Journal of Economic Literature
Vol. XXXIV (March 1996), pp. 97–114

The Standard Error of Regressions

By DEIRDRE N. MCCLOSKEY
and
STEPHEN T. ZILIAK
University of Iowa

Suggestions by two anonymous and patient referees greatly improved the paper. Our thanks also to seminars at Clark, Iowa State, Harvard, Houston, Indiana, and Kansas State universities, at Williams College, and at the universities of Virginia and Iowa. A colleague at Iowa, Calvin Siebert, was materially helpful.

THE IDEA OF statistical significance is old, as old as Cicero writing on forecasts (Cicero, *De Divinatione*, I. xiii. 23). In 1773 Laplace used it to test whether comets came from outside the solar system (Elizabeth Scott 1953, p. 20). The first use of the very word "significance" in a statistical context seems to be John Venn's, in 1888, speaking of differences expressed in units of probable error:

> They inform us which of the differences in the above tables are permanent and significant, in the sense that we may be tolerably confident that if we took another similar batch we should find a similar difference; and which are merely transient and insignificant, in the sense that another similar batch is about as likely as not to reverse the conclusion we have obtained. (Venn, quoted in Lancelot Hogben 1968, p. 325).

Statistical significance has been much used since Venn, and especially since Ronald Fisher.

The problem, and our main point, is that a difference can be permanent (as Venn put it) without being "significant" in other senses, such as for science or policy. And a difference can be significant for science or policy and yet be insignificant statistically, ignored by the less thoughtful researchers.

In the 1930s Jerzy Neyman and Egon S. Pearson, and then more explicitly Abraham Wald, argued that actual investigations should depend on substantive not merely statistical significance. In 1933 Neyman and Pearson wrote of type I and type II errors:

> Is it more serious to convict an innocent man or to acquit a guilty? That will depend on the consequences of the error; is the punishment death or fine; what is the danger to the community of released criminals; what are the current ethical views on punishment? From the point of view of mathematical theory all that we can do is to show how the risk of errors may be controlled and minimised. The use of these statistical tools in any given case, in determining just how the balance should be struck, *must be left to the investigator.* (Neyman and Pearson 1933, p. 296; italics supplied)

Wald went further:

> The question as to how the form of the weight [that is, loss] function . . . should be determined, *is not a mathematical or statistical one.* The statistician who wants to test

> certain hypotheses must first determine the relative importance of all possible errors, *which will depend on the special purposes of his investigation*. (1939, p. 302, italics supplied)

To date no empirical studies have been undertaken measuring the use of statistical significance in economics. We here examine the alarming hypothesis that ordinary usage in economics takes statistical significance to be the same as economic significance. We compare statistical best practice against leading textbooks of recent decades and against the papers using regression analysis in the 1980s in the *American Economic Review*.

I. *An Example*

The usual test of purchasing power parity regresses prices at home (P) on prices abroad ($P*$), allowing for the exchange rate (e). Thus: $P = a + \beta\ (eP*)$ + error term (cf., McCloskey and J. Richard Zecher 1984). The equation can be in levels or rates of change or in some more complex functional form. An estimated coefficient β is of course a random variate, and the accuracy of its estimated mean depends on the properties of the error term, the specification of the model, and so forth. But to fix ideas suppose that all the usual econometric problems have been solved. In tests of purchasing power parity the null hypothesis is usually thought of as "β equal to 1.0." Suppose an unbiased estimator of β yields not exactly 1.000 but very close, say 0.999. That is, prices at home rise by very nearly the same rate as prices abroad for most purposes of science or policy. (Not for all purposes: the point of thinking as Wald did in terms of loss functions is that for *some* purposes a difference that in some metric looks "small" might in another metric be important.) If the sample size were large enough, however, even a coefficient of 0.999 might prove to be a *statistically* significant divergence from exactly 1.000. Under purely statistical procedures the investigator would conclude, as many have, that the hypothesis of purchasing power parity had failed.

The hypothesis does not in truth predict that the coefficient will be 1.000 to many decimal places. It predicts that β will be "about 1." The economically relevant null hypothesis is a range around 1.000, not the point itself in isolation from its neighborhood. The investigator would not want to assert that if $\beta = 0.999$ with a standard error of 0.00000001 we should abandon purchasing power parity, or run our models of the American economy without the world price level. Yet the literature on purchasing power parity has ordinarily used the null of 1.000 exactly. The procedure is not defensible in statistical theory. The table of t will not tell what is "close." Closeness depends, in Wald's words, on the special purposes of the investigation—good enough for inflation control, say, if $\beta = 0.85$, though not good enough to make money on the foreign exchange market unless $\beta = 0.99998$.

Just how the balance should be struck, as Neyman and Pearson put it, must be left to the investigator. A coefficient of 0.15, say, would for most purposes reject "β = about 1." Accepting the null hypothesis may be reasonable or unreasonable, but it depends on economic context. The point is that it does not mainly depend on the value of the test statistic. The uncertainty of the estimate that arises from sampling error—the only kind of uncertainty the test of significance deals with—is still of scientific interest. But low or high uncertainty (more or less "permanence" in Venn's terms) does not by itself answer the question how important the variable is, how large is large. In tests of purchasing power parity, for example, one should ask if $\beta =$

0.999 is close enough for scientific purposes to the null. How should the answer be adjusted if there are 20,000 observations (cf. Richard Rudner 1953, p. 3; Scott Gordon 1991, pp. 664–65)? If the estimate is not taken to be close to the null, what makes it "interestingly different" or what is the "scientific intuition" of one's "public" (Edwin Boring 1919, p. 337)? These are not easy questions. But they are the questions relevant to scientific discovery.

II. *The Evidence: Textbooks*

The late Morris DeGroot ([1975] 1989), a statistician with sophistication in economics, was emphatic on the point:

> It is extremely important . . . to distinguish between an observed value of U that is statistically significant and an actual value of the parameter . . . In a given problem, the tail area corresponding to the observed value of U might be very small; and yet the actual value . . . might be so close to [the null] that, for practical purposes, the experimenter would not regard [it] as being [substantively] different from [the null]. (p. 496)
>
> [I]t is very likely that the t-test based on the sample of 20,000 will lead to a statistically significant value of U . . . [The experimenter] knows in advance that there is a high probability of rejecting [the null] even when the true value . . . differs [arithmetically] only slightly from [the null]. (p. 497)

But few other econometrics textbooks distinguish economic significance from statistical significance. And fewer emphasize economic significance. In the econometrics texts widely used in the 1970s and 1980s, when the practice was becoming standard, such as Jan Kmenta's *Elements of Econometrics* (1971) and John Johnston's *Econometric Methods* ([1963] 1972, 1984), there is no mention of economic as against statistical significance. Peter Kennedy, in his *A Guide to Econometrics* (1985), briefly mentions that a large enough sample always gives statistically significant differences. This is part of the argument but not all of it, and Kennedy in any case relegates the partial argument to an endnote (p. 62).

