darby, Michael R. (1984) "The U.S. Productivity Slowdown: A Case for Statistical Myopia." *American Economic Review* 74(3), 301-322.
- Economist*. (2004) "Signifying nothing?" *The Economist*, 29 Jan 2004,
- Edgeworth, Francis Ysidro. (1885) "Methods of Statistics," *Jubilee Volume of the Statistical Society*. Royal Statistical Society of Britain, pp. 181-217.
- Elliott, Graham and Clive W.J. Granger. (2004) "Evaluating Significance: Comments on 'Size Matters,'" *Journal of Socio-Economics* 33(5), 547-550.
- Eye, Alexander von, & Mun, Eun Young. (2004) *Analyzing Rater Agreement: Manifest Variable Methods*. Mahwah, NJ: Erlbaum.
- Feynman, Richard P. (1985) *Surely You're Joking, Mr. Feynman*. New York: Norton.
- Feynman, Richard P. (1985) *QED The Strange Theory of Light and Matter*. Princeton: Princeton University Press.
- Fisher, Ronald A. (1946) *Statistical Methods for Research Workers*, 10<sup>th</sup> ed. Edinburgh: Oliver and Boyd.
- Fogel, Robert W. (1964) *Railroads and American Economic Growth: Essays in Econometric History*. Baltimore: Johns Hopkins Press.
- Gigerenzer, Gerd. (2004) "Mindless Statistics," *Journal of Socio-Economics* 33(5), 587-606.
- Haavelmo, Trygve. (1943) "The Statistical Implications of a System of Simultaneous Equations," *Econometrica* 11(1), pp. 1-12.
- Hendry, David F. (1980) "Econometrics: Alchemy or Science?," in Hendry *Econometrics: Alchemy or Science*, 2<sup>nd</sup> edition. Oxford: Blackwell, pp. 1-28.

- Hendry, David F. and Hans-Martin Krolzig. (1999) "Improving on 'Data Mining Reconsidered' by K.D. Hoover and S.J. Perez," *Econometrics Journal*, 2(2), 202-218.
- Hendry, D.F. and Krolzig, H.-M. (2004) "Automatic Model Selection: A New Instrument for Social Science," *Electoral Studies* 23, 525-544.
- Hennekens, Charles H. et al. (1988) "Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study," *The New England Journal of Medicine* 318(4), 262-264.
- Hoover, Kevin D. (1994) "Econometrics as Observation: The Lucas Critique and the Nature of Econometric Inference," *Journal of Econometric Methodology* 1(1), 65-80.
- Hoover, Kevin D. (1995) "In Defense of Data Mining: Some Preliminary Thoughts," in Kevin D. Hoover and Steven M. Sheffrin, editors. *Monetarism and the Methodology of Economics: Essays in Honour of Thomas Mayer*. pp. 242-57
- Hoover, Kevin D. (2001) *Causality in Macroeconomics*. Cambridge: Cambridge University Press.
- Hoover, Kevin D. (2004) "Lost Causes," *Journal of the History of Economic Thought*, 26(2), 149-64.
- Hoover, Kevin D. and Stephen J. Perez. (1999) "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search," *Econometrics Journal* 2(2), 167-191.
- Hoover, Kevin D. and Stephen J. Perez. (2000) "Three Attitudes Towards Data Mining," *Journal of Economic Methodology* 7(2), 195-210.
- Hoover, Kevin D. and Stephen J. Perez. (2004) "Truth and Robustness in Cross Country Growth Regressions," *Oxford Bulletin of Economics and Statistics* 66(5), 765-798.
- Horowitz, Joel L. (2004) "Comments on "Size Matters," *Journal of Socio-Economics* 3(5), 551-554.
- Johansen, Søren. (2005) "Confronting the Data," unpublished typescript, Department of Applied Mathematics and Statistics, University of Copenhagen, 1 March 2005.
- Johnston, Jack. (1972) *Econometric Methods*, 2<sup>nd</sup>. edition.
- Kahn, James A., Steven E. Landsburg, and Alan C. Stockman. (1996) "The Positive Economics of Methodology," *Journal of Economic Theory* 68(1), 64-76.
- Keuzenkamp, Hugo A. (2000) *Probability, Econometrics, and Truth*. Cambridge: Cambridge University Press.
- Keuzenkamp, Hugo A., and Jan Magnus. (1995) "On Tests and Significance in Econometrics," *Journal of Econometrics* 67(1), 103-128.
- Krolzig, Hans-Martin and David F. Hendry. (2001) "Computer Automation of General-to-Specific Model Selection Procedures," *Journal of Economic Dynamics and Control*. 25(6-7), 831-66.

