The Cult of Statistical Significance

This series provides a forum for theoretical and empirical investigations of social phenomena. It promotes works that focus on the interactions among cognitive processes, individual behavior, and social outcomes. It is especially open to interdisciplinary books that are genuinely integrative.

Titles in the Series

Stephen T. Ziliak and Deirdre N. McCloskey. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*
Eirik G. Furubotn and Rudolf Richter. *Institutions and Economic Theory: The Contribution of the New Institutional Economics,* Second Edition
Tyler Cowen. *Markets and Cultural Voices: Liberty vs. Power in the Lives of Mexican Amate Painters*
Thráinn Eggertsson. *Imperfect Institutions: Possibilities and Limits of Reform*
Vernon W. Ruttan. *Social Science Knowledge and Economic Development: An Institutional Design Perspective*
Phillip J. Nelson and Kenneth V. Greene. *Signaling Goodness: Social Rules and Public Choice*
Stephen Knack, Editor. *Democracy, Governance, and Growth*
Omar Azfar and Charles A. Cadwell, Editors. *Market-Augmenting Government: The Institutional Foundations for Prosperity*
Randall G. Holcombe. *From Liberty to Democracy: The Transformation of American Government*
David T. Beito, Peter Gordon, and Alexander Tabarrok, Editors. *The Voluntary City: Choice, Community, and Civil Society*
Alexander J. Field. *Altruistically Inclined? The Behavioral Sciences, Evolutionary Theory, and the Origins of Reciprocity*
David George. *Preference Pollution: How Markets Create the Desires We Dislike*
Julian L. Simon. *The Great Breakthrough and Its Cause*
E. L. Jones. *Growth Recurring: Economic Change in World History*
Rosemary L. Hopcroft. *Regions, Institutions, and Agrarian Change in European History*
Lee J. Alston, Gary D. Libecap, and Bernardo Mueller. *Titles, Conflict, and Land Use: The Development of Property Rights and Land Reform on the Brazilian Amazon Frontier*
Daniel B. Klein, Editor. *Reputation: Studies in the Voluntary Elicitation of Good Conduct*
Richard A. Easterlin. *Growth Triumphant: The Twenty-first Century in Historical Perspective*

(continues on last page)

THE CULT OF STATISTICAL SIGNIFICANCE

How the Standard Error Costs Us Jobs, Justice, and Lives

By Stephen T. Ziliak
and
Deirdre N. McCloskey

The University of Michigan Press ~ Ann Arbor

Published in the United States of America by
The University of Michigan Press
Manufactured in the United States of America
♾ Printed on acid-free paper

2011 2010 2009 2008 4 3 2 1

A CIP catalog record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Ziliak, Stephen Thomas, 1963–
The cult of statistical significance : how the standard error costs us jobs, justice, and lives / by Stephen T. Ziliak and Deirdre N. McCloskey.
p. cm. — (Economics, cognition, and society series)
Includes index.
ISBN-13: 978-0-472-07007-7 (cloth : alk. paper)
ISBN-10: 0-472-07007-X (cloth : alk. paper)
ISBN-13: 978-0-472-05007-9 (pbk. : alk. paper)
ISBN-10: 0-472-05007-9 (pbk. : alk. paper)
1. Economics—Statistical methods. 2. Statistics—Social aspects. 3. Statistical hypothesis testing—Social aspects. I. McCloskey, Deirdre N. II. Title.

HB137.Z55 2007
330.01'5195—dc22 2007035401

To Lawrence *and* Barbara Ziliak,
the older generation

∴

To Connor *and* Lily McCloskey,
the younger

∴

And to the memory of
William H. Kruskal
(1919–2005)

The History of Science has suffered greatly from the use by teachers of second-hand material, and the consequent obliteration of the circumstances and the intellectual atmosphere in which the great discoveries of the past were made. A first-hand study is always instructive, and often . . . full of surprises.

Ronald A. Fisher, 1955

Contents

Preface xv
Acknowledgments xix

A Significant Problem 1

In many of the life and human sciences the existence/whether question of the philosophical disciplines has substituted for the size-matters/how-much question of the scientific disciplines. The substitution is causing a loss of jobs, justice, profits, environmental quality, and even life. The substitution we are worrying about here is called "statistical significance"—a qualitative, philosophical rule that has substituted for a quantitative, scientific magnitude and judgment.

