190

ELSEVIER

The Journal of Socio-Economics 33 (2004) 527–546

The Journal of Socio-Economics

www.elsevier.com/locate/econbase

# Size matters: the standard error of regressions in the American Economic Review

Stephen T. Ziliak[a,*], Deirdre N. McCloskey[b]

[a] *School of Policy Studies, College of Arts and Sciences, Roosevelt University, 430 S. Michigan Avenue, Chicago, IL 60605, USA*

[b] *College of Arts and Sciences (m/c 198), University of Illinois at Chicago, 601 South Morgan, Chicago, IL 60607-7104, USA*

## Abstract

Significance testing as used has no theoretical justification. Our article in the *Journal of Economic Literature* (1996) showed that of the 182 full-length papers published in the 1980s in the *American Economic and Review* 70% did not distinguish economic from statistical significance. Since 1996 many colleagues have told us that practice has improved. We interpret their response as an empirical claim, a judgment about a fact. Our colleagues, unhappily, are mistaken: significance testing is getting worse. We find here that in the next decade, the 1990s, of the 137 papers using a test of statistical significance in the *AER* fully 82% mistook a merely statistically significant finding for an economically significant finding. A super majority (81%) believed that looking at the sign of a coefficient sufficed for science, ignoring size. The mistake is causing economic damage: losses of jobs and justice, and indeed of human lives (especially in, to mention another field enchanted with statistical significance as against substantive significance, medical science). The confusion between fit and importance is causing false hypotheses to be accepted and true hypotheses to be rejected. We propose a publication standard for the future: "Tell me the oomph of your coefficient; and do not confuse it with merely statistical significance."

© 2004 Published by Elsevier Inc.

*JEL code:* C12; C10; B23; A20

*Keywords:* Standard error; Regression; American Economic Review; Significance; Testing

* Corresponding author. Tel.: +1 312 341 3763.
*E-mail addresses:* sziliak@roosevelt.edu (S.T. Ziliak), deirdre2@uic.edu (D.N. McCloskey).

1053-5357/$ – see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.socec.2004.09.024

> Sophisticated, hurried readers continue to judge works on the sophistication of their surfaces. ... I mean only to utter darkly that in the present confusion of technical sophistication and significance, an emperor or two might slip by with no clothes.
>
> Annie Dillard, Living by Fiction
> Harper and Row, New York, 1988 ed., p. 31.

Eight years ago, in "The Standard Error of Regressions," we showed how significance testing was used during the 1980s in the leading general interest journal of the economics profession, the American Economic Review (McCloskey and Ziliak, 1996). The paper reported results from a 19-item "questionnaire" applied to all of the full-length papers using regression analysis. Of the 182 papers 70% did not distinguish statistical significance from policy or scientific significance—that is, from what we call "economic significance" (Question 16, Table 1, p. 105). And fully 96% misused a statistical test in some (shall we say) significant way or another. Of the 70% that flatly mistook statistical significance for economic significance, further, again about 70% failed to report even the magnitudes of influence between the economic variables they investigated (1996, p. 106). In other words, during the 1980s about one-half of the empirical papers published in the *AER* did not establish their claims as *economically* significant.

Some economists have reacted to our finding by saying in effect, "Yes, we know it's silly to think that fit is the same thing as substantive importance; but *we* do not do it: only bad economists do." (Such as, it would seem, the bad ones who publish in the *AER*, an implied evaluation of our colleagues that we do not accept.) And repeatedly in the several score of seminars we have given together and individually on the subject since 1996 we have heard the claim that "after the 1980s, the decade you examined in your 1996 paper, best practice improved."

All the better econometricians we have encountered, of course, agree with our point in substance. This is unsurprising, since the point is obviously true: fit is not *the same thing* as scientific importance; a merely statistical significance cannot substitute for the judgment of a scientist and her community about the largeness or smallness of a coefficient by standards of scientific or policy oomph. As Harold Jeffreys remarked long ago, to reject a hypothesis because the data show "large" departures from the prediction "requires a quantitative criterion of what is to be considered a large departure" (Jeffreys, 1967, p. 384, quoted in Zellner, 1984, p. 277n). Just so. Scientific judgment requires quantitative *judgment*, not endlessly more machinery. As lovely and useful as the machinery is, at the end, having skillfully used it, the economic scientist needs to *judge* its output. But the economists and calculators reply, "Do not fret: things are getting better. Look for example at this wonderful *new* test I have devised."

We are very willing to believe that our colleagues have since the 1980s stopped making an elementary error. But like them we are empirical scientists. And so we applied the same 19-item questionnaire of our 1996 paper to all the full-length empirical papers of the *next* decade of the *AER*, just finished, the 1990s.

Significance testing violating the common sense of first-year statistics and the refined common sense of advanced decision theory, we find here, is not in fact getting better.

Table 1
The American Economic Review had numerous errors in the use of statistical significance, 1980–1999

| Survey question | Percent yes in 1990s | Percent yes in 1980s |
|---|---|---|
| Does the paper ... | | |
| 8. Consider the power of the test? | 8.0 | 4.4 |
| 6. Eschew reporting all standard errors, *t*-, *p*-, and *F*- statistics, when such information is irrelevant? | 12.4 | 8.3 |
| 16. Consider more than statistical significance decisive in an empirical argument? | 18.2 | 29.7 |
| 11. Eschew "sign econometrics," remarking on the sign but not the size of the coefficient? | 19.0 | 46.7 |
| 14. Avoid choosing variables for inclusion solely on the basis of statistical significance? | 25.5 | 68.1 |
| 15. Use other criteria of importance besides statistical significance after the crescendo? | 28.5 | 40.7 |
| 10. Eschew "asterisk econometrics," the ranking of coefficients according to the absolute value of the test statistic? | 32.8 | 74.7 |
| 17. Do a simulation to determine whether the coefficients are reasonable? | 35.0 | 13.2 |
| 7. At its first use, consider statistical significance to be one among other criteria of importance? | 36.5 | 47.3 |
| 19. Avoid using the word "significance" in ambiguous ways? | 37.2 | 41.2 |
| 9. Examine the power function?[a] | 45.5 | 16.7 |
| 18. In the conclusions, distinguish between statistical and economic significance? | 52.6 | 30.1 |
| 13. Discuss the scientific conversation within which a coefficient would be judged large or small? | 54.0 | 28.0 |
| 2. Report descriptive statistics for regression variables? | 66.4 | 32.4 |
| 1. Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample? | 67.9 | 85.7 |
| 12. Discuss the size of the coefficients? | 78.1 | 80.2 |
| 5. 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? | 81.0 | 44.5 |
| 4. Test the null hypotheses that the authors said were the ones of interest? | 83.9 | 97.3 |
| 3. Report coefficients in elasticities, or in some other useful form that addresses the question of "how large is large"? | 86.9 | 66.5 |

*Source:* All the full-length papers using tests of statistical significance and published in the *American Economic Review* in the 1980s ($N = 182$) and 1990s ($N = 137$). Table 1 in McCloskey and Ziliak (1996) reports a small number of papers for which some questions in the survey do not apply.