Among recent econometrics books Arthur Goldberger's is the most explicit. His *A Course in Econometrics* (1991) gives the topic "Statistical versus Economic Significance" a page of text (pp. 240–41), quoting McCloskey's little article of 1985. Goldberger's page has been noticed as unusual. Clive Granger, reviewing in the March 1994 issue of this *Journal* four leading books (Goldberger; Russell Davidson and James G. MacKinnon 1993; William H. Greene 1993; William E. Griffiths, R. Carter Hill, and George G. Judge 1993), notes that

> when the link is made [in Goldberger between the economics and the technical statistics] some important insights arise, as for example the section discussing "statistical and economic significance," a topic not mentioned in the other books. (1994, p. 118)

That is, most beginning econometrics books even now, unlike DeGroot and Goldberger and before them the modern masters of statistics, do not contrast economic and statistical significance.

Nor do the present-day advanced handbooks and textbooks. The three volumes of the *Handbook of Econometrics* contain one mention of the point, unsurprisingly by Edward Leamer (p. 325 of Volume I, Zvi Griliches and Michael D. Intriligator, ed. 1983). In the 762 pages of the recent companion work, Volume 11 of the *Handbook of Statistics* (1993), there is one sentence about the level of the test in its relation to sample size (Jean-Pierre Florens and Michel Mouchart 1993, p. 321).

One might defend contemporary usage by arguing that the advanced texts assume their readers already grasp the difference between economic and statistical significance. Economy of style would dictate the unqualified word "significance," its exact meaning, economic or

statistical, to be supplied by the sophisticated reader. Under such a hypothesis the contemporary usage would be no more than a shorthand way to refer to an estimated coefficient. The implied reader would be educated enough to supply the appropriate caveats about *economic* significance.

The hypothesis is not borne out by the evidence. To take one example among many, Takeshi Amemiya's advanced textbook in econometrics does not itself draw a distinction between economic and statistical significance (Amemiya 1985). The book makes little claim to teaching empirical methods, but presumably the theory of econometrics is supposed to connect to empirical work. Amemiya recommends that the student prepare "at the level of Johnston, 1972" (preface). Does the recommendation cover the matter of statistical versus substantive significance?

No. Johnston as we have seen makes no mention of the point. He uses the term "economic significance" once only, without contrasting it to the statistical significance on which he lavishes attention: "It is even more difficult to attach economic significance to the linear combinations arising in canonical correlation analysis than it is to principal components" (p. 333). In an extended example of hypothesis testing, spanning pages 17 to 43, Johnston tests in the conventional way the hypothesis that "sterner penalties" for dangerous driving caused fewer road deaths, concluding "[t]he computed value [of the *t*-statistic] *is suggestive of a reduction*, being *significant* at the 5 per cent, *but not at the one per cent*, level" (p. 43, italics supplied). He is saying that at a high level of rigor the policy of sterner penalties might be doubted to have desirable effects. Statistically the usage is unobjectionable (except that he uses the universe of road casualties in the United Kingdom 1947–1957 as though it were a sample of size 11 from some universe). But the 100,000 lives that were saved in the reduction as measured are not acknowledged as "significant." Johnston has merged statistical and policy significance. At what level the significance level should be set, considering the human cost of ignoring the effect of sterner penalties, is none of Johnston's concern. He leaves the question of how large is large to statistics. As Wald said in 1939, however, the question "is not a mathematical or statistical one."

Johnston does recommend "The Cairncross Test" (1984, pp. 509–10). That is, after computing assorted test statistics the researcher should ask if the model would satisfy the discerning judgment of Sir Alec Cairncross. "Would Sir Alec be willing to take this model to the Riyadh?" But that is our point. If judgments about economic significance are not made at the keyboard they need to be brought into the open, before reaching Sir Alec. The researcher wastes the time of Cairncross if the statistically significant does not correspond to what Sir Alec, and the Riyadh, want: economic significance.

A tenacious defender of contemporary usage might argue further that Johnston, in turn, presumes the reader already understands the difference between economic and statistical significance, having acquired it in elementary courses on statistics. The argument is testable. In his preface Johnston directs the reader who has difficulty with his first chapter to examine a "good introductory" book on statistics, mentioning Paul G. Hoel's *Introduction to Mathematical Statistics* (1954), Alexander M. Mood's *Introduction to the Theory of Statistics* (1950), and Donald A. S. Fraser's *Statistics: An Introduction* (1958) (p. ix). These are fine books: Mood, for example, gives a good treatment of power functions, pointing to their relevance in applied

work. But none of them make a distinction between substantive and statistical significance. Hoel writes that

> [t]here are several words and phrases used in connection with testing hypotheses that should be brought to the attention of students. When a test of a hypothesis produces a sample value falling in the critical region of the test, the result is said to be *significant*; otherwise one says that the result is *not significant*. (p. 176, his italics)

The student from the outset of her statistical education, therefore, is led to believe that economic (or substantive) significance and statistical significance are the same thing. Hoel explains: "This word ['not significant'] arises from the fact that such a sample value is not compatible with the hypothesis and therefore signifies that some other hypothesis is necessary" (p. 176). The elementary point that "[t]here is no sharp border between 'significant' and 'insignificant,' only increasingly strong evidence as the *P*-value decreases" (David S. Moore and George P. McCabe 1993, p. 473) is not found in most of the earlier books from which most economists learned statistics and econometrics. The old classic by W. Allen Wallis and Harry V. Roberts, *Statistics: A New Approach*, first published in 1956, is an exception:

> It is essential not to confuse the statistical usage of "significant" with the everyday usage. In everyday usage, "significant" means "of practical importance," or simply "important." In statistical usage, "significant" means "signifying a characteristic of the population from which the sample is drawn," regardless of whether the characteristic is important. (Wallis and Roberts [1956] 1965, p. 385)

The point has been revived in elementary statistics books, though most still do not emphasize it. In their leading elementary book the statisticians David Freedman, Robert Pisani, and Roger Purves (1978) could not be plainer. In one of numerous places where they make the point they write:

> This chapter . . . explains the limitations of significance tests. The first one is that "significance" is a technical word. A test can only deal with the question of whether a difference is real [permanent in Venn's sense], or just a chance variation. *It is not designed to see whether the difference is important.* (p. 487, italics supplied)

The distinction is also emphatic in Ronald J. Wonnacott and Thomas H. Wonnacott (1982, p. 160) and in Moore and McCabe (1993, p. 474).

III. *The Instrument: A Survey of Practice in Significance*

The evidence, then, is that econometricians are not in their textbooks emphasizing the difference between economic significance and statistical significance. What is practice?