1. Dieting "Significance" and the Case of Vioxx 23

Since R. A. Fisher (1890–1962) the sciences that have put statistical significance at their centers have misused it. They have lost interest in estimating and testing for the actual effects of drugs or fertilizers or economic policies. The big problem began when Fisher ignored the size-matters/how-much question central to a statistical test invented by William Sealy Gosset (1876–1937), so-called Student's *t*. Fisher substituted for it a qualitative question concerning the "existence" of an effect, by which he meant "low sampling error by an arbitrary standard of variance." Forgetting after Fisher what is known in statistics as a "minimax strategy," or other "loss function," many sciences have fallen into a sizeless stare. They seek sampling precision only. And they end by asserting that sampling precision just *is* oomph, magnitude, practical significance. The minke and sperm whales of Antarctica and the users and makers of Vioxx are some of the recent victims of this bizarre ritual.

2. The Sizeless Stare of Statistical Significance 33

Crossing frantically a busy street to save your child from certain death is a good gamble. Crossing frantically to get another mustard packet for your hot dog is not. The size of the potential loss if you don't hurry to save your child is larger, most will agree, than the potential loss if you don't get the mustard. But a majority of scientists in economics, medicine, and other statistical fields appear not to grasp the difference. If they have been trained in exclusively Fisherian methods (and nearly all of them have) they look only for a probability of success in the crossing—the existence of a probability of success better than .99 or .95 or .90, and this within the restricted frame of sampling—ignoring in any spiritual or financial currency the value of the prize and the expected cost of pursuing it. In the life and human sciences a majority of scientists look at the world with what we have dubbed "the sizeless stare of statistical significance."

3. What the Sizeless Scientists Say in Defense 42

The sizeless scientists act as if they believe the *size* of an effect does not matter. In their hearts they do care about size, magnitude, oomph. But strangely they don't measure it. They substitute "significance" measured in Fisher's way. Then they take the substitution a step further by limiting their concern for error to errors in sampling only. And then they take it a step further still, reducing all errors in sampling to one kind of error—that of excessive skepticism, "Type I error." Their main line of defense for this surprising and unscientific procedure is that, after all, "*statistical* significance," which they have calculated, is "objective." But so too are the digits in the New York City telephone directory, objective, and the spins of a roulette wheel. These are no more relevant to the task of finding out the sizes and properties of viruses or star clusters or investment rates of return than is statistical significance. In short, statistical scientists after Fisher neither test nor estimate, really, truly. They "testimate."

4. Better Practice: β-Importance vs. α-"Significance" 57

The most popular test was invented, we've noted, by Gosset, better known by his pen name "Student," a chemist and brewer at Guinness in Dublin. Gosset didn't think his test was very important to his main goal, which was of course brewing a good beer at a good price. The test, Gosset warned right from the beginning, does *not* deal with substantive importance. It does not begin to measure what Gosset called "real error" and "pecuniary advantage," two terms worth reviving in current statistical practice. But Karl Pearson and especially the amazing Ronald Fisher didn't listen. In two great books written and revised during the 1920s and 1930s, Fisher imposed a Rule of Two: if a result departs from an assumed hypothesis by two or more standard deviations of its own sampling variation, regardless of the size of the prize and the expected cost of going for it, then it is to be called a "significant" scientific finding. If not, not. Fisher told the subjectivity-phobic scientists that if they wanted to raise their studies "to the rank of sciences" they must employ his rule. He later urged them to ignore the size-matters/how-much approaches of Gosset, Neyman, Egon Pearson, Wald, Jeffreys, Deming, Shewhart, and Savage. Most statistical scientists listened to Fisher.