[a] *Note:* Of the papers that mention the power of a test, this is the fraction that examined the power function or otherwise corrected for power.

It is getting worse. Of the 137 relevant papers in the 1990s, 82% mistook statistically significant coefficients for economically significant coefficients (as against 70% in the earlier decade). In the 1980s, 53% had relied exclusively on statistical significance as a criterion of importance at its first use; in the 1990s 64% did.

## 1. Significance testing is a cheap way to get marketable results

William Kruskal, an eminent statistician long at the University of Chicago, an editor of the *International Encyclopedia of the Social Sciences*, and a former president of the American Statistical Association, agrees. "What happened?" we asked him in a recent interview at his home (William Kruskal, 2002). "Why did significance testing get so badly mixed up, even in the hands of professional statisticians?" "Well," said Kruskal, who long ago had published in the *Encyclopedia* a devastating survey on "significance" in theory and practice (Kruskal, 1968a), "I guess it's a cheap way to get marketable results."

Bingo. Finding statistical significance is simple, and publishing statistically significant coefficients survives at least that market test. But cheap *t*-tests, becoming steadily cheaper with the Moore's-Law fall in computation cost, have in equilibrium a marginal scientific product equal to their cost. Entry ensures it. In the 1996 paper we discussed the history of statistical versus economic significance. Viewed from the sociology and economics of the discipline the notion of statistical significance has been a smashing success. Many careers have prospered on testing, testing, testing (as David Hendry likes to put it). But intellectually the testing has been a disaster, as indeed Edgeworth had warned at the dawn.[1] He corrected Jevons, who had concluded that a "3 or 4%" difference in the volume of commercial bills is not economically important: "[b]ut for the purpose of science, the discovery of a difference in condition, a difference of 3% and much less may well be important" (Edgeworth, 1885, p. 208). It is easy to see why: a statistically *in*significant coefficient in a financial model, for example, may nonetheless give its discoverer an edge in making a fortune; and a statistically *significant* coefficient in the same model may be offset in its exploitation by transactions costs. Statistical significance, to put it shortly, is neither necessary nor sufficient for a finding to be *economically* important. Yet an overwhelming majority of economists, we have shown for the 1980s and now again still more for the 1990s, believe statistical significance *is* necessary; and a simple majority believe it is sufficient.

Economists are skeptics, members of the tribe of Hume. But Ronald Aylmer Fisher (1890–1962), who codified the usage we are objecting to, was a rhetorical magician (as Kruskal once noted, the inventor of such enchanting phrases as "efficiency" and "analysis of variance"; "significance" was older). Long-lived and persistent, he managed to implant for example a "rule of 2" in the minds of economic and other scientists. Listen, for example, as Fisher computes for the masses in 1925 a first test of significance in his *Statistical Methods for Research Workers*:

> The value for which $p = 0.05$ or 1 in 20, is 1.96 or nearly 2; it is *convenient* to take this point 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*. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will

[1] Edgeworth (1885, p. 187), we believe, is the first source of the word "significance" in a context of hypothesis testing. Our earlier paper claimed erroneously that John Venn was first (McCloskey and Ziliak, 1996, p. 97; see Baird, 1988, p. 468). Anyway, the 1880s: for some purposes not a meaningful difference.

> still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.
>
> Fisher 1925 [1941], p. 42; emphasis added).

Notice how a standard of "convenience" rapidly became in Fisher's prose an item to be "formally regarded." With Fisher there's no loss function. There's no thinking beyond the statistic. We are "to take this point as a limit." Fisher's famous and influential book nowhere confronts the difference between scientific and substantive significance (pp. 123–124, 139–140, concerning soporific drugs and algae growth). He provided (and then stoutly defended for the rest of his long life against the decision-theoretic ideas of Neyman, Pearson, and Wald) the cheapest way to get marketable results.

Our policy recommendation is this: that the profession adopt the standards set forth 120 years ago by Edgeworth, and in the years intervening by a small but distinguished list of dissenters from the mechanical standard of 5% (and no loss function about it).

## 2. Practice has improved in a few ways, but not in the crucial matter of significance

Table 1 reports the results distinguished by decade, the 319 full-length papers using regression from January 1980 to December 1999. (We have at hand the whole population, not a sample; the urn of nature is poured out before us; unlike many of our colleagues, therefore, we will refrain from calculating statistics relevant only to inference from *samples* to a population, such as the "statistical significance" of the differences between the two decades.) Like Table 1 in McCloskey and Ziliak (1996) Table 1 here ranks in ascending order each item of the questionnaire according to "Percent Yes." A "yes" means that the paper took what every statistical theorist since Edgeworth (with the significant exception of R.A. Fisher) has regarded as the "correct" action on the matter. For example, in the 1980s 4.4% of the papers considered the power of the tests (and we do not believe it accidental that every paper considering power also considered "a quantitative criterion of what is to be considered a large departure.") That is, 4.4% did the correct thing by considering also the probability of a Type II error. In the 1990s 8% did. That's an encouraging trend.

The change in practice is more easily seen in Tables 2 and 3, which isolate improvement and decline. In the 1980s, only 44.5% of the papers paid careful attention to the theoretical and accounting meaning of the regression coefficients (Question 5). That is, in the 1980s the reader of an empirical paper in the *AER* was nearly six times out of 10 left wondering how to interpret the economic meaning of the coefficients. In the 1990s the share taking the correct action rose to 81%, a net improvement of about 36 percentage points. (This is what we mean by oomph: a big change, important for the science.) Similarly, the percentage of papers reporting units and descriptive statistics for regression variables of interest rose by 34 percentage points, from 32.4 to 66.4% (Question 2). And gains of more than 20 percentage points were made in the share of papers discussing the scientific conversation in which a coefficient would be judged large or small, the share of papers keeping statistical and economic significance distinct in the "conclusions" section, and the share of papers

Table 2
The economic significance of the American Economic Review has in some regards improved (measured by net percentage difference, 1980–1999)

| Survey question | Percent yes in 1990s | Net improvement since 1980s |
|---|---|---|
| Does the paper . . . | | |
| 5. 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? | 81.0 | +36.5 |
| 2. Report descriptive statistics for regression variables? | 66.4 | +34.0 |
| 9. Examine the power function?[a] | 45.5 | +28.8 |
| 13. Discuss the scientific conversation within which a coefficient would be judged large or small? | 54.0 | +26.0 |
| 18. In the conclusions, distinguish between statistical and economic significance? | 52.6 | +22.5 |
| 17. Do a simulation to determine whether the coefficients are reasonable? | 35.0 | +21.8 |

[a] *Notes:* Of the papers that mention the power of a test, this is the fraction that examined the power function or otherwise corrected for power.