We take the full-length papers published in the *American Economic Review* as an unbiased selection of best practice (we will not say "sample" and will not therefore use tests of statistical significance). We read all the 182 papers in the 1980s that used regression analysis (and record our impression that in most matters these are superb examples of economic science). Each paper was asked 19 questions about its use of statistical significance, to be answered "yes" (sound statistical practice) or "no" (unsound practice) or "not applicable."

The survey questions are:

1. *Does the paper use a small number of observations, such that statistically significant differences are not found at the conventional levels merely by choosing a large number of observations*? The power of a test is high if the significance level at $N = 30{,}000$ is carried over from situations in which the sample is 30 or 300. For example, in Glen C. Blomquist, Mark C. Berger, and John P. Hoehn, $N =$

34,414 housing units and 46,004 individuals (Mar. 1988, p. 93). At such large sample sizes the authors need to pay attention to the tradeoff between power and the size of the test, and to the economic significance of the power against alternatives.

2. *Are the units and descriptive statistics for all regression variables included?* Empirical work in economics is measurement. It is elementary to include units of the variables, and then also to give means.

3. *Are coefficients reported in elasticity form, or in some interpretable form* relevant for the problem at hand and consistent with economic theory, so that readers can discern the economic impact of regressors? Wallis and Roberts long ago complained that "sometimes authors are so intrigued by tests of significance that they fail even to state the actual *amount* of the effect, much less to appraise its practical importance" (1956, p. 409). In some fields (not much in economics, though we did find one example) the investigator will publish tables that consist only of asterisks indicating levels of significance.

4. *Are the proper null hypotheses specified?* The commonest problem would be to test against a null of zero when some other null is to the point. Such an error would be the result of allowing a canned program to make scientific decisions. If a null hypothesis is $\beta_1 + \beta_2 = 1$, there is not much to be gained from testing the hypothesis that each coefficient is statistically significantly different from zero. The most fruitful application of the Neyman-Pearson test specifies the null hypothesis as something the researcher believes to be true. The only result that leads to a definitive conclusion is a rejection of the null hypothesis. Failing to reject does not of course imply that the null is therefore true. And rejecting the null does not imply that the alternative hypothesis is true: there may be other alternatives (a range that investigators agree is relevant, for example) which would cause rejection of the null. The current rhetoric of rejection promotes a lexicographic procedure of "regress height income country age"; inspect t-values; discard as unimportant if $t < 2$; circulate as important if $t > 2$.

5. *Are coefficients carefully interpreted?* Goldberger has an illustration similar to many issues in economic policy (Goldberger 1991, p. 241). Suppose the dependent variable is "weight in pounds," the large coefficient is on "height," the smaller coefficient is on "exercise," and the estimated coefficients have the same standard errors. Neither the physician nor the patient would profit from an analysis that says height is "more important" (its coefficient being more standard errors away from zero in this sample), offering the overweight patient in effect the advice that he's not too fat, merely too short for his weight. "The moral of this example is that statistical measures of 'importance' are a diversion from the proper target of research—estimation of relevant parameters—to the task of 'explaining variation' in the dependent variable" (Goldberger, p. 241).

6. *Does the paper eschew reporting all t- or F-statistics or standard errors, regardless of whether a significance test is appropriate?* Statistical computing software routinely provide t-statistics for every estimated coefficient. But that programs provide it does not mean that the information is relevant for science. We suspect that referees enforce the proliferation of meaningless t- and F-statistics, out of the belief that statistical and substantive significance are the same.

7. *Is statistical significance at the first use, commonly the scientific crescendo of the paper, the only criterion of "impor-*

tance"? By "crescendo" we mean that place in the paper where the author comes to what she evidently considers the crucial test.

8. *Does the paper mention the power of the tests*? For example, Frederic S. Mishkin does, unusually, in two footnotes (June 1981, pp. 298 n11, 305 n27; lack of power is a persistent difficulty in capital-market studies, but is seldom faced). As DeGroot pointed out, the power of a test may be low against a nearby and substantively significant alternative. On the other hand, power may be high against a nearby and trivial alternative.

9. *If the paper mentions power, does it do anything about it*? It is true that power can only be discussed relative to an explicit alternative hypothesis, making power analysis difficult for some of the alternatives. An example is the Durbin-Wu-Hausman test for whether two estimators are consistent. (The survey accounts for the difficulty by coding the relevant papers "not applicable.")

10. *Does the paper eschew "asterisk econometrics,"* that is, ranking the coefficients according to the absolute size of *t*-statistics?

11. *Does the paper eschew "sign econometrics,"* that is, remarking on the sign but not the size of the coefficients? There is a little statistical theory in the econometrics books lying behind this customary practice (Goldberger, ch. 22; Greene, ch. 8), though for the most part the custom outstrips the theory. But sign is not *economically* significant unless the magnitude is large enough to matter. Statistical significance does not tell whether the size is large enough to matter. It is not true, as custom seems to be arguing, that sign is a statistic independent of magnitude.

12. *Does the paper discuss the size of the coefficients*? That is, once regression results are presented, does the paper make the point that some of the coefficients and their variables are *economically* influential, while others are not? Blomquist, Berger, and Hoehn do in part, by giving their coefficients on housing and neighborhood amenities in dollar form. But they do not discuss whether the magnitudes are scientifically reasonable, or in some other way important. Contrast Christina Romer, in a 19–page, exclusively empirical paper: "Indeed, correcting for inventory movements reduces the discrepancy . . . by approximately half. This suggests that inventory movements are [economically] important" (June 1986, p. 327). M. Boissiere, J. B. Knight, and R. H. Sabot reflect the more typical practice: "In both countries, cognitive achievement bears a highly significant relationship to educational level . . . In Kenya, secondary education raises *H* by 11.75 points, or by 35 percent of the mean" (Dec. 1985, p. 1026). They make ambiguous use of the word "significance," then draw back to the relevant question of economic significance. Later in the paragraph they recur to depending on statistical significance alone: "significantly positive" and "almost significantly positive" become again their only criteria of importance.

Daniel Hamermesh, by contrast, estimates his crucial parameter *K*, and at the first mention says, "The estimates of *K* are quite large, implying that the firm varies employment only in response to very large shocks. . . . Consider what an estimate this large means" (Sept. 1989, p. 683). The form is here close to ideal: it gets to the scientific question of what the size of a magnitude means. Two paragraphs down he speaks of "fairly large," "very important," "small," and "important" without merging these with statistical significance. In Goldberger's terms, he focuses on "the proper target of research—estimation of relevant parameters." (Later, though, Hamermesh

falls back to average practice: "The $\hat{K}$ for the aggregated data in Table 2 are insignificant," though he adds wisely, "and very small; and the average values of the $\hat{p}$ are much higher than in the pooled data"; *p. 685.*)

13. *Does the paper discuss the scientific conversation within which a coefficient would be judged "large" or "small"*? Romer, for example, remarks that "The existence of the stylized fact [that is, the scientific consensus] that the economy has stabilized implies a general consensus" (p. 322).