- Kruskal, William H. (1968a) "Tests of Significance," *International Encyclopedia of Social Sciences*, vol. 14. David Sills, editor. Macmillan and Free Press, pp. 238-250.
- Kruskal, William H. (1968b) "Statistics: The Field," *International Encyclopedia of Social Sciences*, vol. 15. David Sills, editor. Macmillan and Free Press, pp. 206-224.
- Leamer, Edward E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. John Wiley: Boston.
- Leamer, Edward E. (1983) "Let's Take the Con Out of Econometrics," *American Economic Review* 73(1), 31-43.
- Levine, Ross and David Renelt. (1992) "A Sensitivity Analysis of Cross-Country Growth Regressions," *American Economic Review* 82(4), 942-963.
- Lovell, Michael C. (1983) "Data Mining," *The Review of Economics and Statistics*, 65(1), 1-12.
- Mayo, Deborah. (1996) *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
- McAleer, Michael, Adrian R. Pagan, and Paul A. Volker. (1985) "What Will Take the Con Out of Econometrics," *American Economic Review*, 75(3), 293-307.
- McCloskey, Donald N. (1985a) *The Rhetoric of Economics*, 1st edition. Madison: University of Wisconsin Press.
- McCloskey, Donald N. (1985b) "The Loss Function Has Been Misaid: The Rhetoric of Significance Tests," *American Economic Review* 75(2), 201-205.
- McCloskey, Donald N. (1992) "Other Things Equal: The Bankruptcy of Statistical Significance," *Eastern Economic Journal* 18(3), 359-361.
- McCloskey, Donald N. (1994) *Knowledge and Persuasion in Economics*. Cambridge: Cambridge University Press.
- McCloskey, Donald N. (1995a) "The Insignificance of Statistical Significance," *Scientific American* 272(4), 32-33.
- McCloskey, Donald N. (1995b). "Computation Outstrips Analysis," *Scientific American* 272(7), 26.
- McCloskey, Deirdre N. (1997) "Other Things Equal: Aunt Deirdre's Letter to a Graduate Student," *Eastern Economic Journal* 23(2), 241-244.
- McCloskey, Deirdre N. (1998) *The Rhetoric of Economics*, 2<sup>nd</sup> edition. Madison: University of Wisconsin Press.
- McCloskey, Deirdre N. (1999) "Other Things Equal: Cassandra's Open Letter to Her Economist Colleagues," *Eastern Economic Journal* 25(3), 357-363.
- McCloskey, Deirdre N. (2002) *The Secret Sins of Economics*. Chicago, IL: Prickly Paradigm Press. Available online at <http://www.prickly-paradigm.com/paradigm4.pdf>
- McCloskey, Deirdre N. (2005) "The Trouble with Mathematics and Statistics in Economics," unpublished typescript, University of Illinois, Chicago and Erasmus University Rotterdam, January.

