forthcoming, *Journal of Economic Methodology*, March 2008

Signifying Nothing:

Reply to Hoover and Siegler

by Deirdre N. McCloskey and Stephen T. Ziliak

University of Illinois at Chicago and Roosevelt University, April 2007

deirdre2@uic.edu, sziliak@roosevelt.edu

We invite distribution of the paper in this form,
and especially any correspondence, and most especially any quarrels with it.

Abstract

After William Gosset (1876-1937), the "Student" of Student's *t*, the best statisticians have distinguished *economic* (or agronomic or psychological or medical) significance from merely *statistical* "significance" at conventional levels. A singular exception among the best was Ronald A. Fisher, who argued in the 1920s that statistical significance at the .05 level is a necessary and sufficient condition for establishing a scientific result. After Fisher many economists and some others—but rarely physicists, chemists, and geologists, who seldom use Fisher-significance—have mixed up the two kinds of significance. We have been writing on the matter for some decades, with other critics in medicine, sociology, psychology, and the like. Hoover and Siegler, despite a disdainful rhetoric, agree with the logic of our case. Fisherian "significance," they agree, is neither necessary nor sufficient for scientific significance. But they claim that economists already know this and that Fisherian tests can still be used for specification searches. Neither claim seems to be true. Our massive evidence that economists get it wrong appears to hold up. And if rhetorical standards are needed to decide the importance of a coefficient in the scientific conversation, so are they needed when searching for an equation to fit. Fisherian "significance" signifies nearly nothing, and empirical economics as actually practiced is in crisis.

JEL codes: C10, C12, B41

We thank Professors Hoover and Siegler (2008) for their scientific seriousness, responding as none before have to our collective 40 person-years of ruminations on significance testing in economics and in certain other misled sciences.[1] We are glad that *someone* who actually believes in Fisherian significance has finally come forward to try to defend the status quo of loss-functionless null-hypothesis significance testing in economics. The many hundreds of comments on the matter we have received since 1983 have on the contrary all agreed with us, in essence or in detail, reluctantly or enthusiastically.

Yet Fisherian significance has not slowed in economics, or anywhere else. Before Hoover and Siegler we were beginning to think that *all* our thousands upon thousands of significance-testing econometric colleagues, who presumably do *not* agree with us, were scientific mice, unwilling to venture a defense. Or that they were merely self-satisfied—after all, they control the journals and the appointments. One eminent econometrican told us with a smirk that he agreed with us, of course, and never used mechanical *t*-testing in his own work (on this he spoke the truth). But he remained unwilling to teach the McCloskey-Ziliak

---

[1] And our thanks to Philippe Burger of the Department of Economics of the University of the Free State, Bloemfontein, South Africa, for a very helpful discussion of these issues at a crux. The paper was drafted while McCloskey was Professor Extraordinary (i.e. briefly visiting) at the University of the Free State in March, 2007.

point to his students in a leading graduate program because “they are too stupid to understand it.” Another and more amiable but also eminent applied econometrican at a leading graduate program, who long edited a major journal, told us that he “tended to agree” with the point. “But,” he continued, “young people need careers,” and so the misapplication of Fisher should go on and on and on.

We do not entirely understand, though, the hot tone of the Hoover and Siegler paper, labeling our writings “tracts” and “hodge-podges” and “jejune” and “wooden” and “sleight of hand” and so forth. Their title, and therefore ours in reply, comes from Macbeth’s exclamation when told that the queen was dead: Life “is a tale / Told by an idiot, full of sound and fury, / Signifying nothing.” Hoover and Siegler clearly regard us as idiots, full of sound and fury. They therefore haven’t listened self-critically to our argument. Their tone says: why listen to idiots? Further, they do not appear to have had moments of doubt, entertaining the null hypothesis that they might be mistaken. Such moments lead one, sometimes, to change ones mind—or at any rate they do if ones priors are non-zero. Our reply is that significance testing, not our criticism of it, signifies nothing. As Lear said in another play, "nothing will come of nothing."

Nor do we understand the obsessive and indignant focus throughout on “McCloskey” (“*né* Donald,” modifying her present name by a French participle with a deliberately chosen male gender). For the past fifteen years the case that economists do in fact commit the Fisherian error, and that *t* statistics signify

nearly nothing, has been built by McCloskey always together with Ziliak, now in fuller form as *The Cult of Statistical Significance: How the Standard Error is Costing Jobs, Justice, and Lives* (2008). The book contains inquiries mainly by Ziliak into the criticism of *t* tests in psychology and medicine and statistical theory itself, in addition to extensive new historical research by Ziliak into "Student" (William Sealy Gosset), his friend and enemy Sir Ronald Fisher, the American Fisher-enthusiast Harold Hotelling, and the sad history, after Fisher and Hotelling developed an anti-economic version of it, of Student's *t*.[2] More than half of the time that McCloskey has been writing on the matter it has been "Ziliak and McCloskey."

Whatever the source of the McCloskey-itis in Hoover and Siegler, however, it does simplify the task they have set themselves. Instead of having to respond to the case against Fisherian significance made repeatedly over the past century by numerous statisticians and users of statistics— ignorable idiots full of sound and fury such as "Student" himself, followed by Egon Pearson, Jerzy

[2] S. T. Ziliak and D. N. McCloskey, *The Cult of Statistical Significance: How the Standard Error is Costing Jobs, Justice, and Lives* [Ann Arbor: University of Michigan Press), 480 pp., 2008; and S. T. Ziliak, "Guinness is Good for You (and So is Gosset): The Economic Origins of 'Student's' *t*," Department of Economics, Roosevelt University, 100 pp., April 20, 2007, http://faculty.roosevelt.edu/Ziliak.

Neyman, Harold Jeffreys, Abraham Wald, W. Edwards Deming, Jimmie Savage, Bruno de Finetti, Kenneth Arrow, Allen Wallis, Milton Friedman, David Blackwell, William Kruskal [whom Hoover and Siegler quote but misunderstand], David A. Freedman, Kenneth Rothman, and Arnold Zellner, to name a few—they can limit their response to this apparently *just awful, irritating* woman. An economic historian. Not even at Harvard. And, in case you hadn't heard, a former man.

But after all we agree that something serious is at stake. The stakes could generate a lot of understandable heat. If McCloskey and Ziliak are right—that merely "statistical," Fisherian significance is scientifically meaningless in almost all the cases in which it is presently used, and that economists don't recognize this truth of logic, or act on it—then econometrics is in deep trouble.

Most economists appear to believe that a test at an arbitrary level of Fisherian significance, appropriately generalized to time series or rectangular distributions or whatever, just *is* empirical economics. The belief frees them from having to bother too much with simulation and accounting and experiment and history and surveys and common observation and all those other methods of confronting the facts. As we have noted in our articles, for example, it frees them from having to provide the units in which their regressed variables are measured. Economists and other misusers of "significance" appear to want to be

free from making an “evaluation in any currency” (Fisher 1955, p. 75). *Economic* evaluation in particular, as we show in our book, was detested by Fisher.[3]

And so—if those idiots Ziliak and McCloskey are right—identifying "empirical economics" with econometrics means that economics as a factual science in deep trouble. If Ziliak and McCloskey are right the division of labor between theorem-proving theory and Fisherian-significance-testing econometrics that Koopmans laid down in 1957 as The Method of Modern Economics, and which Hoover and Siegler so courageously defend, was a mistake. What you were taught in your econometrics courses was a mistake. We economists will need to redo almost all the empirical and theoretical econometrics since Hotelling and Lawrence Klein and Trygve Haavelmo first spoke out loud and bold.