14. *Does the paper avoid choosing variables for inclusion solely on the basis of statistical significance*? The standard argument is that if certain variables enter the model significantly, the information should not be spurned. But such an argument merges statistical and substantive significance.

15. *After the crescendo, does the paper avoid using statistical significance as the criterion of importance*? The referees will have insisted unthinkingly on a significance test, the prudent author will have acceded to their insistence, but will after reporting them turn to other and scientifically relevant criteria of importance.

16. *Is statistical significance decisive, the conversation stopper, conveying the sense of an ending*? Romer and Jeffrey Sachs (Mar. 1980) both use statistical significance, and misuse it—in both cases looking to statistical significance as a criterion for how large is large. But in neither paper does statistical significance run the empirical work. The misuse in Michael Darby (June 1984) is balder: his only argument for a coefficient when he runs a regression is its statistical significance (pp. 311, 315), but on the other hand his findings do not turn on the regression results.

17. *Does the paper ever use a simulation (as against a use of the regression as an input into further argument) to determine whether the coefficients are reasonable*? To some degree Blomquist, Berger, and Hoehn do. They simulate the rankings of cities by amenity, and if the coefficients were quite wrong the rankings would be themselves unreasonable. Santa Barbara does rank high, though the differential value of amenities worst to best, at $5,146, seems low (Mar. 1988, p. 96). Simulations using regression coefficients can be informative, but of course should not use statistical significance as a screening device for input.

18. *In the "conclusions" and "implications" sections, is statistical significance kept separate from economic, policy, and scientific significance*? In Boissiere, Knight, and Sabot (Dec 1985) the effect of ability is isolated well, but the economic significance is not argued.

19. *Does the paper avoid using the word "significance" in ambiguous ways, meaning "statistically significant" in one sentence and "large enough to matter for policy or science" in another*? Thus Darby (June 1984): "First we wish to test whether oil prices, price controls, or both has a significant influence on productivity growth" (p. 310). The meanings are merged.

IV. *Results of the Survey of the American Economic Review*

Some of the *AER* authors, such as Romer and Hamermesh, show that they are aware of the substantive importance of the questions they ask, and of the futility of relying on a test of statistical significance for getting answers. Thus Kim B. Clark: "While the union coefficient in the sales specification is twice the size of its standard error, it is substantively small; moreover, with over 4,600 observations, the power of the evidence that the effect is different from zero is not

TABLE 1
THE *AMERICAN ECONOMIC REVIEW* IN THE 1980s HAD NUMEROUS ERRORS IN THE USE OF STATISTICAL SIGNIFICANCE

Survey Question	Total for which the question applies	Percent Yes
Does the Paper . . .		
8. Consider the power of the test?	182	4.4
6. Eschew reporting all standard errors, *t*-, and *F*-statistics, when such information is irrelevant?	181	8.3
17. Do a simulation to determine whether the coefficients are reasonable?	179	13.2
9. Examine the power function?	12	16.7
13. Discuss the scientific conversation within which a coefficient would be judged large or small?	181	28.0
16. Consider more than statistical significance decisive in an empirical argument?	182	29.7
18. In the conclusions, distinguish between statistical and substantive significance?	181	30.1
2. Report descriptive statistics for regression variables?	178	32.4
15. Use other criteria of importance besides statistical significance after the crescendo?	182	40.7
19. Avoid using the word "significance" in ambiguous ways?	180	41.2
5. Carefully interpret coefficients? For example, does it pay attention to the details of the units of measurement, and to the limitations of the data?	181	44.5
11. Eschew "sign econometrics," remarking on the sign but not the size of the coefficients?	181	46.7
7. At its first use, consider statistical significance to be one among other criteria of importance?	182	47.3
3. Report coefficients in elasticities, or in some other useful form that addresses the question of "how large is large"?	173	66.5
14. Avoid choosing variables for inclusion solely on the basis of statistical significance?	180	68.1
10. Eschew "asterisk econometrics," the ranking of coefficients according to the absolute size of the test statistic?	182	74.7
12. Discuss the size of the coefficients?	182	80.2
1. Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample?	182	85.7
4. Test the null hypotheses that the authors said were the ones of interest?	180	97.3

Source for Tables 1–5: All full-length papers using regression analysis in the *American Economic Review*, 1980–1989, excluding the Proceedings.
Notes: "Percent Yes" is the total number of Yes responses divided by the relevant number of papers (never exceeding 182). Some questions are not generally applicable to particular papers and some questions are not applicable because they are conditional on the paper having a particular characteristic. Question 3, for example, was coded "not applicable" for papers which exclusively use nonparametric statistics. Question 19 was coded "not applicable" for papers that do not use the word "significance."

overwhelming" (Dec 1984, p. 912). And Griliches:

> Here and subsequently, all statements about statistical "significance" should not be taken literally. Besides the usual issue of data mining clouding their interpretation, the "sample" analyzed comes close to covering completely the relevant population. Tests of significance are used here as a metric for discussing the relative fit of different versions of the model. In each case, the actual magnitude of the estimated coefficients is of more interest than their precise "statistical significance." (Dec 1986, p. 146)

Griliches understands that populations should not be treated as samples, and that statistical significance is not a substitute for economic significance. (He does not say why statistical significance is a scientifically relevant "metric for discussing the relative fit of the different versions of the model.")

But most authors in the *AER* do not understand these points. The results of applying the survey to the papers of the 1980s are displayed in Table 1.

The principal findings of the survey are:

- 70 percent of the empirical papers in the *American Economic Review* papers did not distinguish statistical significance from economic, policy, or scientific significance.
- At the first use of statistical significance, typically in the "Estimation" or "Results" section, 53 percent did not consider anything but the size of *t*- and *F*-statistics. About one third used only the size of *t*- and *F*-test statistics as a criterion for the inclusion of variables in future work.
- 72 percent did not ask "How large is large?" That is, after settling on an estimate of a coefficient, 72 percent did not consider what other authors had found; they did not ask what standards other authors have used to determine "importance"; they did not provide an argument one way or another whether the estimate $\beta = 0.999$ is economically close to 1.0 and economically important even though "statistically different from one." Awareness that scientific inquiry takes place in a conversation about how large is large seemed to improve the econometric practice. Of 131 papers that did *not* mention the work of other authors as a quantitative context for their own, 78 percent let statistical significance decide questions of substantive significance. Of 50 papers that did mention the work of other authors as a context, only 20 percent let statistical significance decide.
- 59 percent used the word "significance" in ambiguous ways, at one point meaning "statistically significantly different from the null," at another "practically important" or "greatly changing our scientific opinions," with no distinction.
- Despite the advice proffered in theoretical statistics, only 4-percent considered the power of their tests. One percent examined the power function.
- 69 percent did not report descriptive statistics—the means of the regression variables, for example—that would allow the reader to make a judgment about the economic significance of the results.
- 32 percent admitted openly to using statistical significance to drop variables (question 14). One would have to have more evidence than explicit admissions to know how prevalent the practice is in fact. One-third is a lower bound.
- Multiple-author papers, as one might expect from the theory of common property resources, more