Table 3
. . . But the essential confusion of statistical and economic significance is getting worse (measured by net percentage difference, 1980–1999)

| Survey question | Percent yes in 1990s | Net decline since 1980s |
|---|---|---|
| Does the paper . . . | | |
| 14. Avoid choosing variables for inclusionsolely on the basis of statistical significance? | 25.5 | −42.6 |
| 10. Eschew "asterisk econometrics," the ranking of coefficients according to the absolute value of the test statistic? | 32.8 | −41.9 |
| 11. Eschew "sign econometrics," remarking on the sign but not the size of the coefficient? | 19.0 | −27.7 |
| 1. Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample? | 67.9 | −17.8 |
| 4. Test the null hypotheses that the authors said were the ones of interest? | 83.9 | −13.4 |
| 15. Use other criteria of importance besides statistical significance after the crescendo? | 28.5 | −12.2 |
| 16. Consider more than statistical significance decisive in an empirical argument? | **18.2** | **−11.5** |
| 7. At its first use, consider statistical significance to be one among other criteria of importance? | **36.5** | **−10.8** |

*Source:* All the full-length papers using tests of statistical significance and published in the *American Economic Review* in the 1980s ($N = 182$) and 1990s ($N = 137$). Table 1 in McCloskey and Ziliak (1996) reports a small number of papers for which some questions in the survey do not apply.

doing a simulation to determine whether the estimated coefficients are reasonable. (Our definition of "simulation" is broad. It includes papers that check the plausibility of the regression results by making, for example, Harberger-Triangle-type calculations on the basis of descriptive data. But a paper that uses statistical significance as the sole criterion

for *including* a coefficient in a later simulation is coded "No," which is to say that it does *not* do a simulation *to determine whether the coefficients are reasonable*.)

These few gains are commendable. Whether they are scientifically significant is something only we scientists can judge, in serious conversation with each other (for example: that 8% rather than 4% consider power is nice, but still leave 92% of the papers risking high levels of a Type II error). In almost every question (that is, in all except perhaps Question 5 concerning the interpretation of theoretical coefficients, in which the improvement approaches levels that most people would agree are good practice) the improved levels of performance are still less than impressive. For example, in the 1990s *two-thirds* of the papers did not make calculations to determine whether the estimated magnitude of the coefficients made sense (Question 17)—only a third, we found, had simulated the effect of their coefficients with at least the elementary force of Ec 1. Skepticism of alleged effect is by contrast normal practice in sciences like chemistry and physics. (By the way, we 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; and in their own journals the medical scientists, like the social scientists, are hopelessly confused about substantive error as against sampling error. Bald examples of this last may be found in the technical notes enclosed with medicines such as Rogaine.)

Milton Friedman from 1943 to 1945 was a statistician for the Statistical Research Group of the Division of War Research at Columbia University (there is still a non-parametric test named after him). Listen to his experience with statistical versus substantive significance:

> One project for which we provided statistical assistance was the development of high-temperature alloys for use as the lining of jet engines and as blades of turbo superchargers—alloys mostly made of chrome, nickel, and other metals .... Raising the temperature a bit increases substantially the efficiency of the turbine, turbo supercharger, or jet engine .... I computed a multiple regression from a substantial body of data relating the strength of an alloy at various temperatures to its composition. My hope was that I could use the equations that I fitted to the data to determine the composition that would give the best result. On paper, my results were splendid. The equations fitted very well [note: statistically; with high $R^2$] and they suggested that a hitherto untried alloy would be far stronger than any existing alloy .... The best of the alloys at that time were breaking at about 10 or 20 h; my equations predicted that the new alloys would last some 200 h. Really astounding results! ... So, I phoned the metallurgist we were working with at MIT and asked him to cook up a couple of alloys according to my specifications and test them. I had enough confidence in my equations to call them F1 and F2 but not enough to tell the metallurgist what breaking time the equations predicted. That caution proved wise, because the first one of those alloys broke in about 2 h and the second one in about three.
>
> Friedman, 1985, reprinted in Friedman and Schwartz, 1991, pp. 48–49.

Friedman learned that statistical significance is not the same as metallurgical significance.

The core confusion over the meaning of significance testing is reported in Table 3. One problem, which is often taken to be our main objection (it is not, though bad enough on its own), is that statistical *non*significance is nonpublic. In the 1990s only one-fourth of the papers avoided choosing variables *for inclusion* (pretests, that is) solely on the basis of statistical significance, a net *decline* in best practice of fully 43 percentage points (Question 14). As Kruskal put it in his 1968 article.

> Negative results are not so likely to reach publication as are positive ones. In most significance-testing situations a negative result is a result that is not statistically significant, and hence one sees in published papers and books many more statistically significant results than might be expected .... The effect of this is to change the interpretation of published significance tests in a way that is hard to analyze quantitatively (1968a, p. 245).

The response to Question 14 shows that economists made it hard in the 1990s to analyze quantitatively, in Kruskal's sense, the real-world relevance of their "significant" results. It is the problem of searching for significance, which numerous economists have noted, in cynical amusement or despairing indignation, is encouraged by the incentives to publish.

"Asterisk econometrics," the ranking of coefficients according to the absolute value of the test statistic, and "sign econometrics," remarking on the sign but not the size of coefficient, were widespread in the 1980s. But they are now a plague. Eighty-one percent of the papers in the 1990s engaged in what we called "sign econometrics" (in the 1980s, 53% did [Question 11]). In their paper "Tax-based Test of the Dividend Signaling Hypothesis" Bernheim and Wantz (June 1995, p. 543) report that "the coefficients [in four regressions on their crucial variable, high-rated bonds] are all negative .... However, the estimated values of these coefficients," they remark, "are not statistically significant at conventional levels of confidence." The basic problem with sign econometrics, and with the practice of Bernheim and Wantz, can be imagined with two price elasticities of demand for, say, insulin, both estimated tightly, one at size −0.1 and the other at −4.0. Both are negative, and would both be treated as "success" in establishing that insulin use responded to price; but the policy difference between the two estimates is of course enormous. Economically (and medically) speaking, for most imaginable purposes −0.1 is virtually zero. But when you are doing sign econometrics you ignore the *size* of the elasticity, or the *dollar effect* of the bond rating, and say instead, "the sign is what we expected."

Sign econometrics is worse when the economist does not report confidence intervals. Perhaps because they were not trained in the error-regarding traditions of engineering or chemistry, economists seldom report confidence intervals. Thus, Hendricks and Porter, on "The Timing and Incidence of Exploratory Drilling on Offshore Wildcat Tracts" (June 1996, p. 404): "In the first year of the lease term, the coefficient of HERF is positive, but not significant. This is consistent with asymmetries of lease holdings mitigating any information externalities and enhancing coordination, and therefore reducing any incentive to delay." Yet the reader does not know how much "HERF"—Hendricks' and Porter's Herfindahl index of the dispersion of lease holdings among bidders at auction—contributed to the probability the winners would then engage in exploratory oil drilling. In *Life on the Mississippi* Mark Twain noted that "when I was born [the city of] St. Paul had a population of three persons; Minneapolis had just a third as many" (p. 390). The sign is what a St.-Paul-enthusiast would

want and expect. But the sign gives no guidance as to whether a size of 1 is importantly different from 3. No oomph.