Of course—we note by the way—our assertion that Fisherian significance is simply beside the scientific point is not the only thing wrong with Fisherian procedures. We have tallied more than twenty-two non-Fisherian kinds of non-

[3] In most statistical results in economics “what you really want to know," Gosset said in 1937 to Egon Pearson, ”is can you [or someone else] make money by it?” Such economism drove Fisher mad. See, for example, Fisher 1925a, 1935, 1955, 1956; Hotelling 1927-1939, 1951, 1958; Neyman 1956, 1957, 1961; Pearson 1939, 1990; Kruskal 1980; McCloskey 1998, chp. 8; Ziliak 2007; and Ziliak and McCloskey 2008, chps. 20-23.

sampling error—each kind, from Gosset's "*a priori* bias from fertility slopes" in agriculture to Deming's "bias of the auspices" in survey questionnaires, causing in most applications far more trouble than Type I error does at, say, the .11 or even .20 level.[4] Hoover and Siegler mention this old and large criticism of Fisherian procedures only once, at the end of their paper, though there they mix it up. The analysis of "real" error was by contrast the heart of the scientific work of Morgenstern and Deming and Gosset himself.

* * * *

But anyway, are Ziliak and McCloskey right in their elementary claim that Fisherian significance has little or nothing to do with economic significance?

It appears so, *and Hoover and Siegler agree.* Their paper is not a defense of Fisherian procedures at all, as they forthrightly admit at the outset: "we accept the main point without qualification: a parameter . . . may be statistically significant and, yet, economically unimportant or it may be economically important and statistically insignificant." Let's get this straight, then: we all agree on *the main point* that Ziliak and McCloskey have been making now since the mid-1980s. We all agree that it is simply a mistake to think that statistical significance in R. A. Fisher's sense is either necessary or sufficient for scientific importance. *This is our central point, noted over and over again in a few of the best*

[4] Ziliak and McCloskey, *Cult of Significance,* Introduction.

statistical textbooks, and noted over and over again by the best theoretical statisticians since the 1880s, but ignored over and over again right down to the present in econometric teaching and practice.

Hoover and Siegler, it appears, would therefore agree—since economic scientists are supposed to be in the business of proving and disproving economic importance—that Fisherian significance is not in logic a preliminary screen through we can mechanically put our data, after which we may perhaps go on to examine the Fisher-significant coefficients for their *economic* significance. Of course any economist knows that what actually happens is that the data are put through a Fisherian screen at the 5 percent level of fineness in order to (in most cases illogically) determine what the important, relevant, keepable variables are, and then afterwards, roughly three-quarters to four-fifths of the time even in the best, *AER* economics, and in nearly every textbook, all is silence.

But wait. Hoover and Siegler call our logical truth "jejune"—that is, "dull." Fisherian significance is without question, they admit, a logical fallacy. Its fallacious character is not taught in most econometrics courses (one wonders whether it is in Hoover's and in Siegler's, for example), is seldom acknowledged in econometric papers, and is mentioned once if at all in 450-page econometrics textbooks. Acknowledging the mistake would change the practice of statistics in twenty different fields. And every one of the hundred or so audiences of economists and calculators to whom we have noted it since 1983 have treated it

as an enormous, disturbing, confusing, anger-provoking, career-changing surprise. "Dull"?

After their preparatory sneer they *take back their agreement*: "Our point is the simple one that, while the economic significance of the coefficient does not depend on the statistical significance [there: right again], *our certainty about the accuracy of the measurement surely does*."

No it doesn't. Hoover and Siegler say that they understand our point. But the sneering and the taking-back suggests they don't actually. They don't actually understand, here and throughout the paper, that after any calculation the crucial scientific decision, *which cannot be handed over to a table of Student's* t, is to answer the question of how large is large. The scientists must assess the oomph of a coefficient—*or assess the oomph of a level of certainty about the coefficient's accuracy*. You have to ask what you lose in jobs or justice or freedom or profit or persuasion by lowering the limits of significance from .11 to .05, or raising them from .01 to .20. Estimates and their limits in turn require a scale along which to decide whether a deviation as large as one standard deviation, or a difference in *p* of .05 as against .11 or .20, does in fact matter for something that matters. Not its probability alone, but its probable cost.

You do not evade the logical criticism that fit is simply not the same thing as importance by using statements about probability rather than statements about dollar amounts of national income or millions of square feet of housing. The point is similar to that in measuring utility within a single person by looking

at her choices in the face of this or that wager. Turning Ms. Jones' utility into a probability ranging from zero to one does indeed give economists a coherent way of claiming to "measure" Jones's utility. But of course it does not, unhappily, make it any more sensible to compare Jones' utility with Mr. Smith's. That requires an ethical judgment. Likewise the determination of "accuracy" requires a scientific judgment, not a *t* test equal or greater than the .05 level.

But ever since Fisher's *Statistical Methods* the economists—including now it would seem Hoover and Siegler—choose instead to "*ignore entirely all results* [between Jones and Smith or "accuracy" and "inaccuracy"] *which fail to reach this* [arbitrary, non-economic] *level*" (Fisher 1926, p. 504). Late in the paper Hoover and Siegler claim that "the" significance test "tells us where we find ourselves along the continuum from the impossibility of measurement. . . to. . . perfect accuracy." No: the twenty-two or more kinds of measurement error cannot be reduced to Type I sampling error. And—our only point—on the continuum of Type I error alone, short of literally 0 and literally 1.00000 (on which Hoover and Siegler lavish theoretical attention), there is still a scientific judgment necessary as to where on the continuum one wishes to be. The decision needs to be made in the light of the scientific question we are asking, not delivered bound and gagged to a table of "significance."

Think about that little word "accuracy," accorded such emphasis in Hoover and Siegler's rhetoric, as in "our certainty about the accuracy of the measurement." If an economist is making, say, a calculation of purchasing

power parity between South Africa and the United States over the past century she would not be much troubled by a failure of fit of, say, plus or minus 8 percent. If her purpose were merely to show that prices corrected for exchange rates do move roughly together, and that therefore a country-by-country macroeconomics of inflation would be misleading for many purposes, such a crude level of accuracy does the job. Maybe plus or minus 20 percent would do it. But someone arbitraging between the dollar and the rand over the next month would not be so tranquil if his prediction were off by as little as 1 percent, maybe by as little as 1/10 of 1 percent, especially if he were leveraged and unhedged and had staked his entire net wealth on the matter.