TABLE 2
MULTIPLE AUTHORS APPEAR TO HAVE COORDINATION PROBLEMS, MAKING THE ABUSES WORSE
MEASURED BY PERCENT YES

Survey Question	Multiple Author Papers	Single Author Papers
Does the paper . . .		
7. At its first use, consider statistical significance to be one among other criteria of importance?	42.2	53.4
10. Eschew "asterisk econometrics," the ranking of coefficients according to the absolute size of the test statistic?	68.8	79.2
12. Discuss the size of the coefficients?	76.7	84.1
1. Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample?	77.8	84.8

Notes: "Percent Yes" is the total number of Yes responses divided by the relevant number of papers.

often spoke of "significance" in ambiguous ways, used sign econometrics, did not discuss the size of estimated coefficients, and found nothing more than the size of test statistics to be of importance at the first use of statistical significance (Table 2).

- Authors from "Tier 1" schools did in some respects a little better, but whether the difference justifies the invidious terminology of "tiers" is a scientific, not a statistical, question and must be left to the investigator (Table 3; the terminology is that of the most recent National Research Council assessment and includes Chicago, Harvard, MIT, Princeton, Stanford, and Yale.)

Though we do not here report the results, we found on the other hand that papers written by faculty at Tier 1 schools were proportionally more likely to use sampling theory on entire populations, and to treat as probability samples what are in fact samples of convenience.

The substantive significance of such practices can be made more vivid by examining a few of the papers in some depth.

The first is a case of not thinking about the economic meaning of a coefficient. The authors estimate benefit-cost ratios for the state of Illinois following the implementation of an unemployment insurance experiment. In one experiment a control group was given a cash bonus for getting a job quickly and keeping it for several months. In another experiment, the "Employer Experiment," *employers* were given a cash-bonus if claimants found a job quickly and retained it for some specified amount of time (Sept. 1987, p. 517). The intent of the "Employer Experiment" was to "provide a marginal wage-bill subsidy, or training subsidy, that might reduce the duration of insured unemployment" (p. 517). Here is how the conclusion is presented:

> The fifth panel also shows that the overall benefit-cost ratio for the Employer Experiment is 4.29, but it is not statistically different from zero. The benefit-cost ratio for white women in the Employer Experiment, however, is 7.07, and is statistically different from zero. Hence, a program modeled on the Employer Experiment also might be attrac-

TABLE 3
AUTHORS AT TIER 1 DEPARTMENTS DO BETTER THAN OTHERS IN MANY CATEGORIES MEASURED BY PERCENT YES

Survey Question	Tier 1 Departments	Other Departments
Does the paper . . .		
1. Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample?	91.3	83.9
12. Discuss the size of the coefficients?	87.0	78.9
10. Eschew "asterisk econometrics," the ranking of coefficients according to the absolute size of the test statistic?	84.8	71.4
7. At its first use, consider statistical significance to be one among other criteria of importance?	65.5	41.2
5. Carefully interpret coefficients? For example, does it pay attention to the details of the units of measurement, and to the limitations of the data?	60.0	37.5
19. Avoid using the word "significance" in ambiguous ways?	52.4	37.5
18 In the conclusions, distinguish between statistical and substantitive significance?	50.0	23.1

Notes: According to the most recent National Research Council assessment, the tier 1 departments are Chicago, Harvard, MIT, Princeton, Stanford, and Yale.
"Percent Yes" is the total number of Yes responses divided by the relevant number of papers.

> tive from the state's point of view if the program did not increase unemployment among nonparticipants. Since, however, the Employer Experiment affected only white women, it would be essential to understand the reasons for the uneven effects of the treatment on different groups of workers before drawing conclusions about the efficacy of such a program. (p. 527)

Here "affected" means that the estimated coefficient is statistically significantly different from a value the authors believe to be the relevant one. The 4.29 benefit-cost ratio for the whole Employer Experiment is, according to the authors, *not useful or important for public policy*. The 7.07 ratio for white women is said to "affect"—to be important—because it passed an arbitrary significance test. That is, 7.07 *affects*, 4.29 does not. It is true that 4.29 is a realization from a noisy random variable, whereas 7.07 is from a more quiet one. Though the authors do not say so, the 4.29 benefit-cost ratio is marginally discernible from zero at about the 12 percent level (p. 527). Yet for policy purposes even a noisy benefit-cost ratio is worth talking about. The argument that the 4.29 figure does not "affect" is unsound, and could be costly in employment foregone.

Another paper offers "an alternative test of the CAPM and report[s] . . . test results that are free from the ambiguity imbedded in the past tests" (Jan. 1980, p. 660). The authors are taking exception, they say, to Richard Roll's comment that "there is practically no possibility that such a test can be accomplished in the future" (p. 660). So they test five hypotheses: the intercept equals zero; the slope coefficients differ from zero; the adjusted coefficient of determination should be near one; there is no trend in

the intercept; and there is no trend in the adjusted coefficient of determination (pp. 664–65). On several time-series they run least squares regressions to estimate coefficients. Nowhere in the text is the size of the estimated coefficients discussed (a common mistake in the capital-market literature). Instead, the authors *rank* their results according to the number of times the absolute value of the t-statistic is greater than two (p. 667). Three out of four of their tables of estimation results have a column called "No. of Times $t > 2$," another column with "Average t-statistics," and one with "Adjusted R^2." They do not report coefficient estimates in the three tables, merely the t-statistics (Tables 1, 2, and 3, pp. 667–68). The only "Yes" that the paper earned in our survey was for specifying the null according to what their theory suggests.

Using ambiguously the very word "significance" implies there is no difference between economic significance and statistical significance, that nothing or little else matters. Of the 96 papers that use only the test of statistical significance as a criterion of importance at its first use, 90 percent imply—or state—that it is decisive in an empirical argument, and 70 percent use the word "significance" ambiguously. Of the other 86 papers in the survey less than half use the word ambiguously. The 96 unsound papers continue making inappropriate decisions at a higher rate than the 86 papers that acknowledge some criterion other than statistical significance. Only seven of the 96 distinguish statistical significance from economic or policy or scientific significance in the conclusions and implications sections, while 47 of the 86 make the distinction (Table 4).