- McCloskey, Deirdre N., and Stephen T. Ziliak. (1996) "The Standard Error of Regressions," *Journal of Economic Literature* 34(1), 97-114.
- McCloskey, Donald N. and J. Richard Zecher. (1984) "The Success of Purchasing Power Parity," in *A Retrospective on the Classical Gold Standard, 1821-1931*. Michael D. Bordo and Anna J. Schwartz, eds. Chicago: University of Chicago Press and the National Bureau of Economic Research.
- Mizon, Grayham E. (1995) "Progressive Modelling of Economic Time Series: The LSE Methodology," in Kevin D. Hoover, editor. *Macroeconometrics: Developments, Tensions, and Prospects*. Dordrecht: Kluwer, 1995, pp. 107-170.
- Morgan, Mary S. (1990) *The History of Econometric Ideas*. Cambridge: Cambridge University Press.
- Peirce, Charles S. (1885) "Note on the Theory of the Economy of Research," *Science and Philosophy*, volume 7 of the *Collected Papers of Charles Sanders Peirce*, Arthur W. Burks, editor, pp. 76-88.
- Peirce, Charles S. (1934) *Pragmatism and Pragmaticism*, volume 5 of the *Collected Papers of Charles Sanders Peirce*, Charles Hartshorne and Paul Weiss, editors. Cambridge, MA: Belknap Press.
- Wible, James R. (1994) "Charles Sanders Peirce's Economy of Research," 1(1), 135-60.
- Wible, James R. (1998) *The Economics of Science: Methodology and Epistemology As If Economics Really Mattered*. London: Routledge.
- Savage, Leonard J. (1954 [1972]) *The Foundations of Statistics*. New York: Dover.
- Solon, Gary. (1992) "Intergenerational Income Mobility in the United States," *American Economic Review* 82(3), 393-408.
- Staley, Kent W. (2004) *The Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation*. Cambridge: Cambridge University Press.
- J.R. Southon, J.W. Stark, J.S. Vogel, and J.C. Waddington. "Upper Limit for Neutron Emission from Cold *d-t* Fusion," *Physical Review C*, vol. 41, November 1990, pp. R1899-R-1900.
- Spanos, Aris. (1995) "On Theory Testing in Econometrics: Modeling with Nonexperimental Data," *Journal of Econometrics* 67(1), 189-226.
- Stigler, Stephen M. (1986) *The History of Statistics: Measurement of Uncertainty Before 1900*. Cambridge, MA: Belknap Press.
- Stigler, Stephen M. (1999) *Statistics on the Table*. Cambridge, MA: Harvard University Press.
- Summers, Lawrence H. (1991) "The Scientific Illusion in Empirical Macroeconomics," *Scandinavian Journal of Economics* 93(2), 129-48.
- White, Halbert. (1990) "A Consistent Model Selection Procedure Based on *m*-testing," in C.W.J Granger (ed.) (1990) *Modelling Economic Series: Readings in Econometric Methodology*. Oxford: Clarendon Press, pp. 369-383.

Hoover and Siegler, "McCloskey and Significance Testing in Economics"  
4 August 2005

Woodbury, Stephen A. and Robert G. Spiegelman. (1987) "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois, 513-530.

Woodward, James. (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

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

Ziliak, Stephen T. and Deirdre N. McCloskey. (2004a) "Size Matters: The Standard Error of Regressions in the *American Economic Review*," *Journal of Socio-Economics* 33(5), 527-546. (Also published in *Econ Journal Watch* 1(2), 331-358. Available online at [http://www.econjournalwatch.org/main/index.php?issues\\_id=3](http://www.econjournalwatch.org/main/index.php?issues_id=3))