5. A Lot Can Go Wrong in the Use of Significance Tests in Economics 62

We ourselves in our home field of economics were long enchanted by Fisherian significance and the Rule of Two. But at length we came to wonder why the correlation of prices at home with prices abroad must be "within two standard deviations of 1.0 in the sample" before one could speak about the integration of world markets. And we came to think it strange that the U.S. Department of Labor refused to discuss black teenage unemployment rates of 30 or 40 percent because they were, by Fisher's circumscribed definition, "insignificant." After being told repeatedly, if implausibly, that such mistakes in the use of Gosset's test were *not* common in economics, we developed in the 1990s a questionnaire to test in economics articles for economic as against statistical significance. We applied it to the behavior of our tribe during the 1980s.

6. A Lot Did Go Wrong in the *American Economic Review* during the 1980s 74

We did not study the scientific writings of amateurs. On the contrary, we studied the *American Economic Review* (known to its friends as the *AER*), a leading journal of economics. With questionnaire in hand we read every full-length article it published that used a test of statistical significance from January 1980 to December 1989. As we expected, in the 1980s more than 70 percent of the articles made the significant mistake of R. A. Fisher.

7. Is Economic Practice Improving? 79

We published our article in 1996. Some of our colleagues replied, "In the old days [of the 1980s] people made that mistake, but [in the 1990s] we modern sophisticates do not." So in 2004 we published a follow-up study, reading all the articles published in the *AER* in the next decade, the 1990s. Sadly, our colleagues were again mistaken. Since the 1980s the practice in important respects got worse, not better. About 80 percent of the articles made the mistaken Fisherian substitution, failing to examine the magnitudes of their results. And less than 10 percent showed full concern for oomph. In a leading journal of economics, in other words, nine out of ten articles in the 1990s acted as if size doesn't matter for deciding whether a number is big or small, whether an effect is big or small enough to matter. The significance asterisk, the flickering star of *, has become a totem of economic belief.

8. How Big Is Big in Economics? 89

Does globalization hurt the poor, does the minimum wage increase unemployment, does world money cause inflation, does public welfare undermine self-reliance? Such scientific questions are always matters of economic significance. *How much* hurt, increase, cause, undermining? Size matters. Oomph is what we seek. But that is not what is found by the statistical methods of modern economics.

9. What the Sizeless Stare Costs, Economically Speaking 98

Sizeless economic research has produced mistaken findings about purchasing power parity, unemployment programs, monetary policy, rational addiction, and the minimum wage. In truth, it has vitiated most econometric findings since the 1920s and virtually all of them since the significance error was institutionalized in the 1940s. The conclusions of Fisherian studies might occasionally be correct. But only by accident.

10. How Economics Stays That Way: The Textbooks and the Referees 106

New assistant professors are not to blame. Look rather at the report card of their teachers and editors and referees—notwithstanding cries of anguish from the wise Savages, Zellners, Grangers, and Leamers of the economics profession. Economists received a quiet warning by F. Y. Edgeworth in 1885—too quiet, it seems—that sampling precision is not the same as oomph. They ignored it and have ignored other warnings, too.

11. The Not-Boring Rise of Significance in Psychology 123

Did other fields, such as psychology, do the same? Yes. In 1919 Edwin Boring warned his fellow psychologists about confusing so-called statistical with actual significance. Boring was a famous experimentalist at Harvard. But during his lectures on scientific inference his colleagues appear to have dozed off. Fisher's 5 percent philosophy was eventually codified by the *Publication Manual of the American Psychological Association,* which dictated the erroneous method worldwide to thousands of academic journals in psychology, education, and related sciences, including forensics.

12. Psychometrics Lacks Power 131

"Power" is a neglected statistical offset to the "first kind of error" of null-hypothesis significance testing. Power assigns a likelihood to the "second kind of error," that of undue gullibility. The leading journals of psychometrics have had their power examined by insiders to the field. The power of most psychological science in the age of Fisher turns out to have been

embarrassingly low or, in more than a few cases, spuriously "high"—as was found in a seventy-thousand-observation examination of the matter. Like economists the psychologists developed a fetish for testimation and wandered away from powerful measures of oomph.

13. The Psychology of Psychological Significance Testing 140

Psychologists and economists have said for decades that people are "Bayesian learners" or "Neyman-Pearson signal detectors." We learn by doing and staying alert to the signals. But when psychologists and others propose to test those very hypotheses they use Fisher's Rule of Two. That is, they erase their own learning and power to detect the signal. They seek a foundation in a Popperian falsificationism long known to be philosophically dubious. What in logic is called the "fallacy of the transposed conditional" has grossly misled psychology and other sizeless sciences. An example is the overdiagnosis of schizophrenia.