About two-thirds of the papers *ranked* the importance of their estimates according to the absolute values of the test statistics, ignoring the estimated size of the economic impact (Question 10). In other words, asterisk econometrics (which is what we call this bizarre but widespread practice), became in the 1990s a good deal more popular in economics (it has long been popular in psychology and sociology), increasing over the previous decade by 43 percentage points. Bernanke and Blinder (1992), Bernheim and Wantz (1995), and Kachelmeier and Shehata (1992), for example, published tables featuring a hierarchy of *p*-, *F*-, and *t*-statistics, the totems of asterisk econometrics (pp. 905, 909; 547, 1130). The asterisk, the flickering star of *, has become a symbol of vitality and authority in economic belief systems. Twenty years ago Arnold Zellner pointed out that economists then (in a sample of 18 articles in 1978) never had "a discussion of the relation between choice of significance levels and sample size" (one version of the problem we emphasize here) and usually did not discuss *how far* from 5% the test statistic was: "there is room for improvement in analysis of hypotheses in economics and econometrics" (Zellner, 1984, pp. 277–280). Yes.

What is most distressing about Table 3, however, is the rising conflation of statistical and economic significance, indicated by the responses to Questions 16 and 7. Our main points are:

- Eighty-two percent of the empirical papers published in the 1990s in the *American Economic Review* did not distinguish statistical significance from economic significance (Question 16). In the 1980s, 70% did not—scandalous enough (McCloskey and Ziliak, 1996, p. 106).
- At the first use of statistical significance, typically in the "Estimation" or "Results" section, 64% in the 1990s did not consider *anything but* the size of the test statistics as a criterion for the inclusion of variables in future work. In the 1980s, 53%—11 percentage points fewer—had done so (Question 7, p. 106).

## 3. Following the wrong decision rule has large scientific costs

Of course, not everyone gets it wrong. The *American Economic Review* is filled with examples of superb economic science (in our opinion most of the papers can be described this way—even though most them, we have seen, make elementary mistakes in the use of statistical significance; in other words, we do *not* accept the opinion of one eminent econometrician we consulted, who dismissed our case by remarking cynically that after all such idiocy is to be expected in the *AER*). Table 4 reports the author rankings by economic significance, in five brackets. If a paper chose between 15 and 19 actions correctly, as Gary Solon's paper did (June 1992), then it is in the top bracket, the best if not perfect practice. If the paper chose between 6 and 8 actions correctly, as Gary Becker, Michael Grossman, and Kevin Murphy did (June 1994), then it is in the fourth bracket, second to last.

Joshua D. Angrist does well in his "The Economic Returns to Schooling in the West Bank and Gaza Strip" (December 1995, pp. 1065–1087). "Until 1972," Angrist writes,

Table 4
Author rankings by economic significance (measured by number yes, that is, good, in the 19-Question Survey of the 1990s) [year and month of publication in brackets]

| | |
|---|---|
| 15–19 | Solon [9206] |
| | Zimmerman [9206] |
| | Goldin [9109] |
| | Craig and Pencavel [9212] |
| | Anderson and Holt [9712] |
| | Ransom [9303] |
| | Allen [9203] |
| | Ausubel [9012] |
| 12–14 | Simon [9812] |
| | Angrist and Evans [9806] |
| | Berk, Hughson, and Vandezande [9609] |
| | Myagkov and Plott [9712] |
| | Gordon and Bovenberg [9612] |
| | Angrist [9512] |
| | Gilligan [9212] |
| | Hoover and Sheffrin [9203] |
| | Benhabib and Jovanovic [9103] |
| | Angrist [9006] |
| | Cecchetti, Lam, and Mark [9006] |
| | Baker and Benjamin [9709] |
| | Paxson [9203] |
| | Blank [9112] |
| | Froot and Obstfeld [9112] |
| 9–11 | Brainerd [9812] |
| | Calomiris and Mason [9712] |
| | Morrison and Schwartz [9612] |
| | Landers, Rebitzer, and Taylor [9606] |
| | Guiso, Jappell, and Terlizzese [9603] |
| | Borjas [9506] |
| | Kaminsky [9306] |
| | Calvo and Leiderman [9203] |
| | Fair and Shiller [9006] |
| | Sauer and Leffler [9003] |
| | Schachar and Nalebuff [9906] |
| | Craft [9812] |
| | Dyck [9709] |
| | Genesove and Meyer [9706] |
| | Pontiff [9703] |
| | Rosenszweig and Wolpin [9412] |
| | Currie and McConnell [9109] |
| | Hendry and Ericsson [9103] |
| | Pitt, Rosenzwieg, and Hassan [9012] |
| | Berry, Levinsohn, and Pakes [9906] |
| | Yano and Nugent [9906] |
| | Ham, Sveinar, and Terrell [9812] |
| | Hallock [9809] |
| | Rajan and Zingales [9806] |
| | Ichnowski, Shaw, and Prennushi [9706] |

Table 4 (*Continued*)

| | |
|---|---|
| | Nalbantian and Schotter [9706] |
| | Wilhelm [9609] |
| | Fuchs [9603] |
| | Rotemberg and Woodford [9603] |
| | Griliches and Cockburn [9412] |
| | James [9309] |
| | Forsythe, Nelson, Neumann, and Wright [9212] |
| | Stratman [9212] |
| | Lin [9203] |
| | Viscusi and Evans [9006] |
| 6–8 | Mendelson, Nordhaus, Shaw [9409] |
| | Fernald [9906] |
| | Gali [9903] |
| | Murray, Evans, and Schwab [9809] |
| | Alesina and Peroti [9712] |
| | Harrigan [9709] |
| | Dorwick and Quiggin [9703] |
| | Chevalier and Scharfstein [9609] |
| | Levin, Kagel, and Richard [9606] |
| | Trefler [9512] |
| | Feldstein [9506] |
| | Mark [9503] |
| | Ashenfelter and Krueger [9412] |
| | Gale and Scholz [9412] |
| | Cohen [9306] |
| | Altonji, Hayashi, and Kotlikoff [9212] |
| | Bernanke and Blinder [9209] |
| | Card [9009] |
| | Aitken and Harrison [9906] |
| | Levine and Zervos [9806] |
| | Blonigen [9706] |
| | Hines [9612] |
| | Henderson [9609] |
| | Laitner and Juster [9609] |
| | Grinblatt, Titman, and Wermers [9512] |
| | Lemieux, Fortin, and Frechette [9403] |
| | Hanes [9309] |
| | Blundell, Pashardes, and Weber [9306] |
| | Kachelmeier and Shehata [9212] |
| | Wolff [9106] |
| | Hardouvelis [9009] |
| | Wright [9009] |
| | Card and Krueger [9409] |
| | Burman and Randolph [9409] |
| | Palfrey and Prisbrey [9712] |
| | Peek and Rosengren [9709] |
| | Levitt [9706] |
| | Cardia [9703] |
| | Hamilton [9703] |
| | Foster and Rosenweig [9609] |
| | Hendricks and Porter [9606] |
| | Ayers and Siegelman [9506] |