Now it's true that we can make statements about the probability of a deviation of so many standard units from the mean. That's nice. In other words, we can pretend to shift substantive statements over into a probability space. Hoover and Siegler say this repeatedly, and think they are refuting our argument. (It's a measure, we suspect, of their evident conviction that we are idiots that they say it so often and with such apparent satisfaction, as if finally *that* issue is settled.) They declare that Fisherian calculations can provide us with "a measure of the precision of his estimates," or can tell us when a sample "is too small to get a precise estimate," or provide us with "a tool for the assessment of signal strength," or is "of great utility" in allowing us to take whole universes as samples for purposes of measuring "the precision of estimates," or can give us a yes/no answer to whether "the components are too noisily measured to draw

firm conclusions," or whether "its signal rises measurably above the noise," or "whether data from possibly different regimes could have been generated by the same model."

No it doesn't. Unless there is a relevant scientific or policy standard for precision or signal strength or firmness or measurability or difference, the scientific job has been left undone. The probability measure spans a so far *arbitrary* space, and does not on its own tell us, without human judgment, what is large or small. The 5 percent level of significance—buried in the heart of darkness of every canned program in econometrics—is not a relevant scientific standard, because it is unconsidered. A *p* of .10 or .40 or for that matter .90 may be in the event the scientifically persuasive or the policy-relevant level to choose. And in any case the precision in a sample may not be the scientific issue at stake. Usually it is not. Occasionally it is, and in this case a considered level of *p* together with a consideration of power would be worth calculating. It is never the issue when one wants to know how large an effect is, its oomph.

We realize that since 1927 a growing number of economists—upwards of 95 percent of them by our survey during the 1980s and 1990s—have fervently believed that the so-called test settles "whether" an effect "is there" or not—after which, you see, one can go on to examine the *economic* significance of the magnitudes. But we—and the numerous other students of statistics who have made the same point—are here to tell the economists that their belief is mistaken. The sheer probability statement about one or two standard errors is useless,

unless you have judged *by what scale a number is large or small for the scientific or policy or personal purpose you have in mind.* This applies to the so-called "precision" or "accuracy" of the estimate, too, beloved of Hoover and Siegler—the number we calculate as though that very convenient sampling theory did in fact apply.

Scores of medical statisticians, psychometricians, and theoretical statisticians have complained that their people, like ours, do *not* think it's worth the trouble after R. A. Fisher to defend the mixing up of Fisherian and substantive significance. Their people, and our economists, just go ahead cheerfully mixing them up, killing cancer patients and misdiagnosing schizophrenia and failing to recognize the salience of world prices for policy on inflation at home. The cheerful mixers have no justification in statistical theory—remember, Hoover and Siegler say they agree. Still they just do it.

* * * *

And so here's the first big point in the Hoover-Siegler defense of current practice. They claim to think that economists do *not* mix up substantive and Fisherian significance. We contend that economists do, in almost all the textbooks (we have examined scores of them, contrary to what Hoover and Siegler imply) and in the great bulk of the papers in leading journals (which we and others such as Zellner have also examined). So the first big issue between us is a matter of fact.

Hoover and Siegler are denying a fact about significance testing as used that any economist with his eyes open would not venture to deny. The economist will on occasion explain away the admitted fact, in various ways, not all of them noble or just. More than one eminent economist has replied to us in private, "Yes, it *is* silly to mix up substantive and Fisherian significance—silly and common. But we barons and baronesses at Cornell and Princeton don't do it. Only third-raters at state universities do." (This by the way turns out *not* to be factually true—thank the Lord. Our democratic principles were offended by such remarks, which is one reason we did the surveys in the first place. So we were relieved to discover that at Cornell and Princeton they mix things up, too, and in certain ways worse than do the peasants at Roosevelt or the University of Illinois at Chicago.)

Hoover and Siegler have set themselves the task of denying the obvious. Their rhetoric, therefore, betrays a certain sweaty desperation.

For example they take the failure of significance-using economists to defend the mixing up of substantive significance and Fisherian significance as evidence that the economists *already understand* the "uncontroversial" point that the two should not be mixed up. So by analogy, for example, the statistical economists of the 1920s who failed to defend the mixing up of the joint effects of unidentified demand and supply curves may be taken as evidence that the pre-Holbrook-Working, pre-Cowles economists already understood the "uncontroversial" point that the two curves need identifying restrictions. And

likewise before Arnold Zellner the average economist knew how to compute the power function for her posterior estimates of inflation and unemployment, since after all she failed to defend her non-use of power.

Or again, late in the paper Hoover and Siegler assert that "the power of the test is not typically ignored." Their evidence? Power "is a major consideration among specialist econometricians," an assertion not backed by evidence, but one we are willing to stipulate. Then they concede that "it is less frequently discussed by workaday users of econometric methods." That's putting it mildly: *four* percent of economists in our *AER* sample "discussed" it, and *one or two percent* did anything about it. Hoover and Siegler would have done well to ask our friends Zellner and Horowitz and Wurtz about how "major" the use of power is.

Maybe we have here the source of the hotly defended conviction that Fisherian significance is OK. Like the Cornell economist excusing it as the practice of mere peasants, Hoover and Siegler believe that at the commanding heights of the profession—those smart specialist econometricians who have nearly all failed to teach their students the "uncontroversial" point that fit is not the same thing as importance—things are fine, because Fisherian tests are used there with discernment. (That, by the way, is an alleged fact we are *not* willing to stipulate: we too have read some of the books and articles published on the commanding heights). And so everything's fine. Two percent of the people who

do econometrics do it right. So stop complaining about the 98 percent who do it wrong.

Or yet again, Hoover and Siegler attack our survey instrument applied to the *AER* in the 1980s and 1990s as a "hodge-podge" (more heat), addressing idiotic questions such as whether economists report the magnitudes of the fitted coefficients or whether they consider power or whether they think statistical significance all by itself can "decide" a result. Hoover and Siegler are wrong, of course to think our survey questions are off the main point: magnitudes and power are two ways among many of getting beyond the routine of Fisherian significance, as taken together are all the items in our Edgeworth-Gosset-Wald-Savage-Kruskal-Granger-Horowitz-and-Zellner calibrated instrument. Our point is that most economists haven't gotten a single step beyond the routine of Fisherian significance.

Hoover and Siegler further complain that if one looks into *many* connected mistakes, one has an index number problem in weighting them. Well, yes. Got it. Surprisingly, the same point occurred to the economists McCloskey and Ziliak. But so what? Does that mean that it's better to measure auto theft alone when looking into crime?

In a footnote they assert that our embarrassing omission of many of the papers in the survey of the *AER* in 1990s (corrected in *The Cult of Significance*; since the sample was large, of course, nothing much changed) is "emblematic" of our disgraceful carelessness in argument. We might reply with similar heat that

to call items such as power or magnitude "tangential" to the primary mistake and part of a "hodge-podge" of questions and a source of an (apparently always hopeless) index number problem, all of which anyway are "uncontroversial" and "jejune," is "emblematic" of the quality of *their* argument.

Also "emblematic" is their distaste for the subjective character of our textual measurement, a distaste articulated at the same time, however, with an admission that, of course, "subjectivity alone does not rule out scientific or reproducible procedures." That's right. One does not have to refer to psychology for cases in point. After all, the unemployment rate is the result of a survey, turning on the subjective matter for instance of whether an activity is judged "paid work" in the mind of the respondent. (Paid to babysit your brother? Paid in cash? "Work"?)