14. Medicine Seeks a Magic Pill 154

We found that medicine and epidemiology, too, are doing damage with Student's *t*—more in human terms perhaps than are economics and psychology. The scale along which one would measure oomph is very clear in medicine: life or death. Cardiovascular epidemiology, to take one example, combines with gusto the fallacy of the transposed conditional and the sizeless stare of statistical significance. Your mother, with her weak heart, needs to know the oomph of a treatment. Medical testimators aren't saying.

15. Rothman's Revolt 165

Some medical editors have battled against the 5 percent philosophy. But even the *New England Journal of Medicine* could not lead medical research back to William Sealy Gosset and the promised land of real science. Neither could the International Committee of Medical Journal Editors, though covering worldwide hundreds of journals. Kenneth Rothman, the founder of *Epidemiology,* forced change in his journal. But only his journal. Decades ago a sensible few in education, ecology, and sociology initiated a "significance test controversy." But grantors, journal referees, and tenure committees in the statistical sciences had faith that probability spaces can judge—the "judgment" merely that $p < .05$ is "better" for variable *X* than $p < .11$ for variable *Y*. It's not. It depends on the oomph of *X* and *Y*.

16. On Drugs, Disability, and Death 176

The upshot is that because of Fisher's standard error you are being given dangerous medicines, and are being denied the best medicines. The Centers for Disease Control is infected with *p*-values in a grant, for example, to study drug use in Atlanta. Public health has been infected, too. An outbreak of salmonella in South Carolina was studied using significance tests. In consequence a good deal of the outbreak was ignored. In 1995 a Cancer Trialists' Collaborative Group came to a rare consensus on effect size: ten different studies agreed that a certain drug for treating prostate cancer can increase patient survival by 12 percent. An eleventh study published in the *New England Journal of Medicine* dismissed the drug. The dismissal was based not on effect size bounded by confidence intervals based on what Gosset called "real" error but on a single *p*-value only, indicating, the Fisherian authors believed, "no clinically meaningful improvement" in survival.

17. Edgeworth's Significance 187

The history of this persistent but mistaken practice is a social study of science. In 1885 an eccentric and brilliant Oxford don, Francis Ysidro Edgeworth, coined the very term *significance.* Edgeworth was prolific in science and philosophy, but was especially interested in

watching bees and wasps. In measuring their behavioral differences, though, he focused on the sizes and meanings of the differences. He never depended on *statistical* significance.

18. "Take 3σ as Definitely Significant": Pearson's Rule 193

By contrast, Edgeworth's younger colleague in London, the great and powerful Karl Pearson, used "significance" very heavily indeed. As such things were defined in 1900 Pearson was an advanced thinker—for example, he was an imperialist and a racist and one of the founding fathers of neopositivism and eugenics. Seeking to resolve a tension between passion and science, ethics and rationality, Pearson mistook significance for "revelations about the objective world." In 1901 he believed 1.5 to 3 standard deviations were "definitely significant." By 1906, he tried to codify the sizeless stare with a Rule of Three and tried to teach it to Gosset.

19. Who Sits on the Egg of *Cuculus Canorus?* Not Karl Pearson 203

Pearson's journal, *Biometrika* (1901–), was for decades a major nest for the significance mistake. An article on the brooding habits of the cuckoo bird, published in the inaugural volume, shows the sizeless stare at its beginnings.

20. Gosset: The Fable of the Bee 207

Gosset revolutionized statistics in 1908 with two articles published in this same Pearson's journal, "The Probable Error of a Mean" and "The Probable Error of a Correlation Coefficient." Gosset also independently invented Monte Carlo analysis and the economic design of experiments. He conceived in 1926 the ideas if not the words of "power" and "loss," which he gave to Egon Pearson and Jerzy Neyman to complete. Yet most statistical workers know nothing about Gosset. He was exceptionally humble, kindly to other scientists, a good father and husband, altogether a paragon. As suits an amiable worker bee, he planted edible berries, blew a pennywhistle, repaired entire, functioning fishing boats with a penknife, and—though a great scientist—was for thirty-eight years a businessman brewing Guinness. Gosset always wanted to answer the how-much question. Guinness needed to know. Karl Pearson couldn't understand.