Table 4 (*Continued*)

| | |
|---|---|
| | Jones and Kato [9506] |
| | Meyer, Viscusi, and Durbin [9506] |
| | Fuhrer and Moore [9503] |
| | Shea [9503] |
| | Becker, Grossman, and Murphy [9406] |
| | Persson and Tabellini [9406] |
| | Alogoskoufis and Smith [9112] |
| | Fair and Dominguez [9112] |
| <6 | Frankel and Romer [9906] |
| | Krozner and Stratman [9812] |
| | Bernard and Jones [9612] |
| | Munnell, Tootell, Browne, and McEneany [9603] |
| | Attanasio and Browning [9512] |
| | Marin and Schnitzer [9512] |
| | Chevalier [9506] |
| | Currie and Thomas [9506] |
| | Bronars and Grogger [9412] |
| | Krozner and Rajan [9409] |
| | Kim and Singal [9306] |
| | Bronars and Deere [9303] |
| | Kashyap, Stein, and Wilcox [9303] |
| | Falvey and Gemmell [9112] |
| | Keeley [9012] |
| | Ramey and Ramey [9512] |
| | Hamermesh and Biddle [9412] |
| | Keane [9309] |
| | Grossman [9209] |
| | Cukierman, Edwards, and Tabellini [9206] |
| | Wolak and Kolstad [9106] |
| | Keane and Runkle [9009] |
| | Roberts and Tybout [9709] |
| | Engel and Rogers [9612] |
| | Besley and Case [9503] |
| | Levine and Renelt [9209] |
| | Trejo [9109] |
| | Brainard [9709] |
| | Bernheim and Wantz [9506] |

"there were no institutions of higher education in these territories. Beginning in 1972 .... higher education began to open in the West Bank. Previously, Palestinian residents of the territories had to obtain their advanced schooling abroad. But by 1986, there were 20 institutions granting post-high school degrees in the territories. As a consequence, in the early and mid 1980s, the labor market was flooded with new college graduates. This paper studies the impact of this dramatic influx of skilled workers on the distribution of wages in the occupied territories" (p. 1064). In a first regression Angrist estimates the magnitude of wage premia earned by Israelis and Palestinians who work in Israel:

> The first column of Table 2 shows that the daily wage premium for working in Israel fell from roughly 18% in 1981 to zero in 1984. Beginning in 1986, the Israel wage premium rose steeply. By 1989, daily wages paid to Palestinians working in Israel were

> 37% higher than local wages, nearly doubling the 1987 wage differential. The monthly wage premium for working in Israel increased similarly. These changes parallel the pattern of Palestinian absences from work and are consistent with movements along an inelastic demand curve for Palestinian labor (p. 1072).

The reader is told magnitudes. She knows the oomph.

Yet even Angrist falls back into asterisk econometrics. On page 1079 he is testing alternative models, and emphasizes that:

> The alternative tests are not significantly different in five out of nine comparisons ($p < 0.02$), but the joint test of coefficient equality for the alternative estimates of $\theta_t$ leads to rejection of the null hypothesis of equality (p. 1079).

To which his better nature would say, "So?"

David Zimmerman, in his "Regression Toward Mediocrity in Economic Stature" (1992), and especially the well-named Gary Solon, in his "Intergenerational Income Mobility in the United States" (1992), have set an admirable if rare standard for the field. Line by line Solon asks the question "How much?" and then gives an answer. How much, he wonders, is a son's economic well-being fated by that of his father? The sign, the star, the sign-and-the-star-together, do not tell. Previous estimates, observes Solon, had put the father-son income correlation at about 0.2 (p. 394). A new estimate, a tightly fit correlation of $0.20000000001^{***}$, would say nothing new of *economic* significance. And a poorly fit correlation with the "expected sign" would say nothing. Nothing at all. Solon's attempts at a new estimate, on pages 397–405, refer only once to statistical significance (p. 404). Instead, Solon writes 18 paragraphs on *economic* significance: why he believes the "intergenerational income correlation in the United States is [in fact] around 0.4" (p. 403) and how the higher correlation changes American stories about mobility. Solon's paper is three standard deviations above the average of the *AER*.

"Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania" by David Card and Alan B. Krueger (1994a), is above the median in several measures of scientific seriousness. Card and Krueger designed their own surveys, collected their own data, talked on the telephone with firms in their sample, and visited firms that did and did not respond to their survey, all of which is most unusual among economists, and seems to have raised scientific standards in the field. It matches the typical procedure in economic history, for example, or the best in empirical sociology and experimental physics. Their sample was designed to study prices, wages, output, and employment in the fast food industry in Eastern Pennsylvania and Western New Jersey before and after New Jersey raised its minimum wage above the national and Pennsylvania levels. On pages 775–776 of the article (and pages 30–33 in their widely cited book [1994b]), Card and Krueger report their crucial test of the conventional labor market model. The chief prediction of the conventional model is that full-time equivalent employment in New Jersey relative to Pennsylvania would fall following the increase in the New Jersey minimum wage. Specifically Card and Krueger's null hypothesis says that the difference-in-difference is zero—that "change in employment in New Jersey" minus "change in employment in Pennsylvania" should equal zero if as they suppose the minimum wage is *not* oomphul. If they find that the difference-in-difference is zero (other things equal), then New Jersey gets the wage gains

without loss of employment: a good thing for workers. Otherwise, New Jersey employment under the raised minimum wage will fall, perhaps by a lot: a bad thing for workers.

Yet Card and Krueger do not test the null they have claimed. They test two distinct nulls, "change in employment in New Jersey = zero" and (in a separate test) "change in employment in Pennsylvania = zero." In other words, they compute *t*-tests for each state, examining average full-time equivalent employment before and after the increase in the minimum wage. But they do not test here the (relevant) difference-in-difference null of zero. Card and Krueger report on page 776 a point estimate suggesting employment in New Jersey increased by "0.6" of a worker per firm (from 20.4 to 21; rather than falling as enemies of the minimum wage would have expected). Then they report a second point estimate suggesting that employment in Pennsylvania fell by 2.1 workers per firm (from 23.3 to 21.2). "Despite the increase in wages," they conclude from the estimates, "full-time equivalent employment *increased* in New Jersey relative to Pennsylvania. Whereas New Jersey stores were initially smaller, employment gains in New Jersey coupled with losses in Pennsylvania led to a small and statistically insignificant interstate difference in wave 2" (776; their emphasis). The errors are multiple: Card and Krueger run the wrong test (testing the wrong null, by the way, was less common in the *AER* during the 1980s [Table 1, Question 4]); they "reject" a null of zero change in employment in New Jersey, having found an average difference, estimated noisily at $t = 0.2$, of 0.6 workers per firm; they do not discuss the power of their tests, though the Pennsylvania sample is larger by a factor of 5; they practice asterisk econometrics (with a "small and statistically insignificant interstate difference"); and yet they emphasize *acceptance* of their favored alternative, with italics. Further attempts to measure with multiple regression analysis the size of the employment effect, the price effect, and the output effect, though technically improved, are rarely argued in terms of economic significance.