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

## Table 1

### Survey Questions from McCloskey and Ziliak (1996) and Ziliak and McCloskey (2004a)

1. Does the paper use a small number of observations, such that statistically significant differences are not found at the conventional levels merely by choosing a large number of observations. (*Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample?*)
2. Are units and descriptive statistics for all regressions included? (*Report descriptive statistics for regression variables?*)
3. Are coefficients reported in elasticity form, or in some interpretable form relevant to the problem at hand and consistent with economic theory, so that readers can discern the economic impact of regressors? (*Report coefficients in elasticities, or in some other useful form that addresses the question of "how large is large"?*)
4. Are proper nulls specified? (*Test the null hypotheses that the authors said were the ones of interest?*)
5. Are coefficients carefully interpreted? (*Carefully interpret the theoretical meaning of the coefficients? For example, does it pay attention to the details of the units of measurement, and to the limitations of the data?*)
6. Does the paper eschew reporting all  $t$ - or  $F$ -statistics or standard errors, regardless of whether a significance test is appropriate? (*Eschew reporting all standard errors,  $t$ -,  $p$ -, and  $F$ - statistics, when such information is irrelevant?*)
7. Is statistical significance at the first use, commonly the scientific crescendo of the paper, the only criterion of "importance"? (*At its first use, consider statistical significance to be one among other criteria of importance?*)
8. Does the paper mention the power of the tests? (*Consider the power of the test?*)
9. If the paper mentions power, does it do anything about it? (*Examine the power function?*)
10. Does the paper eschew "asterisk econometrics," that is ranking the coefficients according to the absolute size of  $t$ -statistics. (*Eschew "asterisk econometrics," the ranking of coefficients according to the absolute value of the test statistic?*)
11. Does the paper eschew "sign econometrics," that is, remarking on the sign but not the size of the coefficients? (*Eschew "sign econometrics," remarking on the sign but not the size of the coefficient?*)
12. Does the paper discuss the size of the coefficients? (*Discuss the size of the coefficients?*)

continued next page

Table 1 continued

13. Does the paper discuss the scientific conversation within which a coefficient would be judged "large" or "small"? (*Discuss the scientific conversation within which a coefficient would be judged large or small?*)
14. Does the paper avoid choosing variables for inclusion solely on the basis of statistical significance? (*Avoid choosing variables for inclusion solely on the basis of statistical significance?*)
15. After the crescendo, does the paper avoid using statistical significance as the criterion of importance? (*Use other criteria of importance besides statistical significance after the crescendo?*)
16. Is statistical significance decisive, the conversation stopper, conveying the sense of ending? (*Consider more than statistical significance decisive in an empirical argument?*)
17. Does the paper ever use a simulation (as against a use of the regression as an input into further argument) to determine whether the coefficients are reasonable? (*Do a simulation to determine whether the coefficients are reasonable?*)
18. In the "conclusions" and "implications" sections, is statistical significance kept separate from economic policy, and scientific significance? (*In the conclusions, distinguish between statistical and economic significance?*)
19. Does the paper avoid using the word "significance" in ambiguous ways, meaning "statistically significant" in one sentence and "large enough to matter for policy or science" in another? (*Avoid using the word "significance" in ambiguous ways?*)

*Notes:* The questions in upright font are verbatim quotations from McCloskey and Ziliak's (1996, pp. 101-104) description of their survey instrument. The parenthetical questions in italic font are their capsule summaries from Table 1 (p. 105). Ziliak and McCloskey (2004a, Table 1, p. 529) refer only to the capsule summaries with inconsequential differences in wording from the earlier survey.



**Table 2. The Relative Use of Some Statistical Terminology in Scientific Journals**

Search for at least one statistical keyword					
and	but not	in	Hits	Total Articles	Ratio (percent)
chemistry	biology, biological, medicine, or medical	7 General Science Journals	139	1,988	7
physics		7 General Science Journals	348	4,127	8
astronomy		7 General Science Journals	80	896	9
geology		7 General Science Journals	156	1,447	11
chemistry	biology or biological	7 General Science Journals	299	2,707	11
chemistry		7 General Science Journals	948	6,828	14
		<i>Physical Review</i>	2,113	14,670	14
cosmology		7 General Science Journals	26	167	16
		7 General Science Journals	8,670	37,441	23
biology			3,445	14,485	24
medicine			3,308	9,799	34
		<i>American Economic Review</i>	796	1,945	41
		39 Economics Journals	9,598	18,200	53