21. Fisher: The Fable of the Wasp 214

The tragedy in the fable arose from Gosset the bee losing out to R. A. Fisher the wasp. All agree that Fisher was a genius. Richard Dawkins calls him "the greatest of Darwin's successors." But Fisher was a genius at a certain kind of academic rhetoric and politics as much as at mathematical statistics and genetics. His ascent came at a cost to science—and to Gosset.

22. How the Wasp Stung the Bee and Took over Some Sciences 227

Fisher asked Gosset to calculate Gosset's tables of *t* for him, gratis. He then took Gosset's tables, copyrighted them for himself, and in the journal *Metron* and in his *Statistical Methods for Research Workers,* later to be published in thirteen editions and many languages, he promoted his own circumscribed version of Gosset's test. The new assignment of authorship and the faux machinery for science were spread by disciples and by Fisher himself to America and beyond. For decades Harold Hotelling, an important statistician and economist, enthusiastically carried the Fisherian flag. P. C. Mahalanobis, the great Indian scientist, was spellbound.

23. Eighty Years of Trained Incapacity: How Such a Thing Could Happen 238

R. A. Fisher was a necessary condition for the standard error of regressions. No Fisher, no lasting error. But for null-hypothesis significance testing to persist in the face of its logical and practical difficulties, something else must be operating. Perhaps it is what Thorstein Veblen called "trained incapacity," to which might be added what Robert Merton called the "bureaucratization of knowledge" and what Friedrich Hayek called the "scientistic prejudice." We suggest that the sizeless sciences need to reform their scientistic bureaucracies.

24. What to Do 245

What, then? Get back to size in science, and to "real error" seriously considered. It is more difficult than Fisherian procedures, and cannot be reduced to mechanical procedures. How big is big is a necessary question in any science and has no answer independent of the conversation of scientists. But it has the merit at least of being relevant to science, business, and life. The Fisherian procedures are not.

A Reader's Guide 253
Notes 255
Works Cited 265
Index 289

Preface

The implied reader of our book is a significance tester, the keeper of numerical things. We want to persuade you of one claim: that William Sealy Gosset (1876–1937)—aka "Student" of Student's *t*-test—was right and that his difficult friend, Ronald A. Fisher, though a genius, was wrong. Fit is not the same thing as importance. Statistical significance is not the same thing as scientific finding. R^2, *t*-statistic, *p*-value, *F*-test, and all the more sophisticated versions of them in time series and the most advanced statistics are misleading at best.

No working scientist today knows much about Gosset, a brewer of Guinness stout and the inventor of a good deal of modern statistics. The scruffy little Gosset, with his tall leather boots and a rucksack on his back, is the heroic underdog in our story. Gosset, we claim, was a great scientist. He took an economic approach to the logic of uncertainty. For over two decades he quietly tried to educate Fisher. But Fisher, our flawed villain, erased from Gosset's inventions the consciously economic element. We want to bring it back.

We lament what could have been in the statistical sciences if only Fisher had cared to understand the full import of Gosset's insights. Or if only Egon Pearson had had the forceful personality of his father, Karl. Or if only Gosset had been a professor and not a businessman and had been positioned therefore to offset the intellectual capital of Fisher.

But we don't consider the great if mistaken Fisher and his intellectual descendants our enemies. We have learned a great deal from Fisher and his followers, and still do, as many have. We hope you, oh significance tester, will read the book optimistically—with a sense of how "real" significance can transform your science. Biometricians who study AIDS and economists who study growth policy in poor countries are causing damage with

a broken statistical instrument. But wait: consider the progress we can make if we fix the instrument.