The cost of following the wrong decision rule is especially clear in "An Empirical Analysis of Cigarette Addiction" by Gary Becker, Michael Grossman, and Kevin Murphy (June 1999; 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.

> Our estimation strategy is to begin with the myopic model. We then test the myopic model by testing whether future prices are significant predictors of current consumption as they would be in the rational-addictive model, but not under the myopic model (p. 403). . .. According to the parameter estimates of the myopic model presented in Table 2, cigarette smoking is inversely related to current price and positively related to income.

And then: "the highly significant effects of the smuggling variables (ldtax, sdimp, and sdexp) indicate the importance of interstate smuggling of cigarettes."

But as Kruskal put it, echoing Neyman and Pearson from 1933, "the adverb 'statistically' is often omitted, and this is unfortunate, since statistical significance of a sample bears no necessary relationship to possible subject-matter significance of whatever true departure from the null hypothesis might obtain" (Kruskal, 1968a, p. 240). At $N$ = about 1400 with high power they can reject a nearby alternative to the null—an alternative different, *but trivially different*, from the null (at high sample sizes, after all $s/\sqrt{N}$ approaches zero: all

hypotheses are rejected, and in mathematical fact, without having to look at the data, you know they will be rejected at any pre-assigned level of significance). Yet they conclude that "the positive and significant past-consumption coefficient is consistent with the hypothesis that cigarette smoking is an addictive behavior" (p. 404). It is sign econometrics, with policy implications. When sign econometrics meets asterisk econometrics the mystification redoubles:

> When the one-period lead of price is added to the 2SLS models in Table 2, its coefficient is negative and significant at all conventional levels. The absolute $t$ ratio associated with the coefficient of this variable is 5.06 in model (i) 5.54 in model, (ii) and 6.45 in model (iii). These results suggest that decisions about current consumption depend on future price. They are inconsistent with a myopic model of addiction, but consistent with a rational model of this behavior in which a reduction in expected future price raises expected future consumption, which in turn raises current consumption. While the tests soundly reject the myopic model [and so forth] (p. 404).

Eventually they report (though never interpret) the estimated magnitudes of the price elasticities of demand for cigarettes. But their way of finding the elasticities is erroneous. Cigarette smoking may be addictive. But Becker, Grossman, and Murphy have not shown why, or how much. (They are, incidentally, inferring individual behavior from state-wide data; sociologists call this the ecological fallacy.) Perhaps what they have shown is that statistics play multiple roles:

> There are some other roles that activities called "statistical" may, unfortunately, play. Two such misguided roles are (1) to sanctify or provide seals of approval (one hears, for example, of thesis advisors or journal editors who insist on certain formal statistical procedures, whether or not they are appropriate); (2) to impress, obfuscate, or mystify (for example, some social science research papers contain masses of undigested formulas [or tests of significance] that serve no purpose except that of indicating what a bright fellow the author is).
>
> Kruskal (1968b), p. 209.

Table 5 shows what happens if statistical significance is the only criterion of importance at first use. In a large number of cases, if only statistical significance is said to be of importance as its first use, then statistical significance tends to decide the entire empirical argument. Of the 137 full length papers in the 1990s, 80 papers made both mistakes (Question 7 = 0 and Question 16 = 0). To put it differently, of the 87 papers using only statistical significance as a criterion of importance at first use, fully 92% considered statistical significance the last word. Cross tabulations on the 1980s data reveal a similar though slightly better record (Table 5).

## 4. We are not original

We are not the first social scientists to make the distinction between economic and statistical significance. One of us has been making the point since 1985 (McCloskey, 1985a,

Table 5
If only statistical significance is said to be of importance at its first use (Question 7), then statistical significance tends to decide the entire argument

| In the 1990s . . . | | Does not consider the test decisive (Question 16) | | |
|---|---|---|---|---|
| | | 0 | 1 | |
| Considers more than the test at the first use (Question 7) | 0 | 80 | 7 | 87 |
| | 1 | 32 | 18 | 50 |
| | Total | 112 | 25 | 137 |

*Notes:* '0' means "no, did the wrong thing;" '1' means "yes, did the right thing." In the 1980s data, when Question 7 = 0 and Question 16 = 0 the first row is by contrast 86-10-96 [McCloskey and Ziliak, 1996, Tables 1 and 5]; in other words, practice was perhaps in this additional sense slightly better in the 1980s.

1985b, 1992, 1995), but she learned it from a long, long line of distinguished if lonely protesters of the arbitrary procedures laid down in the 1920s by the blessed Fisher. We have pointed out before that in the 1930s Neyman and Pearson and then especially Abraham Wald had distinguished sharply between practical and statistical significance (McCloskey and Ziliak, 1996, pp. 97–98; McCloskey, 1985a). But Wald died young, and Neyman and Pearson carried the day only at the level of high-brow statistical theory (and Fisher we have just noted failed to measure or mention the matters of substantive significance that occupied Wald and Neyman and Pearson [Fisher, 1925 (1941), pp. 42, 123–124, 138–140, 326–329]). Statistical practice on the ground stayed with a predetermined level of 5% significance (mainly), regardless of the loss function, misleading even the Supreme Court of the United States.

Yet some simple souls got it right. Educators have written about the difference between substantive and statistical significance early and late (Tyler, 1931; Shulman, 1970; Carter, 1978). Psychologists have known about the difference for nearly a century, though most of them continue like economists to ignore it (a committee of the American Psychological Association was recently charged to re-open the question). In 1919, an eminent experimental psychologist, the alarmingly named Edwin Boring, published an article unmasking the confusion between arbitrarily-judged-statistical significance and practical significance (Boring, 1919). And empirical sociology would be less easy for economists to sneer at if more realized that a good many sociologists grasped the elementary statistical point decades before even a handful of the economists did (Morrison and Henkel (Eds.), 1970).