In 2004 we invited Hoover and Siegler to sit down with us to discuss the necessarily "subjective" scoring of particular papers, but they declined. They demanded that we *write down* the 7,000 or so decisions we had made, complete with page and sentence citations. For reasons of the opportunity cost of time we declined. We instead invited them to come to Chicago, where we both live, to examine photocopies of the original articles which we had made and written notes on, January 1980 to December 1999. We offered to sit down with them to discuss the data and the notes. Hoover waxed wroth.

But, we asked mildly, if you want to confute our results, why don't *you* rescore the 370-odd papers, or even a modest sample of them?[5] The *AER* articles are in the libraries: so go ahead. They didn't, and haven't, not for any sample size—not for $N = 10$ (one that McCloskey used in her very first attempt back in 1985 to persuade doubters like Hoover and Siegler that water flows downhill) or $N = 369$ (our "sample" size after correcting for some missing papers of the 1990s). It seems to be another case of not applying a standard of argument to ones own procedures that, in the style of the blessed Fisher himself, one so stridently demands others follow.

An electronic version of Hoover and Siegler's paper has been circulating for some years. In a widely read comment on it the RAND economist Kevin Brancato, who seemed at first glad to see a defense of the conventional wisdom, remarked, "I must say that I'm disappointed in H&S. I don't think H&S have much new to say other than the problem is not as bad as M&Z claim. However,

---

[5] As Kevin Brancato put it, "I was with H&S much of the way in that [empirical] section. . . until they equate the refusal of M&Z to *reproduce* a representative sample of the now lost paper-to-dataset mappings with a refusal to 'share' them. . . . [H&S] lose my respect with what I fear is not just poor word choice." Our frequent e-mail exchanges with Hoover about this matter confirms Brancato's hypothesis. H&S intended then to impugn our scientific integrity, and intend so now.

this is an empirical question, that in my mind, H&S fail to address thoroughly—in fact, not even in a cursory fashion."[6] One of Kevin Hoover's former colleagues, Thomas Mayer, who has for a long time been making the same point about econometric practice as we make here, commented, "I don't understand the relevance of the table with the additional papers. Isn't the main issue whether authors pay attention to the size of the coefficients? And that is not in the table. What am I missing?" Brancato and Mayer are not missing anything: the main issue is, as Mayer suggests, not whether the Ziliak and McCloskey survey as originally published contained "the entire population" but whether, on the evidence of any reading of statistical practice in the *American Economic Review* (100% of the papers, or an apparently random sample of less than 100%), the authors pay attention to the size of the coefficients when making decisions about what's important. Most don't.

But give credit where credit is due. Hoover and Siegler roused themselves at least to defend the papers we discussed by actual quotation or reference. Well, at any rate they defended five of the dozens of papers we discussed. They defend for example Michael Darby's low-scoring paper of 1984. (Let us state warmly, by the way, that we admire many of the economists who

[8] Kevin Brancato's comments, with Mayer's, may be found by googling such combinations as "Kevin" and "Truck and Barter" and "Hoover and Siegler."

scored low—Ben Bernanke, for example, and Gary Becker, and Michael Darby. Michael was a colleague of one of us years ago, and we agree in substance with much of his work. The same can be said of Gary. But Michael and Ben and Gary do not understand that Fisherian significance without a loss function is ordinarily useless for science.)

Hoover and Siegler defend Darby's standard error, and the other four papers they discuss, by *assuming the very hypothesis under dispute*. That's known in logic as "begging the question," as on one matter they later, and with their usual heat, claim that we do. Darby performs *F*-tests, which in Hoover and Siegler's words constitute a "specification exercise [which] suggests his final specification."

So Hoover and Siegler *do* believe that Fisher significance is an initial filter through which one can sift ones data, for the "final specification." We are to add and drop variables, *that is, to determine the substantive importance of the variables*, on the basis of *t*- and *F*-tests. Fit *is* to be taken, they now claim—in ignorance of the *economic* approach to the logic of uncertainty taken by "Student" himself—as the *same thing* as importance. Darby, in Hoover and Siegler's words, "uses statistical significance as a measure of the precision of his estimates." Darby found, in his own words, that "no direct effect [of oil-price increases on productivity] is directly detectable." The standard of "detectable" oomph here? Statistical significance at the 5 percent level. A variable, Darby and Hoover and Siegler are claiming, is either "detectable" or not, *by an absolute standard determined by*

consulting a table of Student's t *or Fisher's* F *at an arbitrary level of significance.* So Hoover and Siegler do *not* agree that it's simply a mistake to think that statistical significance in R. A. Fisher's sense is necessary for importance.

They defend Stephen Woodbury and Robert Spiegelman (1987) in the same way, begging the question of whether running ones intellectual life with *t*-tests is a good idea. Woodbury and Siegelman, after doing an experiment that reaches in other ways very high standards of scientific persuasiveness, proposed not to advise the State of Illinois that a dollar spent on an employment subsidy for Black women would on average save the state over four dollars in unemployment benefits because, in Hoover and Siegler's paraphrase (our italics in this and the next quotation), "the number of experimental subjects in three of the categories is too small to get a *precise* estimate." Again in Hoover and Siegler's words, "some of components are too noisily measured to draw *firm* conclusions." They are assuming what was to be proven, that (sampling) precision or (sampling) firmness at an arbitrary level of significance is the same thing as importance. Don't tell the state the best guess about the Black women—regardless of the loss function in human lives diminished. If the regression doesn't fit / Then you must acquit. Set Fisher-"insignificant" variables at zero. "Ignore all results that fail to attain this level," as Fisher literally said. In the final specification, throw 'em out. Fit is what we seek, not oomph.

They defend Benjamin Bernanke and Alan Blinder (1992) against our observation that these two eminent and excellent economists used tests of

significance without reporting the magnitudes of coefficients or asking whether the variables were substantively important. Their "defense" is that the Granger-causality test "does not imply [that a variable] is important…, but only that its signal rises *measurably* above the noise *according to a conventional threshold.*" That's right, alas: "measurably" by Fisherian convention above the "conventional [Fisherian] threshold." It's why someone who actually grasped that Fisherian significance is neither necessary nor sufficient for the scientific importance of a variable, whether in slope or in sampling variance, as Hoover and Siegler so forthrightly claimed they did grasp some pages earlier, would want the contents of the canned program yielding mechanical, 5-percent judgments on Granger-causality to be opened up. (By the way, Clive Granger is one of the many econometricians who pretty much agree with McCloskey and Ziliak, in print and especially in private.)

The criterion for "measurable" cannot be handed over to "a conventional threshold." Scientific judgment is not like that. Numbers matter, of course, as inputs into a scientific judgment. But they have to be assessed every time on a scale of relevant importance. That's what most of the physicists and historians do. A table of Student's *t* simply doesn't tell us how large is large. Fit to an arbitrary standard is not the same thing as importance, and is commonly irrelevant to it. As *The Onion* once put it in one of its crazy headlines, "Standard Deviation Not Enough for Perverted Statistician." Nor should it be for you normal statisticians. All kidding aside: all the econometric findings since the

1930s need to be done over again. Almost all have depended on mixing substantive and Fisherian significance.