Can so many scientists have been wrong over the eighty years since 1925? Unhappily, yes. The mainstream in science, as any scientist will tell you, is often wrong. Otherwise, come to think of it, science would be complete. Few scientists would make that claim, or would want to. Statistical significance is surely not the only error in modern science, although it has been, as we will show, an exceptionally damaging one. Scientists are often tardy in fixing basic flaws in their sciences despite the presence of better alternatives. Think of the half century it took American geologists to recognize the truth of drifting continents, a theory proposed in 1915 by—of all eminently ignorable people—a German meteorologist. Scientists, after all, are human. What Nietzsche called the "twilight of the idols," the fear of losing a powerful symbol or god or technology, haunts us all.

In statistical fields such as economics, psychology, sociology, and medicine the idol is the test of significance. The alternative, Gossetian way is a uniformly more powerful test, but it has been largely ignored. Unlike the Fisherian idol, Gosset's approach is a rational guide for decision making and easy to understand. But it has been resisted now for eighty years.

Our book also addresses implied readers outside the statistical fields themselves such as intellectual historians and philosophers of science. The history and philosophy of applied statistics took a wrong turn in the 1920s, too. In an admittedly sketchy way—Ziliak himself is working on a book centered on Gosset—we explore the philosophy and tell the history here. We found that the recent historians of statistics, whom we honor in other matters, have not gotten around to Gosset. The historiography of "significance" is still being importantly shaped by R. A. Fisher himself four decades beyond the grave. It is known among sophisticates that Fisher took pains to historicize his prejudices about statistical methods. Yet his history gave little credit to other people and none to those who in the 1930s developed a decision-theoretic alternative to the Fisherian routine. Since the 1940s most statistical theorists, particularly at the advanced level, have not mentioned Gosset. With the notable exception of Donald MacKenzie, a sociologist and historian of science, scholars have seldom examined Gosset's published works. And it appears that no one besides the ever-careful Egon S. Pearson (1895–1980) has looked very far into the Gosset archives—and that was in 1937–39 for the purpose of an obituary.

The evidence on the Gosset-Fisher relationship that Ziliak found in the archives is startling. In brief, Gosset got scooped. Fisher's victory over

Gosset has been so successful and yet so invisible that a 2006 publication on *anti*-Fisherian statistics makes the usual mistake, effectively equating Fisher's approach with Gosset's (Howson and Urbach 2006, 133). In truth it was Gosset, in 1905, not Neyman, in 1938, who gave "the first emphasis of the behavioralistic outlook in statistics" (Savage 1954, 159).

Only slowly did we realize how widespread the standard error had become in sciences other than our home field of economics. Some time passed before we systematically looked into them. Thus the broader intervention here. We couldn't examine every science or subfield. And additional work remains of course to be done, on significance and other problems of testing and estimation. Some readers, for example, have asked us to wade in on the dual problems of specification error and causality. We reply that we agree—these are important issues—but we couldn't do justice to them here.

But we think the methodological overlaps in education and psychology, economics and sociology, agriculture and biology, pharmacology and epidemiology are sufficiently large, and the inheritance in them of Fisherian methods sufficiently deep, that our book can shed some light on all the *t*-testing sciences. We were alarmed and dismayed to discover, for example, that supreme courts in the United States, state and federal, have begun to decide cases on the basis of Fisher's arbitrary test. The law itself is distorted by Fisher. Time to speak up.

We invite a general and nontechnical reader to the discussion, too. If he starts at the beginning and reads through chapter 3 he will get the main point—that oomph, the difference a treatment makes, dominates precision. The extended but simple "diet pill example" in chapter 3 will equip him with the essential logic and with the replies he'll need to stay in the conversation. Chapter 17 through to the end of the book provides our brief history of the problem and a sketch of a solution.

Readers may find it strange that two historical economists have intruded on the theory, history, philosophy, sociology, and practice of hypothesis testing in the sciences. We are not professional statisticians and are only amateur historians and philosophers of science. Yet economically concerned people have played a role in the logic, philosophy, and dissemination of testing, estimation, and error analysis in all of the sciences from Mill through Friedman to Heckman. Gosset himself, we've noted, was a businessman and the inventor of an economic approach to uncertainty. Keynes wrote *A Treatise on Probability* (1921), an important if somewhat neglected book on the history and foundations of probability theory.