**Notes:** All searches except the *Physical Review* conducted on 3 June 2005 using the JSTOR Archive (<http://www.jstor.org/>). A list of the journals in each category is found at the Browse link on the JSTOR website. Searches of the *Physical Review* conducted on 20 July 2005 on the *Physical Review Online Archive* (<http://prola.aps.org/>). Searches covered parts A-E, *Special Topics: Accelerator Beams*, and *Physical Review Letters*. A hit is indicated by the presence of the at least one of the statistical keywords as modified according to the terms of the first two columns. The keywords are: "statistically significant," "statistical significance," "significance test(s)," "test(s) of significance," "t test," "F test," and "chi squared." All searches are for the years 1980-1999.

### Table 3

## Examples of Articles from the Physical Sciences that Use Statistical Significance from the 1980s and 1990s

- J.F. Ogilvie. "A General Potential Energy Function for Diatomic Molecules," *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 378, no. 1773, 8 October 1981, pp. 287-300.
- Kathryn A. Whaler. "Geomagnetic Secular Variation and Fluid Motion at the Core Surface," *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 306, no. 1492, *The Earth's Core: Its Structure, Evolution and Magnetic Field*, 20 August 1982, pp. 235-246.
- Y.T. Chen, Alan H. Cook, and A.J.F. Metherell, "An Experimental Test of the Inverse Square Law of Gravitation at Range of 0.1 m," *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 394, no. 1806, 9 July 1984, pp. 47-68.
- Jay D. Goguen and William M. Sinton. "Characterization of Io's Volcanic Activity by Infrared Polarimetry," *Science*, new series, vol. 230, no. 4721, 4 October 1985, pp. 65-69.
- S. Singh *et al.* "Charge Stripping and Delayed Autoionization in Doubly Charged Ions of the Noble Gases," *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 402, no. 1823, 9 December 1985, pp. 373-400.
- J.D. McDowell. "New Physics from the CERN Collider," *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 404, no. 1827, *Quarks and Leptons: The New Elementary Particles?*, 8 April 1986, pp. 213-232.
- Thomas K. Gaisser. "Gamma Rays and Neutrinos as Clues to the Origin of High Energy Cosmic Rays," *Science*, new series, vol. 247, no. 4946, 2 March 1990, pp. 1049-1076.
- J.R. Southon, J.W. Stark, J.S. Vogel, and J.C. Waddington. "Upper Limit for Neutron Emission from Cold *d-t* Fusion," *Physical Review C*, vol. 41, November 1990, pp. R1899-R-1900.
- C.L. Bennett, *et al.* "Scientific Results from the Cosmic Background Explorer (COBE)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 11, 1 June 1993, pp. 4766-4773.
- M.J. Molina *et al.* "Physical Chemistry of the H<sub>2</sub>SO<sub>4</sub>/HNO<sub>3</sub>/H<sub>2</sub>O System: Implications for Polar Stratospheric Clouds," *Science*, new series, vol. 261, no. 5127, 10 September 1993, pp. 1418-1423.
- C.J. Gilmore, K. Shankland, G. Bricogne. "Applications of the Maximum Entropy Method to Powder Diffraction and Electron Crystallography," *Proceedings: Mathematical and Physical Sciences*, vol. 442, no. 1914, 8 July 1993, pp. 97-111.
- Wendy L. Freedman. "Measuring Cosmological Parameters," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 1, 6 January 1998, pp. 2-7.
- Eric Gawiser and Joseph Silk. "Extracting Primordial Density Fluctuations," *Science*, new series, vol. 280, no. 5368, 28 May 1998, pp. 1405-1411.
- M.P. Bradley *et al.* "Penning Trap Measurements of the Masses of Cs, Rb, and Na with Uncertainties  $\leq 0.2$  ppb," *Physical Review C*, vol. 83, 4510-4513.