And so it goes. The Hoover and Siegler argument is supposed to show, contrary to what every economist knows and what our survey shows and what the textbooks demonstrate, that "there is little evidence that economists systematically" mix up Fisherian and scientific significance. Hoover and Siegler *assume* that "a [Fisherian, unadorned] test of whether data come from possible different regimes. . . is appropriate." They *assume* that if the economists show any interest anywhere in the actual magnitudes (of, say, cigarette addiction) they are absolved of dropping and adding variables on the basis of a Fisherian "test."

We suspect, actually, that this last why such energetic and intelligent economists as Hoover and Siegler have so completely missed the point. They think we are saying, "You know, economists don't *ever* care about magnitudes, *anywhere* in their papers. When Becker, Grossman, and Murphy look into the facts of cigarette addiction, they *never* actually state the magnitudes." That would be a silly thing to say, contradicted by the merest glance at what Becker et al. wrote. You can see how high is the prior in Hoover and Siegler that we are idiots. Only idiots ($P = .98$) would say such a thing.

What we do say is that the typical economist doesn't care about magnitudes *when "formally," "statistically" testing and deciding upon the importance of a variable.* For that job the economist strongly tends to substitute fit for oomph, that is, R. A. Fisher for common sense. "Student," in his short life, always said so.

Then maybe somewhere later in the paper the economist will get around to talking about the oomph (a step we commend in our questionnaire, and on which Becker et al. in fact scored well). Becker et al., as Hoover and Siegler triumphantly report with numerous quotations, do get around to talking about magnitudes. Yeah, we realized they did. We read the paper, too, and 368 other ones, 1980-1999. But meanwhile the way Becker et al. have decided which magnitudes to focus on is Fisherian. Pure, unadorned R. A. Fisher, c. 1925. And it's wrong, which Hoover and Siegler said they understood. It's the wrong way to decide, and leaves the wrong variables in the regressions, and results in biased and inconsistent estimates of the coefficients. That's elementary.

In the course of several pages copying out the usual logic of significance tests—we suppose Hoover and Siegler mean to signal by this that they are sophisticated theorists of statistics—they claim that when we recommend "Edgeworth's standard" we mean his conventional level of significance. That's actually what we *don't* like about Edgeworth, 1885, in contrast to, say, Edgeworth, 1907. We meant, and said, that even Edgeworth, the very inventor of the disastrously equivocal term "significance," does distinguish between practical and some other significance. When he recommends that a "scientist" might judge a 3 percent difference worth looking into if it "repays the trouble" he is making our point: that the decision to attend or not to attend to a difference of this or that magnitude is itself a human and scientific and often indeed economic decision that cannot be handed over to machinery. A speculator on the foreign

exchanges might want one level, a student of the habits of bees and wasps (one of Edgeworth's many hobbies) might want another. This is precisely what R. A. Fisher and the mechanical tradition down to Hoover and Siegler deny.

To adopt in tone a Hoover-Sieglerism, we find their historical research shallow. They have not actually read, it would appear, more than a very few books and articles on the history of the procedure they are confidently defending. They do not grasp what we have said in all our work since 1983, and now especially in *The Cult of Significance*, that null hypothesis testing was adopted in many fields because of the rhetorical and political skills of R. A. Fisher, and was contradicted by most of the people around him in Gower Street. Gosset tried for years to persuade Fisher to acknowledge "real" error and to accept that a statistical decision is a human decision. In 1926 he explained in a letter to Egon Pearson (1895-1980) that benefit and cost should guide use of his *t*, and incidentally there in the 1926 letter gave Pearson and eventually Neyman the idea of power which they went on to formalize.[7] Fisher by contrast looked all his life for a qualitative essence, rather in the style of economics up to the 1870s looking for an essence of value, before opportunity cost was made clear. After the Gosset letter Pearson joined with Fisher's other domestic enemy, Jerzy

[7] Letter from Gosset to E. S. Pearson, May 11, 1926, "Student's Original Letters," Letter #2, Green Box, G7, Pearson papers, Egon file, University College London, Special Collections.

Neyman, in their great series of decision-theoretic papers of the late 1920s and early 1930s. They showed beyond cavil that Fisher was wrong.

* * * *

On the page or so following the block quotation of the example of "kelp, kelp" the Hoover-Siegler attack on our point reaches a sort of crescendo (by the way, read the page and you'll see what we mean by the "crescendo" in the *AER* papers we examined). By then they have thankfully given up the task of denying the obvious—the large overuse of Fisherian significance in economics—and are going after our theoretical objections to substituting fit for importance. Here they get really angry. Noise, they say with even more than their usual heat, can mask a cry for help. The cry of "help" (instead of "kelp"), they say, "may be there or it may not be there. The point is we do not know."

No, the point is that we "know" at a level of significance. The choice of the level depends every time on the cost and benefit, in lives saved or profit gained or scientific persuasions performed. "Clearly," they admit, "if the costs and benefits are sufficiently skewed, we may seek more data." No, not if we must act now on the sample we have. The woman crying "help, help" needs our assistance now, not after we have applied to the National Science Foundation for a grant to buy a good hearing aid ($p < .05$). True, when we rush to her assistance we may find to our embarrassment that she was only crying "kelp, kelp," in a strangely heated argument about the best vegetable to use for true Dutch

stamppot. Against our potential embarrassment is set. . . her life. A level of 5 percent won't do.

Hoover and Siegler are uneasily aware of this. "If the potential cost of Type II error is large," as it is here, "we may choose a large test size." That's right. But it is exactly what the embedded use of Fisherian 5 percent in the *AER*, and even in the most sophisticated econometrics from David Hendry on down, does *not* do. Hoover and Siegler declare that with small samples "the noise overwhelms the signal." *But there is no absolute standard of "overwhelming*." It depends. Nor is there an absolute standard of "smallness" in samples. It depends. Gosset co-invented with Edwin S. Beaven a variety of barley that helped make Guinness good for you. He did it acre-by-acre on an experimental field in Warminster, near Reading, of size *four* acres. That is, Gosset changed the cost and nutrition and taste of Guinness beer, working with samples of size four. No wonder he wanted a small-sample theory. Field experiments are expensive in labor and revenue foregone. Hoover and Siegler don't get what the inventor of Student's *t* himself got: that economic judgment, or else a judgment of what persuades other scientists, must be used in every step of a statistical calculation.

Hoover and Siegler discuss for many pages the justification for taking a population as an ersatz "sample." In our work we have mentioned the matter incidentally, as still another frailty of Fisherian significance, to which we have not devoted sustained attention (along with, for instance, the frequently noted publication bias in reporting "significance": that too was a side point in our

work—though not, by the way, responded to in Hoover and Siegler, and mentioned only once). Ersatz sampling was not our main point. Our main point, you will recall, Hoover and Siegler have long conceded.