Here is an extreme case of ambiguity:

> The statistically significant [read: (1) sampling theory] inequality aversion is in addition to any unequal distribution of inputs resulting from different social welfare weights for different neighborhoods. The KP results allowing for unequal concern yield an estimate of q of -3.4. This estimate is significantly [read: (2) some numbers are smaller than others] less than zero, indicating aggregate outcome is not maximized. At the same time, however, there is also significant [read: (3) a moral or scientific or policy matter] concern about productivity, as the inequality parameter is significantly [read: (4) a joint observation about morality and numbers] greater than the extreme of concern solely with equity. (AER Mar. 1987, p. 46)

In a piece on Ricardian Equivalence, statistical significance decides nearly everything:

> Notice the least significant of the variables in the constrained estimation is the second lagged value of the deficit in the government purchases equation. A natural course would be to reestimate the model for the case of two lagged values of government spending and one lagged value of the government deficit. . . . Although the elimination of [the variable] raises the confidence level at which the null hypothesis can be rejected, it remains impossible to argue that the data provides evidence against the joint proposition of Ricardian equivalence and rational expectations at conventional levels of significance. (AER Mar. 1985, p. 125)

Another paper reports "significant" results on the relation between unemployment and money:

> The coefficient is significant at the 99 percent confidence level. Neither the current money shock nor all 12 coefficients as a group are significantly different from zero. The coefficient on c is negative and significant and the distributed lag on c is significant as well. In column (2) we report a regression which omits the insignificant lags on money shocks. The c distributed lag is now significant at the 1 percent confidence level. . . .
>
> We interpret these results as indicating that the primary factor determining cyclical variations in the probability of leaving unemployment is probably heterogeneity. Inventory innovations appear to play some role and surprisingly, money shocks have no significant impact. (AER Sept. 1985, p. 630)

TABLE 4
IF ONLY STATISTICAL SIGNIFICANCE IS SAID TO BE OF IMPORTANCE AT ITS FIRST USE (QUESTION 7), THEN MANY OTHER INAPPROPRIATE DECISIONS ARE MADE
MEASURED BY PERCENT YES

Survey Question	If only statistical significance is important	If more than statistical significance is important
Does the paper . . .		
12. Examine the power function?	0	28.6
6. Eschew reporting all standard errors, *t*-, and *F*-statistics, when such information is irrelevant	3.2	14.0
8. Consider the power of the test?	4.2	4.7
17. Do a simulation to determine whether the coefficients are reasonable	6.3	17.9
18. In the conclusions, distinguish between statistical and substantive significance	7.3	55.3
16. Consider more than statistical significance decisive in an empirical argument?	10.4	51.2
5. Carefully interpret coefficients? For example, does it pay attention to the units of measurement, and to the limitations of the data?	13.7	77.9
13. Discuss the scientific conversation within which a coefficient would be judged large or small?	17.7	38.8
11. Eschew "sign econometrics," remarking on the sign but not the size of the coefficients?	21.9	74.1
2. Report descriptive statistics for regression variables?	26.3	36.1
15. Use other criteria of importance besides statistical significance after the crescendo?	30.2	52.3
19. Avoid using the word "significance" in ambiguous ways?	29.5	52.9
3. Report coefficients in elasticities, or in some other useful form that addresses the question "how large is large?"	51.6	80.0
14. Avoid choosing variables for inclusion solely on the basis of statistical significance?	59.0	77.7
10. Eschew "asterisk econometrics," the ranking of coefficients according to the size of the test statistic?	66.7	83.7
12. Discuss the size of the coefficients, making points of substantive significance?	66.7	96.5
1. Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample?	86.5	84.8
4. Test the null hypotheses that the authors say are the ones of interest?	94.7	100

Notes: "Percent Yes" is the total number of Yes responses divided by the relevant number of papers. Some questions are not generally applicable because they are conditional on a paper having a particular characteristic. Question 3, for example, was coded "not applicable" for papers which exclusively use nonparametric statistics. Question 19 was coded "not applicable" for papers that do not use the word "significance."

TABLE 5
THE EASE OF COMPUTING STATISTICAL SIGNIFICANCE IN THE LATE 1970s MAY HAVE HAD ILL EFFECTS ON THE USE OF REGRESSION ANALYSIS
MEASURED BY PERCENT YES

	Does the Paper . . .		
Date of Ph.D. Conferral	Distinguish Among Kinds of Significance in the Conclusions (Question 18)	Eschew Ambiguous Usage of the Very Word (Question 19)	Consider More Than Statistical Significance Decisive in Empirical Argument (Question 16)
1940–1969	29	61	26
1970–1974	33	37	31
1975–1979	17	29	13
1980–1984	33	45	33

Notes: The number of papers published by each cohort is 31, 48, 24, and 24. Multiple author papers were dated by the first name listed on the published article.

Such misuses of statistical significance appear to depend in part on a vintage effect, measured by date of Ph.D. conferral. The papers authored by Ph.D.'s conferred between 1975 and 1979, when inexpensively generated *t*-tests first reached the masses, were considerably worse than the papers of others at making a distinction between economic and statistical significance. They used the word "significance" in ambiguous ways more often than did early or later Ph.D.'s and they were less likely to separate statistical significance from other kinds of significance in the sections on scientific and policy implications (Table 5).

V. *Taking the Con Out of Confidence Intervals*

In a squib published in the *American Economic Review* in 1985 one of us claimed that "[r]oughly three-quarters of the contributors to the *American Economic Review* misuse the test of statistical significance" (McCloskey 1985, p. 201). The full survey confirms the claim, and in some matters strengthens it.

We would not assert that every economist misunderstands statistical significance, only that most do, and these some of the best economic scientists. By way of contrast to what most understand statistical significance to be capable of saying, Edward Lazear and Robert Michael wrote 17 pages of empirical economics in the *AER*, using ordinary least squares on two occasions, without a single mention of statistical significance (*AER* Mar. 1980, pp. 96–97, pp. 105–06). This is notable considering they had a legitimate sample, justifying a discussion of statistical significance were it relevant to the scientific questions they were asking. Estimated coefficients in the paper are interpreted carefully, and within a conversation in which they ask how large is large (pp. 97, 101, and throughout).

The low and falling cost of calculation, together with a widespread though unarticulated realization that after all the significance test is not crucial to scientific questions, has meant that statistical significance has been valued at its cost. Essentially no one believes a finding of statistical significance or insignificance.

This is bad for the temper of the field. My statistical significance is a "finding"; yours is an ornamented prejudice. Con-

trary to the decisive rhetoric of rejection in the mechanical test, statistical significance has not in fact changed the minds of economic scientists. In a way the insignificance of significance tests in scientific debate is comforting. Economists have not been fooled, even by their own mistaken beliefs about statistical significance. To put it another way, no economist has achieved scientific success as a result of a statistically significant coefficient. Massed observations, clever common sense, elegant theorems, new policies, sagacious economic reasoning, historical perspective, relevant accounting: these all have led to scientific success. Statistical significance has not.