Of late the protest has grown a little louder, but is still scattered (we detailed in the 1996 paper the evidence that almost all econometrics textbooks teach the students to ignore substantive significance in favor of testing without a loss function and without substantive judgments of the size of coefficients). James Berger and Robert Wolpert in 1988, though making a slightly different point (the Bayesian one that Jeffreys and Zellner emphasize), noted the large number of theoretical statisticians engaging in "discussions of important practical issues such as 'real world' versus 'statistical' significance": Edwards et al. (1963), Good (1981), and the like. What we find bizarre is that in the mainstream statistical literature this "important" point is hardly mentioned (we found

in our 1996 article, though, some honorable exceptions, such as the first edition of the elementary text by Freedman, Pisani, and Purvis [1978; we note with alarm that later editions have soft-peddled the issue]. See also Schlaifer (1959), Raiffa and Schlaifer (1961), and Simonoff (2003) (with no soft-peddling about it). Among economists the roll of honor is likewise short but distinguished. J.M. Keynes (virtually), Oskar Lange, Arnold Zellner, Arthur Goldberger, A.C. Darnell, Clive Granger, Edward Leamer, Milton Friedman, Robert Solow, Kenneth Arrow, Joel Horowitz, Zvi Griliches, Jack Hirshleifer, Glen Cain, Gordon Tullock, Gary Solon, Daniel Hamermesh, Thomas Mayer, David Colander, Jeffrey Wooldridge, Jan Magnus, and Hugo Keuzenkamp are not dunces and they have not minced words (Lange, 1959 [1978], pp. 13–15, 133–157 [on page 151 Lange speaks of "practical significance," his main concern]; Cain and Watts (1970), pp. 229, 231–232; Keuzenkamp and Magnus, 1995; McCloskey and Ziliak, 1996, p. 99 and numerous other references on pp. 112–114; McCloskey's citations in her works cited; Darnell's comprehensive review of 1997; Hamermesh, 1999; Colander, 2000; Wooldridge, 2000, pp. 131–134; Keuzenkamp, 2000, p. 266; Hirshleifer, 2004; Ziliak, 2004; and so forth). Recently, to pick one among the small, bright stream of revisions of standard practice that appear in our mailboxes, Clinton Greene (2003) has applied the argument to time-series econometrics, showing that tests of cointegration based on arbitrary levels of significance miss the economic point: they are neither necessary nor sufficient.

We are sometimes told that "You are rehashing issues decided in the 1950s" or "Sure, sure: but the hot *new* issue is [such and such new form of specification error, say]" or "I have a metaphysical argument for why a universe should be viewed like a sample." When we are able to get such people-in-a-hurry to slow down and listen to what we are saying (which is not often), we discover that in fact they do *not* grasp our main point, and their own practice shows why. It is dangerous, for example, to mention Bayes in this connection, because the reflexive reply of most econometrically minded folk is to say "1950s" and have done with it. Our point is not Bayesian (although we honor the Bayesians such as Leamer and Zellner who have made similar—and also some different—criticisms of econometric practice). Our (idiotically simple) point has nothing to do with Bayes' Theorem: it applies to the most virginal classical regressions.

Our experience is that in the rare cases when people *do* grasp our point—that fit and importance are not the same—they are appalled. They realize that almost everything that has been done in econometrics since the beginning needs to be redone. The wrong variables have been included, for example (which is to say errors in specification have vitiated the conclusions); tiny coefficients have inflated reputations; mistaken policies have been recommended; science has stopped.

We believe we have shown from our evidence in the *American Economic Review* over the two last decades what scientists from Edgeworth to Goldberger have been saying: science is about magnitudes. Seldom is the magnitude of the sampling error the chief scientific issue. (A sympathetic reader might reply it is not the size that counts; it is what you do with it. But that too is mistaken. As Friedman's alloy regression and hundreds of other statistical experiments reveal, what matters is size *and* what you do with it. Scientific judgment, like any judgment, is about loss functions—what R.A. Fisher was most persistent in denying.)

## 5. What should economists do?

We should act more like the Gary Solons and the Claudia Goldins. We should be economic scientists, not machines of walking dead recording 5% levels of significance. In his acceptance speech for the Nobel Prize, Bob Solow put it this way:

> [Economists] should try very hard to be scientific with a small s. By that I mean only that we should think logically and respect fact .... Now, I want to say something about fact. The austere view is that "facts" are just time series of prices and quantities. The rest is all hypothesis testing. I have seen a lot of those tests. They are almost never convincing, primarily because one senses that they have very low power against lots of alternatives. There are too many ways to explain a bunch of time series .... My hunch is that we can make progress only by enlarging the class of eligible facts to include, say, the opinions and casual generalizations of experts and market participants, attitudinal surveys, institutional regularities, even our own judgments of plausibility (Solow, 1988).

Solow recommends we "try very hard to be scientific with a small s"; the authors we have surveyed in the *AER*, by contrast, are trying very hard to be scientific with a small *t*. As Solow says, it's almost never convincing.

What to do? One of us was advised to remove the 1996 article from his CV while job hunting—it was not "serious" research. Shut up and follow R.A. Fisher. The other served fleetingly on the editorial board of the *AER*. Each time she saw the emperor had no clothes of oomph she said so (by the way, in the original Danish of the story the child is *not* identified as to gender: we think it was probably a little *girl*.) The behavior did not endear her to the editors. After a while she and they decided amiably to part company.

The situation is strange: economic scientists, for example those who submit to and publish papers in the *AER*, or serve on hiring committees, routinely violate elementary standards of statistical cogency. And yet it is the messengers who are to be taken out and shot. This should stop. We should revise publication standards, and cease shooting messengers who bring the old news that fit is not the same thing as importance. If the *AER* were to test papers for cogency, and refused to publish papers that used fit irrelevantly as a standard of oomph, economics would in a few years be transformed into a field with empirical standards. At present (we can say until someone starts claiming that in the 2000s practice has improved), we have shown, it has none. Ask: "Is the paper mainly about showing and measuring *economic* significance?" If not, the editor and referees should reject it. It will not reach correct scientific results. Its findings will be biased by misspecification and mistaken as to oomph. (Requiring referees to complete a 19-item questionnaire would probably go against the libertarian grain of the field; a short form would do: "Does the paper focus on the *size* of the effect it is trying to measure, or does it instead recur to irrelevant tests of the coefficient's *statistical* significance?") To do otherwise—continuing to decorate our papers with stars and signs while failing to interpret size—is to discard our best unbiased estimators, and to renege on the promise of empirical economics: measurement. No size, we should say, no significance.

## Acknowledgements

We thank for their amazed attention to the present paper audiences at Baruch College (CUNY), the University of Colorado at Boulder, Denison University, the Georgia Institute of Technology, the University of Georgia, the University of Illinois at Chicago, Macquerie University, the University of Wollongong, the summer institute over many years of EDAMBA, the annual meetings of the Eastern Economic Association (2003), of the Economic Society of Australia (2003), of the Association for Heterodox Economics (University of Leeds, 2004), of the American Economic Association (together with the Association for Social Economics) 2004, and the ICAPE Conference on "The Future of Heterodox Economics" (University of Missouri-Kansas City, 2003). For their generous participation in a session of the 2004 meetings of the AEA, the basis of this special volume of the *JSE*, we thank especially Morris Altman, Kenneth Arrow, Clive Granger, Joel Horowitz, Ed Leamer, Tony O'Brien, Eric Thorbecke, and Arnold Zellner and, for their comments thereabouts, Stephen Cullenberg, David Hendry, and Jack Hirshleifer. Cory Bilton, David McClough, and Noel Winter provided excellent research assistance and we would like to thank them. We dedicate this paper to William Kruskal.