Advanced empirical economics, which we've endured, taught, and written about for years, has become an exercise in hypothesis testing, and is broken. We're saying here that the brokenness extends to many other quantitative sciences—though notably—we could say significantly—not much to physics and chemistry and geology. We don't claim to understand fully the sciences we survey. But we do understand their unhappy statistical rhetoric. It needs to change.

Acknowledgments

We thank above all Morris Altman, who organized a session at the American Economic Association meetings in San Diego on these matters (January 2004) and then edited the articles into a special issue of the *Journal of Socio-Economics* (no. 5, 2004). We have benefited over the years from the comments of a great many scientists, by no means all of them favorable to our views: Theodore W. Anderson, Kenneth Arrow, Orley Ashenfelter, Howard Becker, Yakov Ben-Haim, Mary Ellen Benedict, Nathan Berg, Kevin Brancato, James Buchanan, Robert Chirinko, Ronald Coase, Kelly DeRango, Peter Dorman, Paul Downward, Roderick Duncan, Graham Elliott, Deborah Figart, William Fisher, Edward Fullbrook, Andrea Gabor, Marc Gaudry, Robert Gelfond, Gerd Gigerenzer, Arthur Goldberger, Clive Granger, Daniel Hamermesh, Wade Hands, John Harvey, Reid Hastie, David Hendry, Kevin Hoover, Joel Horowitz, Sanders Korenman, William Lastrapes, Tony Lawson, Frederic Lee, Geoffrey Loftus, Peter Lunt, John Lyons, Andrew Mearman, Peter Monaghan, John Murray, Anthony O'Brien, David F. Parkhurst, John Pencavel, Gregory Robbins, William Rozeboom, David Ruccio, Thomas Schelling, Allan Schmid, George Selgin, Jeffrey Siminoff, John Smutniak, Gary Solon, Dwight Steward, Stephen Stigler, Diana Strassman, Lester Telser, Bruce Thompson, Erik Thorbecke, Geoffrey Tilly, Andrew Trigg, Gordon Tullock, Jeffrey Wooldridge, Allan Würtz, and James P. Ziliak.

Arnold Zellner, the late William Kruskal (1919–2005), Daniel Klein, Stephen Cullenberg, Kenneth Rothman, Edward Leamer, and the late Jack Hirshleifer (1925–2005) have our special thanks. Zellner, Kruskal, Rothman, Leamer, and Hirshleifer have long advocated sanity in significance. It has been inspiriting to have such excellent scientists saying to us, "Yes, after all, you are quite right." We would like especially to thank Arnold

Zellner for plying us with papers and books on Jeffreys's and Bayes's methods that we clearly needed to read. And we thank him, Kenneth Rothman, Regina Buccola, Roger Chase, Charles Collings, Joel Horowitz, Geoffrey Loftus, Shirley Martin, Stephen Meardon, Bruce Thompson, Erik Thorbecke, and the two reviewers, Peter Boettke at George Mason and Julian Reiss, who teaches in Spain, for reading and commenting on substantial parts of the manuscript. Late in the project Pete Boettke saw a need for a wider sociological explanation of an eighty-year-old mistake in science. Thus the penultimate chapter.

Collectively speaking we have been telling versions of our story for some decades now. McCloskey has been dining out on the idea of economic versus statistical significance for over twenty years. J. Richard Zecher first explained the point to her in the early 1980s when they were colleagues at the University of Iowa working on an article on the gold standard. Eric Gustafson had explained it to her when she was an undergraduate at Harvard, but after her "advanced" econometric training in Fisherian methods, vintage 1965, the point slipped away. In 1983 Harry Collins introduced her to the "significance test controversy" in psychology and sociology. She remembers a presentation to a large audience at the American Economic Association meetings in Dallas in 1984, with Edward Leamer and the late Zvi Griliches commenting; and smaller but still crucial seminars at Groningen, Oxford, and the LSE in 1996. The results of all this patient tuition, 1962–96, show up in her *The Rhetoric of Economics* (1985b [1998]).