But we're willing to discuss the matter. We realize that it would be very convenient if a time series of length *N* could be taken routinely, without consideration, as a sample of size *N*. It would be convenient because then, by a happy chance, we could apply all the neat things we know about the mathematics of samples from populations, expressions involving that very *N*. It's the usual routine. One suspects that the Haavelmo assumption is being adopted *because then sampling theory can be used,* not because on sober consideration the "sample" really is plausibly viewed as a sample. A balder case than time series would be a regression of electricity usage in 2008 against national income for all 192 members of the United Nations. The urn of nature is spilled out before us. The relationship is a matter of calculation, counting red and white balls from the spilled urn, not an inference to a universe from a "sample" of size 192.

Hoover and Siegler find it sincerely "puzzling" that two counterfactual-loving economic historians such as Ziliak and McCloskey do not realize that "we must situate observed data in the context of population data that could have been observed, but was not," and then cite Robert Fogel's great work of the mid-1960s on counterfactual U.S. transport in 1890 by water. Yes, so true. It is the justification (which again they find "puzzling" in our mouths) for a much greater

use of simulation in a future economics. But Fogel knew that his "sample" was sized $N = 1$. Hoover and Siegler do not. The relevant sample size is simply *assumed* to be $N = 192$. More begging of the question.

They appear to appreciate that a time series of crime rates in Philadelphia in the 1980s, say, is one instantiation out of an infinite number of "different paths had the errors been realized differently." There is no reason, except the purist metaphysics in the abusive sense of that fraught word, to suppose that the observation for June, 1986 is a properly random sample from all possible universes. In view of autocorrelation and of structure and path-dependence over time, it certainly is not. But anyway: sample size in Philly? One.

* * * *

The failure of Hoover and Siegler to grasp that a number needs a rhetorical standard comes through sharply in the last section of their paper, concerning their online-publication-enabled statistics on the prevalence of this or that scientific practice. They assert boldly that "the claim is false" that, as the idiots McCloskey and Ziliak (and Zellner) claim, economists do not much use confidence intervals.

It is of course not the case that confidence intervals solve the Fisherian problem. More machinery by itself cannot. Only economic judgment and persuasion, two sides of the rhetorical coin, can. Hoover and Siegler find "puzzling" that we hammer away at *t* tests but recommend confidence intervals.

After all, they note, confidence intervals can be derived from the *t* test—at any rate in the minority of cases in which the econometrician has provided enough information about the fitted coefficient to do so. But that's not the point. The point, as we suggested early and late, is that being forced to think about an interval of the variable in question at least encourages the economist to wonder how big is big. A lone asterisk on the fitted coefficient, which is the usual economic practice, does not.

Their evidence for asserting that it is simply "false" that economists underuse such wondering-provoking confidence intervals is a JSTOR search of 39 economics journals over two decades, producing 1788 entries "indicative of reporting or using confidence intervals." They leave it at that. Gosh: 1788 is big, isn't it?

No, it isn't, *not by a relevant standard.* Another scientist in the conversation would not be persuaded, unless she is simply uncritically dazzled by 1788 being "far" from, say, zero.

Here's one relevant standard: the 39 journals were published at a minimum 4 times a year and had perhaps 8 empirical articles in each issue over the 20 years. That's about 25,000 articles, of which 1788 is a mere 7 percent.

And who knows how important the alleged confidence interval was in each paper? Hoover and Siegler do not actually read the articles and apply our questionnaire to the "hits" that a computerized search of JSTOR records. Ask, then, as they do not, of each hit or article, what is the content—a single mention

for some inessential point? Ten mentions in one paper, none in fourteen others? Seven percent, to repeat, is not big by a standard that 100 percent *should* be reporting confidence intervals, and the situation is a lot worse if the intervals should accompany each of the (say) 20 estimated coefficients in each paper. That's *half a million* coefficients that should have confidence intervals reported, if economists in the 39 journals were actually thinking about magnitudes when they report Fisherian significance. So a lower bound on the substantive importance of their 1788 hits is that it is 4 tenths of 1 percent of the ideal of 100 percent.

Seven percent, not to speak of 4 tenths of 1 percent, is substantively far from 100 percent, right? We ask you instead of telling you because the *rhetorical* standard is what matters for science, what persuades a serious economic scientist engaged in the conversation. On this point *we* don't know once and for all. You and we together must consider it. There is no "absolute" standard, of a 5 percent probability of a Type I error, say. You, the serious economic scientist, must decide, in light of the numbers, but not mechanically ruled by the numbers. That's neither arbitrary nor jejune. It's the scientific conversation.

Similarly, Hoover and Siegler believe they falsify our assertion that physicists and chemists do not use statistical significance—much. We admit that our statement that the physicists, say, *never, ever* use statistical significance was an overstatement, and we will gladly send Hoover and Siegler each the check for $50 promised in some of our presentations to anyone who could find physicists

misusing it. But that a very few physicists make the same theoretical mistake that economists make, using an arbitrary level of *t* to "assess the quality of the observations relative to the assumed statistical model," does not mean that economists are right to go on ignoring substance in favor of Fisherian routine. The fact is—look at their useful Table 2—that economists in the 39 economics journals use "some statistical terminology" over 2 times more than cosmologists and 5 times more than non-cosmologist astronomers and 8 times more than non-astronomical physicists.

Hoover and Siegler admit indeed that the role of significance tests in the physical sciences is "a modest one." That, again, is putting it mildly. Their argument shows again how reluctant Hoover and Siegler are to attend to meaningful magnitudes, preferring instead to stick with the Fisherian routine of on/off tests of "whether" something "exists" or "is accurate." We have not done the empirical work, but wouldn't it be reasonable to suppose that the number of such tests per paper in, say, physics is much lower than in the typical economics paper littered with asterisks? Would it surprise you if the typical physicist used the test, say, 2 times in the 8 percent of papers that used it at all and the typical economist in each such paper used it 20 times? In which case, wouldn't you find important the resulting (8 multiplied by [20/2]) = 80 to 1 difference in the usage between economics and physics? Or would you want to base the decision on the standard error of the estimate, substituting fit for oomph?

We ask such rhetorical questions, again, because the issue *is* rhetorical. Hoover and Siegler ask with some heat, "How would. . . physicists define a loss function?" But like jesting Pilate they do not stay for an answer. The answer is not (as Hoover and Siegler indignantly assert we are saying) that every scientific question must have a vulgar application to a world of money. Though many do. The answer is that economic or physical scientists *face an audience of other such scientists*. That is what provides the standard for judging numbers large or small. *There is no non-human standard for the decision*. Deciding, judging, concluding are human activities, and not activities, we repeat, that can be turned over to a machine, however nice it is to have the machines in good working order. Some person in the conversation must propose a considered level of fit, constituting a substantively meaningful scientific improvement over some other fit, and must argue the case. She must tell how the size of a variable matters, and must argue. Fisherian tests in the way they are overwhelming used in economics, or in the exceptionally rare cases that they are so used in physics, do not do anything of the sort. Econometrics must be taken apart and redone from top to bottom, attending now to considered standards of oomph, whether in matters of coefficient size or in matters of fit.

We are not just randomly breaking up the machinery. Hyperplane fitting is lovely and interesting. We, too, are quantitative folk. Numbers are essential for real science. But once the matrices are inverted a human being must judge. Humans who are good at scientific persuasion, such as the Robert Fogel whom

Hoover and Siegler praise, engage in argument with their colleagues. They try to persuade them, as did Fogel, for example, with lower bound estimates, the argument *a fortiori*. They try to persuade them with multiple arguments, commonly called "triangulation," and called in classical rhetoric *copia*. They ask, as we just did, whether something that occurs $1/80^{th}$ of the time in one field as in another might be considered a "detectable difference" by a substantive standard.