## References

American Economic Review, January 1980 to December 1999. The 319 full-length papers using tests of statistical significance. May Supplement excluded.

Angrist, J., 1995. The economic returns to schooling in the West Bank and Gaza Strip. American Economic Review 85 (5), 1065–1086.

Baird, D., 1988. Significance tests, history and logic. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences, vol. 8. John Wiley, New York, pp. 466–471.

Becker, G.S., Grossman, M., Murphy, K.M., 1994. An empirical analysis of cigarette addiction. American Economic Review 84 (3), 396–418.

Berger, J.O., Wolpert, R.L., 1988. The Likelihood Principle, second ed. Institute of Mathematical Statistics, Hayward, CA.

Bernanke, B.S., Blinder, A.S., 1992. The federal funds rate and the channels of monetary transmission. American Economic Review 82 (4), 901–921.

Bernheim, B.D., Wantz, A., 1995. A tax-based test of the dividend signaling hypothesis. American Economic Review 85 (3), 532–551.

Boring, E.G., 1919. Mathematical versus Scientific Significance. Psychological Bulletin 16 (10), 335–338.

Cain, G.G., Watts, H.W., 1970. Problems in making policy inferences from the Coleman report. American Sociological Review 35 (2), 228–242.

Card, D., Krueger, A.B., 1994a. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. American Economic Review 84 (2), 772–793.

Card, D., Krueger, A.B., 1994b. Myth and Measurement: the New Economics of the Minimum Wage. Princeton University Press, Princeton.

Carter, R.P., 1978. The case against statistical significance testing. Harvard Educational Review 48 (3), 378–398.

Colander, D., 2000. New millenium economics: how did it get this way, and what way is it? Journal of Economic Perspectives 14 (1), 121–132.

Darnell, A.C., 1997. Imprecise tests and imprecise hypotheses. Scottish Journal of Political Economy 44 (3), 247–268.

Edgeworth, F.Y., 1885. Methods of statistics. Jubilee Volume of the Statistical Society, 181–217, Royal Statistical Society of Britain, 22–24 June.

Edwards, W., Lindman, H., Savage, L.J., 1963. Bayesian statistical inference for psychological research. Psychological Review 70 (3), 193–242.
Fisher, R.A., 1925. Statistical Methods for Research Workers, Eighth ed. G.E. Stechart and Co., New York.
Friedman, M., 1985. In: W. Breit, R.W. Spencer (Eds.), Lives of the Laureates. MIT Press, Cambridge, 1990. Selection reprinted: M. Friedman, A.J. Schwartz, 1991. Alternative approaches to analyzing data. American Economic Review 81(1), 39–49, 77–92.
Good, I.J., 1981. Some logic and history of hypothesis testing. In: Pitt, J.C. (Ed.), Philosophy in Economics. Reidel, Dordrecht, The Netherlands.
Greene, C.A., 2003. Towards Economic Measures of Cointegration and Non-Cointegration. Department of Economics, University of Missouri, St. Louis. April. clinton_greene@umsl.edu, unpublished paper.
Hamermesh, D.S., 1999. The Art of Labormetrics. National Bureau of Economic Research, Inc., Cambridge, MA.
Hendricks, K., Porter, R.H., 1996. The timing and incidence of exploratory drilling on offshore wildcat tracts. American Economic Review 86 (3), 388–407.
Hirshleifer, J., 2004. Personal Letter of Communication. University of California, Los Angeles, January 5.
Jeffreys, H., 1967. Theory of Probability, Third revised. Oxford University Press, London.
Kachelmeier, S.J., Shehata, M., 1992. Examining risk preferences under high monetary incentives: experimental evidence from the People's Republic of China. American Economic Review 82 (5), 1120–1141.
Keuzenkamp, H.A., 2000. Probability, Econometrics and Truth. Cambridge University Press, Cambridge.
Keuzenkamp, H.A., Magnus, J., 1995. On tests and significance in econometrics. Journal of Econometrics 67 (1), 103–128.
Kruskal, W.S., 2002. Personal Interview. University of Chicago, 16 August.
Kruskal, W.S., 1968a. Tests of statistical significance. In: David Sills (Ed.), International Encyclopedia of the Social Sciences, vol. 14. MacMillan, New York, pp. 238–250.
Kruskal, W.S., 1968b. Statistics: the field. In: David Sills (Ed.), International Encyclopedia of the Social Sciences, vol. 15. MacMillan, New York, pp. 206–224.
Lange, O., 1959. Introduction to Econometrics. PWN—Polish Scientific Publishers/Pergamon Press, Warsaw/Oxford, England.
McCloskey, D., Ziliak, S., 1996. The standard error of regressions. Journal of Economic Literature 34, 97–114.
McCloskey, D., 1985a. The loss function has been mislaid: the rhetoric of significance tests. American Economic Review Supplement 75 (2), 201–205.
McCloskey, D., 1995. The insignificance of statistical significance. Scientific American, 32–33.
McCloskey, D., 1992. The bankruptcy of statistical significance. Eastern Economic Journal 18, 359–361.
McCloskey, D., 1985b. The Rhetoric of Economics. University of Wisconsin Press, Madison, Especially chapters 8 and 9.
Morrison, D.E., Henkel, R.E., 1970. The Significance Test Controversy: A Reader. Aldine, Chicago.
Raiffa, H., Schlaifer, R., 1961. Applied Statistical Decision Theory. Harpercollins, New York.
Schlaifer, R., 1959. Probability and Statistics for Business Decisions. McGraw Hill, New York.
Shulman, L.S., 1970. Reconstruction of educational research. Review of Educational Research 40, 371–393.
Simonoff, J., 2003. Analyzing Categorical Data. Springer-Verlag, New York.
Solon, G., 1992. Intergenerational income mobility in the United States. American Economic Review 82 (3), 393–408.
Solow, R., 1988. In: Breit, W., Spencer, R.W. (Eds.), Lives of the Laureates. MIT Press, Cambridge.
Twain, M., 1883. Life on the Mississippi. Bantam, New York.
Tyler, R.W., 1931. What is statistical significance? Educational Research Bulletin 10, 115–118, 142.
Wooldridge, J.M., 2000. Introductory Econometrics. South-Western College Publishing (Thomson Learning), USA.
Zellner, A., 1984. Basic Issues in Econometrics. University of Chicago Press, Chicago.
Ziliak, S.T., 2004. Why I Left Alan Greenspan to Seek Economic Significance: The Confessions of an α-Male, Rethinking Marxism, forthcoming.
Zimmerman, D.J., 1992. Regression toward mediocrity in economic stature. American Economic Review 82 (3), 409–429.

R00083788_SOCECO_368