What should replace a lessened attention to statistical significance is serious attention to the scientific question. The scientific question is ordinarily "How large is large in the present case?" This is the question that geologists thinking about continental drift and astrophysicists thinking about stellar evolution spend their days answering.

The question "How large is large?" requires thinking about what coefficients would be judged large or small in terms of the present conversation of the science. It requires thinking more rigorously about data—for example, asking what universe they are a "sample" from. (Carelessness in such matters is more common than one might have expected. Of the 107 papers using cross-sectional data, for example, 20 percent used tests of statistical significance on the entire population or on a sample of convenience. Only two of these offered some justification for the usage.)

Most scientists (and historians) use simulation, which in explicit, quantitative form is becoming cheaper in economics, too. It will probably become the main empirical technique, following other observational sciences. Econometrics will survive, but it will come at last to emphasize economic rather than statistical significance. We should of course worry some about the precision of the estimates, but as Leamer has pointed out the imprecision usually comes from sources other than too small a sample.

Simulation, new data sets, and quantitative thinking about the conversation of the science offer a way forward. The first step anyway is plain: stop searching for economic findings under the lamppost of statistical significance.

References

AMEMIYA, TAKESHI. *Advanced econometrics*. Cambridge: Harvard U. Press, 1985.

AMES, EDWARD AND REITER, STANLEY. "Distributions of Correlation Coefficients in Economic Time Series," *J. Amer. Statist. Assoc.*, Sept. 1961, *56*(295), pp. 637–56.

BAKAN, DAVID. "The Test of Significance in Psychological Research," *Psychological Bulletin*, Dec. 1966, *66*(6), pp. 423–37.

BARRETT, WILLIAM. *The illusion of technique*. Garden City, NY: Anchor Press, 1978.

BEHRMAN, JERE R. AND CRAIG, STEVEN G. "The Distribution of Public Services: An Exploration of Local Government Preferences," *Amer. Econ. Rev.*, Mar. 1987, *77*(1), pp. 37–49.

BLOMQUIST, GLENN C.; BERGER, MARK C. AND HOEHN, JOHN P. "New Estimates of Quality of Life in Urban Areas," *Amer. Econ. Rev.*, Mar. 1988, *78*(1), pp. 89–107.

BOISSIERE, M.; KNIGHT, J.B. AND SABOT, R.H. "Earnings, Schooling, Ability, and Cognitive Skills," *Amer. Econ. Rev.*, Dec. 1985, *75*(5), pp. 1016–30.

BORING, EDWIN G. "Mathematical versus Scientific Significance," *Psychological Bulletin*, Oct. 1919, *16*(10), pp. 335–38.

CICERO, MARCUS TULLIUS. *De divinatione* [45 BC]; in *De senectute; De amicitia; De divinatione*. Ed. and trans. WILLIAM A. FALCONER. Cambridge: Harvard U. Press, 1938.

CLARK, KIM B. "Unionization and Firm Performance: The Impact on Profits, Growth, and Productivity," *Amer. Econ. Rev.*, Dec. 1984, *74*(5), pp. 893–919.

COHEN, JACOB. "The Statistical Power of Abnormal-Social Psychological Research: A Review," *J. Abnormal and Social Psychology*, Sept. 1962, *65*(3), pp. 145–53.

COOLEY, THOMAS F. AND LEROY, STEPHEN F. "Identification and Estimation of Money Demand," *Amer. Econ. Rev.*, Dec. 1981, *71*(5), pp. 825–44.

DARBY, MICHAEL R. "The U.S. Productivity Slowdown: A Case of Statistical Myopia," *Amer. Econ. Rev.*, June 1984, *74*(3), pp. 301–22.

DAVIDSON, RUSSELL AND MACKINNON, JAMES G. *Estimation and inference in econometrics*. Oxford: Oxford U. Press, 1993.

DAVIS, PHILIP J. AND HERSH, REUBEN. "Rhetoric and Mathematics," in *The rhetoric of the human sciences*. Eds.: JOHN S. NELSON, ALLAN MEGILL, AND DONALD N. MCCLOSKEY. Madison: U. of Wisconsin Press, 1987, pp. 53–68.

DEGROOT, MORRIS H. *Probability and statistics*. Reading, MA: Addison-Wesley, [1975] 1989.

DENTON, FRANK T. "Data Mining as an Industry," *Rev. Econ. Statist.*, Feb. 1985, *67*(1), pp. 124–27.

———. "The Significance of Significance: Rhetorical Aspects of Statistical Hypothesis Testing in Economics," in *The consequences of economic rhetoric*. Eds.: ARJO KLAMER, DONALD N. MCCLOSKEY, AND ROBERT SOLOW. New York: Cambridge U. Press, 1988, pp. 163–83.

FEIGE, EDGAR. "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *J. Polit. Econ.*, Dec. 1975, *83*(6), pp. 1291–95.

FISHER, RONALD A. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 1925.

FLORENS, JEAN-PIERRE AND MOUCHART, MICHEL. "Bayesian Testing and Testing Bayesians," in *Handbook of statistics*. Vol. 11. *Econometrics*. of Eds.: G. S. MADDALA, C. R. RAO, AND H. D. VINOD. Amsterdam: North-Holland, 1993, pp. 303–91.

FRASER, DONALD A. S. *Statistics: An introduction*. New York: Wiley, 1958.

FREEDMAN, DAVID; PISANI, ROBERT AND PURVES, ROGER. *Statistics*. New York: Norton, 1978.

GIGERENZER, GERD. "Probabilistic Thinking and the Fight Against Subjectivity." Unpublished paper, Department of Psychology, Universitität Konstanz, no date.

GOLDBERGER, ARTHUR S. *A course in econometrics*. Cambridge: Harvard U. Press, 1991.

GORDON, SCOTT. *The history and philosophy of social science*. London: Rontledge, 1991.

GOULD, STEPHEN JAY. *The mismeasure of man*. New York: Norton, 1981.

GRANGER, CLIVE W. J. "A Review of Some Recent Textbooks of Econometrics," *J. Econ. Lit.*, Mar. 1994, *32*(1), pp. 115–22.

GREENE, WILLIAM H. *Econometric analysis*. New York: Macmillan, [1990] 1993.

GRIFFITHS, WILLIAM E.; HILL, R. CARTER AND JUDGE, GEORGE G. *Learning and practicing econometrics*. New York: Wiley, 1993.

GRILICHES, ZVI. "Productivity, R&D, and Basic Research at the Firm Level in the 1970's," *Amer. Econ. Rev.*, Mar. 1986, *76*(1), pp. 141–54.

GRILICHES, ZVI AND INTRILIGATOR, MICHAEL D. *Handbook of econometrics*. Vols. I, II, and III. Amsterdam: North-Holland, 1983, 1984, 1986.

GUTTMAN, LOUIS. "What Is Not What in Statistics?" in *Multidimensional data representations: When and why*. Ed.: INGWER BORG. Ann Arbor: Methesis Press, 1981, pp. 20–46.