One final point. The beliefs of economists don't actually depend on significance testing. The fact is evident from the very large number of tests done each year, on every side of an issue, without consensus. New facts are persuasive in economic science, as generated in cliometrics such as the national income accounting of Kuznets and Maddison. Historical instances are persuasive, offering new stories—the Great Depression, after all, inspired modern macroeconomics. Accounting is persuasive—witness Friedman, Little, Samuelson and the direct vs. indirect tax argument in the late 1940s. New theories, that is, new metaphors, are persuasive—thus Keynes on "animal spirits," Schumpeter on "creative destruction." Becker on children as "durable goods," and Nancy Folbre on the "invisible heart." Theorems are sometimes persuasive, though mainly in a negative way against other theorems, as in Arrow's Impossibility Theorem, or the Folk Theorem demolishing the claims of game theory.

But the sizeless stare of statistical significance—testing without a loss function and without full attention to the question "how big is big"—is not

persuasive. Null hypothesis significance testing is an empty and damaging ceremony. In Fisher's hands, "Student's" original Bayesian test of alternative hypotheses became a one-way test of the null. Well, sort of—in Fisher's hands it is not even the hypothesis that is being tested, but the data. Fisher transposed the conditional probability, creating in daily usage what is known as the fallacy of the transposed conditional. "If Hypothesis, then Data," is not the same as "if Data, then Hypothesis." The great scientist Harold Jeffreys and before him the great brewer Gosset himself tried to persuade Sir Ronald to take his hands off of the dangerously reversed machine. He didn't. The marriage of Fisher's sizeless stare of statistical significance to the fallacy of the transposed conditional we call *testimation,* which has been the ruin of empirical research in economics as in medicine and sociology and psychology. As the psychologist Jacob Cohen has shown, for example, Fisher's testimation has led to the over-diagnosis of adult onset schizophrenia (Cohen 1994). The null procedure does not in the end change rational minds.

In fact, we have argued, it shouldn't. Card and Krueger (1994) changed some minds about the minimum wage with their sample design and their brilliant exploitation of a natural experiment. They did not change minds with their erroneous and mechanical testimations, signifying nothing.

References

Altman, Morris. 2004. "Statistical Significance, Path Dependency, and the Culture of Journal Publication." *Journal of Socio-Economics* 33(5): 651-63.

Berg, Nathan. 2004. "No-Decision Classification: An Alternative to Testing for Statistical Significance." *Journal of Socio-Economics* 33(5): 631-50.

Berger, James O. 2003. "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science* 18(1): 1-32.

Card, David, and Alan B. Krueger. 1994. *Myth and Measurement: the New Economics of the Minimum Wage*. Princeton: Princeton University Press.

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

De Finetti, Bruno. 1971 [1976]. "Comments on Savage's "On Rereading R. A. Fisher." *Annals of Statistics* 4(3): 486-7.

Edgeworth, Francis Y. 1885. "Methods of Statistics." *Jubilee Volume of the Statistical Society*, pp. 181-217. Royal Statistical Society of Britain, June 22-24.

Edgeworth, Francis Y. 1896. "Statistics of Unprogressive Communities," *Journal of the Royal Statistical Society* 59 (2, June): 358-86.

Edgeworth, Francis Y. 1907. "Statistical Observations on Wasps and Bees." *Biometrika* 5(4, June): 365-86.

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

Fisher, R. A. 1921. “On the "probable error" of a coefficient of correlation deduced from a small sample.” *Metron* 1: 3-32.

Fisher, R. A. 1922. “On the Mathematical Foundations of Theoretical Statistics.” *Philos. Trans. Roy. Soc. London Ser. A* 222: 309-368.

Fisher, R. A. 1923. “Statistical tests of agreement between observation and hypothesis.” *Economica* 3: 139-147.

Fisher, R. A. 1925a[1941]. *Statistical Methods for Research Workers*. New York: G. E. Stechart and Co. Originally published Edinburgh: Oliver and Boyd. Eighth of thirteen editions, in at least seven languages.

Fisher, R. A. 1925b. “Applications of ‘Student’s’ distribution.” *Metron* V(3, Dec.): 90-104.

Fisher, R. A. 1925c. “Expansion of ‘Student’s’ Integral in Powers of n^{-1}.” *Metron* V(3, Dec.): 110-120.

Fisher, R. A. 1926. “Arrangement of Field Experiments.” *Journal of Ministry of Agriculture* XXXIII: 503-13.

Fisher, R. A. 1931. “Letter to Arne Fisher.” Page 313. in J. H. Bennett, ed., 1990.

Fisher, R. A. 1932. “Family Allowances in the Contemporary Economic Situation.” *Eugenics Review* 24: 87-95.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd. Reprinted in eight editions and at least four languages.

Fisher, R. A. 1936. “The Significance of Regression Coefficients” [*abstract*], *Colorado College Publ. Gen. Ser. 208*, pp. 63-67. Cowles Commission Research Conference on Economics and Statistics.

Fisher, R. A. 1950. *Contributions to Mathematical Statistics.* New York: Wiley. Ed. Walter A. Shewhart.

Fisher, R. A. 1951. “Letter to W. E. Hick.” P. 145, in J. H. Bennett, ed., *Statistical Inference.*

Fisher, R. A. 1955. “Statistical Methods and Scientific Induction.” *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 17, No. 1, pp. 69-78.

Fisher, R. A. 1956 [1959]. *Statistical Methods and Scientific Inference.* New York: Hafner. Second edition.

Fisher, R. A., and Frank Yates. 1938 [1963]. *Statistical Tables for Biological, Agricultural and Medical Research.* Edinburgh: Oliver and Boyd. Sixth edition.

Hoover, Kevin. 2006. “The Vanity of the Economist: A Comment on Peart and Levy’s *The Vanity of the Philosopher*.” Departments of Economics and Philosophy, Duke University, http://www.econ.duke.edu/~kdh9/

Hoover, Kevin, and Mark Siegler. 2008. “Sound and Fury: McCloskey and Significance Testing in Economics.” *Journal of Economic Methodology %%%%*

Horowitz, Joel L. 2004. “Comments on ‘Size Matters’.” *Journal of Socio-Economics* 33(5): 551-554.

Hotelling, Harold. 1927 to 1939. *Journal of American Statistical Association:*

1927 22, pp. 411-12—Review of Fisher's *SMRW*

1928 23, p. 346 " "

1930 25, pp. 381-2 " "

1933 28, pp. 374-5 " "

1935 30, pp. 771-2— Review of Fisher's *Design*

1937 32, pp. 218-9— *SMRW*

1937 32, pp. 580-2— *Design*

1939 34, pp. 423-4— *SMRW*

Hotelling, Harold. 1930. "British Statistics and Statisticians Today," *Journal of the American Statistical Association* 25 (170, June): 186-90.

Hotelling, Harold. 1931. "The Generalization of Student's Ratio." *Annals of Mathematical Statistics* 2: 360-78.

Hotelling, Harold. 1938. "Review of Fisher and Yates, *Statistical Tables for Biological, Agricultural and Medical Science*." *Science* 88: 596-7

Hotelling, Harold. 1951. "The Impact of R. A. Fisher on Statistics," *Journal of the American Statistical Association* 46 (243, Mar.): 35-46.

Hotelling, Harold. 1958. "The Statistical Method and the Philosophy of Science," *The American Statistician* 12 (5, Dec.): 9-14.

Jeffreys, Harold. 1963. Review of L. J. Savage, et al., *The Foundations of Statistical Inference* (Methuen, London and Wiley, New York, 1962), in *Technometrics* 5(3), August 1963, pp. 407-410.

Kruskal, William H. 1980. “The Significance of Fisher: A Review ofR. A. Fisher*: The Life of a Scientist*, *Journal of the American Statistical Association*, Vol. 75, No. 372 (Dec.):1019-30.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Non-Experimental Data.* New York: Wiley.

Leamer, Edward E. 2004. “Are the Roads Red? Comments on ‘Size Matters,’” *Journal of Socio-Economics* 33(5): 555-558.

Mayer, Thomas. 1979. “Economics as an Exact Science: Realistic Goal or Wishful Thinking?” University of California-Davis, Department of Economics, Working Paper Series, No. 124.

McCloskey, Deirdre N. 1983. “The Rhetoric of Economics,” *Journal of Economic Literature* XXI (June): 481-517.

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

McCloskey, Deirdre N. 1985b [1998]. *The Rhetoric of Economics* (Madison: University of Wisconsin Press).

McCloskey, Deirdre N. 1986. “Why Economic Historians Should Stop Relying on Statistical Tests of Significance, and Lead Economists and Historians into the Promised Land,” *Newsletter of the Cliometric Society*. 2 (2, Nov): 5-7.

McCloskey, Deirdre N. 1987. “Rhetoric Within the Citadel: Statistics,” pp. 485-490 in J.W. Wenzel at al., eds., *Argument and Critical Practice: Proceedings of*

the Fifth SCA/AFA Conference on Argumentation (Annandale, Va.: Speech Communication Association, 1987); reprinted in C. A. Willard and G. T. Goodnight, eds., *Public Argument and Scientific Understanding* (1993).

McCloskey, Deirdre N. 1992. "The Bankruptcy of Statistical Significance," *Eastern Economic Journal* 18 (Summer): 359-361.

McCloskey, Deirdre N. 1995. "The Insignificance of Statistical Significance," *Scientific American* (Apr): 32-33.

McCloskey, Deirdre N. 1997. "Aunt Deirdre's Letter to a Graduate Student" *Eastern Economic Journal* 23 (2, Spring): 241-244.

McCloskey, Deirdre N. 1997. *The Vices of Economists; The Virtues of the Bourgeoisie*. Amsterdam and Ann Arbor: University of Amsterdam Press and University of Michigan Press.

McCloskey, Deirdre N. 1998. "Two Vices: Proof and Significance," unpublished paper presented at an American Economic Association session in Chicago, Jan 3.

McCloskey, Deirdre N. 1999. "Cassandra's Open Letter to Her Economist Colleagues" *Eastern Economic Journal* 25 (3, Summer).

McCloskey, Deirdre N. 2000. "Beyond Merely Statistical Significance." Statement of editorial policy, *Feminist Economics*.

McCloskey, Deirdre N. 2000. *How to Be Human* *Though an Economist*. Ann Arbor: University of Michigan Press.

McCloskey, Deirdre N. 2002. *The Secret Sins of Economics.* Prickly Paradigm Pamphlets (Marshall Sahlins, ed.). Chicago: University of Chicago Press

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. "The Standard Error of Regressions." *Journal of Economic Literature* 34 (March 1996): pp. 97-114.

Neyman, Jerzy. 1956. "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society. Series B (Methodological)* 18 (2):288-94.

Neyman, Jerzy. 1957. "'Inductive Behavior' as a Basic Concept of Philosophy of Science." *Review of the Mathematical Statistics Institute* 25: 7-22.

Neyman, Jerzy. 1960. "Harold Hotelling—A Leader in Mathematical Statistics." Pp. 6-10 in I. Olkin, et al., eds., *Contributions.*

Neyman, Jerzy. 1961. "Silver Jubilee of my Dispute with Fisher," *Journal of Operations Research* (Society of Japan) 3: 145-54.

Pearson, Egon S. 1939. "'Student' as Statistician." *Biometrika* 30 (3/4, Jan.): 210-50.

Pearson, Egon S. 1990 [posthumous]. *'Student': A Statistical Biography of William Sealy Gosset.* Oxford: Clarendon Press. Edited and augmented by R. L. Plackett, with the assistance of G. A. Barnard.

Savage, Leonard J. 1954 [1972]. *The Foundations of Statistics.* New York: John Wiley & Sons [Dover edition].

Savage, Leonard J. 1971 [1976 posthumous]. "On Re-Reading R. A. Fisher." *Annals of Statistics* 4(3): 441-500.

Schelling, Thomas. 2004. “Correspondence” [Comment on Ziliak and McCloskey, “Size Matters”] *Econ Journal Watch* 1 (3, Dec.), pp. 539-40. http://www.econjournalwatch.com

Student [W. S. Gosset]. 1904. “The Application of the ‘Law of Error’ to the Work of the Brewery.” *Report*, Arthur Guinness Son and Co., 3 November, in E. S. Pearson 1939, pp. 212-15.

Student. 1905. “The Pearson Co-efficient of Correlation.” *Supplement to Report of 1904*, Arthur Guinness Son and Co., in E. S. Pearson 1939, p. 217.

Student. 1926 [1942]. “Mathematics and Agronomy,” *Journal of the American Society of Agronomy* 18; reprinted: Pp. 121-34 in E. S. Pearson and J. Wishart, eds., *Student's Collected Papers* (1942).

Student. 1938 [posthumous]. “Comparison between Balanced and Random Arrangements of Field Plots.” *Biometrika* 29 (3/4, Feb.): 363-78.

Student. 1942 [posthumous]. Student's *Collected Papers* (London: Biometrika Office). Eds. E. S. Pearson and John Wishart.

Thorbecke, Erik. 2004. “Economic and Statistical Significance: Comments on ‘Size Matters,’” *Journal of Socio-Economics* 33(5): 571-576.

Wooldridge, Jeffrey M. 2004. “Statistical Significance is Okay, Too: Comment on ‘Size Matters,’” *Journal of Socio-Economics* 33(5): 577-80.

Yates, Frank, and Kenneth Mather. 1963. “Ronald Aylmer Fisher,” *Biograph. Mem. Fell. R. Soc. Lond.* 9: 91–129.

Zellner, Arnold. 2004. "To Test or Not to Test and If So, How?: Comments on 'Size Matters,'" *Journal of Socio-Economics* 33(5): 581-86.

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): 527-46. And "works cited" there.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2004b. "Significance Redux." *Journal of Socio-Economics* 33(5): 665-75. And "works cited" there.

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