———. "The Illogic of Statistical Inference for Cumulative Science," *Applied Stochastic Models and Data Analysis*, July 1985, *1*(1), pp. 3–10.

HAMERMESH, DANIEL S. "Labor Demand and the Structure of Adjustment Costs," *Amer. Econ. Rev.*, Sept. 1989, *79*(4), pp. 674–89.

HOEL, PAUL G. *Elementary statistics*. New York: Wiley, 1966.

HOGBEN, LANCELOT T. *Statistical theory: The relationship of probability, credibility, and error*. New York: Norton, 1968.

HOGG, ROBERT V. AND CRAIG, ALLEN T. *Introduction to mathematical statistics*. 4th ed. New York: Macmillan, 1978.

JOHNSTON, JOHN. *Econometric methods*. 2nd ed.. New York: McGraw-Hill, [1963] 1972.

———. *Econometric methods*. 3rd ed. New York: McGraw-Hill, 1984.

KENDALL, MAURICE G. AND STUART, ALAN.*The advanced theory of statistics*. Vol. 2., 3rd ed. London: Griffin, 1951.

KENDALL, MAURICE G.; STUART, ALAN AND ORD, J. KEITH. *The advanced theory of statistics*. Vol. 3, 4th ed. *Design and analysis, and time-series*. New York: Macmillan, 1983.

KENNEDY, PETER. *A guide to econometrics*. Cambridge: MIT Press, [1979] 1985.

KMENTA, JAN. *Elements of econometrics*. New York: Macmillan, 1971.

KRUSKAL, WILLIAM. "Significance, Tests of," *International encyclopedia of statistics*. Eds.: WILLIAM H. KRUSKAL AND JUDITH M. TANUR. New York: Macmillan, [1968] 1978a, pp. 944–58.

———. "Formulas, Numbers, Words: Statistics in Prose," *The American Scholar*, Spring 1978b, *47*(2), pp. 223–29.

KURTZ, ALBERT K. AND EDGERTON, HAROLD A., eds. *Statistical dictionary of terms and symbols*. New York: Wiley, 1939.

LAZEAR, EDWARD P. AND MICHAEL, ROBERT T. "Family Size and the Distribution of Real Per Capita Income," *Amer. Econ. Rev.*, Mar. 1980, *70*(1), pp. 91–107.

LEAMER, EDWARD E. *Specification searches: Ad hoc inferences with nonexperimental data*. New York: Wiley, 1978.

———. "Let's Take the Con Out of Econometrics," *Amer. Econ. Rev.*, Mar. 1983, *73*(1), pp. 31–43.

LOVELL, MICHAEL C. "Data Mining," *Rev. Econ. Statist.*, Feb. 1983, *65*(1), pp. 1–12.

MADDALA, G. S. *Introduction to econometrics*. New York: Macmillan, [1988] 1992.

MAYER, THOMAS. "Selecting Economic Hypotheses by Goodness of Fit," *Econ. J.*, Dec. 1975, *85*(340), pp. 877–83.

MCCLOSKEY, DONALD N. "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests," *Amer. Econ. Rev.*, May 1985, *75* (2), pp. 201–05.

McCLOSKEY, DONALD N. AND ZECHER, J. RICHARD. "The Success of Purchasing Power Parity," in *A retrospective on the classical gold standard, 1821–1931*. Eds.: MICHAEL D. BORDO AND ANNA J. SCHWARZ. Chicago and London: U. of Chicago Press, 1984, pp. 121–50.

MEEHL, PAUL E. "Theory Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*, June 1967, *34*(2), pp. 103–15.

MISHKIN, FREDERIC S. "Are Market Forecasts Rational?" *Amer. Econ. Rev.*, June 1971, *71*(3), pp. 295–305.

MOOD, ALEXANDER M. *Introduction to the theory of statistics*. 1st ed. New York: McGraw-Hill, 1950.

MOOD, ALEXANDER M. AND GRAYBILL, FRANKLIN A. *Introduction to the theory of statistics*. 2nd ed. New York: McGraw-Hill, 1963.

MOORE, DAVID S. AND McCABE, GEORGE P. *Introduction to the practice of statistics*. 2nd ed. New York: Freeman, 1993.

MORRISON, DENTON E. AND HENKEL, RAMON E. "Significance Tests Reconsidered," *Amer. Sociologist*, May 1969, *4*(2), pp. 131–39.

———. *The significance test controversy: A reader*. Chicago: Aldine, 1970.

MOSTELLER, FREDERICK AND TUKEY, JOHN W. *Data analysis and regression*. Reading, MA: Addison-Wesley, 1977.

NEYMAN, JERZY AND PEARSON, EGON S. "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society* A, 1933, *231*, pp. 289–337.

OHTA, MAKOTA AND GRILICHES, ZVI. "Automobile Prices Revisited: Extensions of the Hedonic Hypothesis," in *Household production and consumption*. Studies in Income and Wealth, vol. 40. Ed.: NESTOR E. TERLECKYJ. New York: National Bureau of Economics Research, 1976, pp. 325–90.

ROMER, CHRISTINA D. "Is the Stabilization of the Postwar Economy a Figment of the Data?" *Amer. Econ. Rev.*, June 1986, *76*(3), pp. 314–34.

RUDNER, RICHARD. "The Scientist Qua Scientist Makes Value Judgments," *Phil. Science*, Jan. 1953, *20*(1), pp 1–6.

SACHS, JEFFREY D. "The Changing Cyclical Behavior of Wages and Prices: 1880–1976," *Amer. Econ. Rev.*, Mar. 1980, *70* (1), pp. 78–90.

SCOTT, ELIZABETH. "Testing Hypotheses," in *Statistical astronomy*. Ed.: ROBERT J. TRUMPLER AND HAROLD F. WEAVER. New York: Dover, 1953, pp. 220–30.

TUKEY, JOHN W. "Sunset Salvo," *The American Statistician*, Feb. 1986, *40*(1), pp. 72–76.

TULLOCK, GORDON. "Publication Decisions and Tests of Significance—A Comment," *J. Amer. Statist. Assoc.*, Sept. 1959, *54*(287), p. 593.

———. *The organization of inquiry*. Durham, NC: Duke U. Press, 1966.

WALD, ABRAHAM. "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *Annals of Mathematical Statistics*, Dec. 1939, *10*(4), pp. 299–326.

WALLIS, W. ALLEN AND ROBERTS, HARRY V. *Statistics: A new approach*. New York: Macmillan, 1956.

WONNACOTT, RONALD J. AND WONNACOTT, THOMAS H. *Statistics: Discovering its power*. New York: Wiley, 1982.

WOOFTER, T. J., JR. "Common Errors in Sampling," *Social Forces*, May 1933, *11*(4), pp. 521–25.