Ziliak, too, has been the recipient of many courtesies. He first learned of the point in 1988, from the elementary book by Ronald and Thomas Wonnacott (1982), while working in cooperation with the U.S. Department of Labor as a labor market analyst for the Indiana Department of Employment and Training Services. When Ziliak pointed out to the chief of his division that black teenage unemployment rates were being concealed from public view he encountered puzzling resistance. Given the small sample sizes, the chief said, the unemployment rates did not reach an arbitrary level of statistical significance. But the Department of Labor, which authorizes the distribution of official labor market statistics, appeared to be saying that an average 30 or 40 percent rate of unemployment was not discussable because the *p*-values exceeded .10, the department's shut-up point. Ziliak was embarrassed to return to the telephone to deliver the news to the citizen whose call had started the inquiry. "Sorry, sir. We do not have any quantitative information about black teenage un-

employment in the cities." In 1989 he read the first edition of McCloskey's *The Rhetoric of Economics* (1985b [1998]), including the then startling chapters on "significance." Two years later he moved to Iowa and the graduate study of economics, where soon McCloskey invited him to join forces. Talks given jointly with McCloskey at Iowa (1993, 1994) and then solo at Indiana (1995) and the Eastern Economic Association (New York, 1995) transformed early puzzlement into action.

Throughout the 1980s and 1990s, then, we were talking and talking, individually and as a tag team, persuading a happy few. In recent years (we sense we have not mentioned all the events and apologize) we have found often appreciative and always attentive and sometimes stunned audiences at the annual meetings of the American Economic Association (Chicago, 1998; San Diego, 2004), Ball State University, Baruch College (School of Public Affairs), Bowling Green State University (a student seminar), the bi-ennial meetings of the Association for Heterodox Economics (University of Leeds, 2004), the Association for Heterodox Economics Post-graduate Workshop on Research Methods (University of Manchester, 2005), the University of Chicago (Center for Population Economics, 2005), the University of Colorado-Boulder, Dennison University, the Eastern Economic Association/Association for Social Economics (New York, 2003), the Elgin Community College/Roosevelt University Faculty Speaker Series, Erasmus University of Rotterdam, the summer institutes over many years of the European Doctoral Association in Management and Business, the First International Congress of Heterodox Economics (University of Missouri, Kansas City, 2003), George Mason University (Philosophy, Politics, and Economics Seminar), the University of Georgia, the Georgia Institute of Technology, Göteborg University (twice), Harvard University (a seminar for graduate students in economics; the faculty was skeptical), the University of Illinois at Chicago, Illinois State University, Macquarie University (Australia), the University of Michigan (another student seminar), the University of Nebraska, Northwestern University (Economic History Workshop), the University of Wisconsin (still another student seminar; the faculty was outraged), the Rhetoric and Economics Conference (organized by Paul Turpin at Milliken University, 2005), and annual meetings of the Southern Economic Association (New Orleans, 2004), the Ratio Institute of Stockholm (2006), and the University of Wollongong.

Ziliak gratefully acknowledges the cooperation of libraries and their staffs: University College London, Special Collections, where Gillian Furlong and Steven Wright gave able and kind access to the Galton Papers

and Pearson Papers (containing files on Karl Pearson, Egon Pearson, and Gosset, Fisher, and Neyman); the Guinness Archives, Diageo (Guinness Storehouse, Dublin, where Eibhlin Roche and Clare Hackett are themselves a "storehouse" of ideas); the Museum of English Rural Life, University of Reading; the University of Illinois at Chicago, Special Collections, Richard J. Daley Library, Science Library, Mathematics Library, and Health Sciences Library; the University of Chicago's Eckhart Library (for providing access to *Letters of William Sealy Gosset to R. A. Fisher, 1915–1936, Vols. 1–5* (private circulation, 1962) and its Regenstein, Yerkes, and Crerar libraries; Roosevelt University's Murray-Green Library; Emory University's Woodruff Library; and the libraries at the Georgia Institute of Technology, the University of Iowa, and Bowling Green State University. For research assistance at various stages of the project we were fortunate to employ Cory Bilton, Angelina Lott, David McClough, and Noel Winter.

Ziliak also thanks the Institute for Humane Studies for a Hayek Scholar Travel Grant and Roosevelt University for two summer grants used to collect primary materials in London, Dublin, Reading, and Chicago. Roosevelt is a rare site of sanity in academic life, serious about justice and freedom. He thanks there many colleagues who have tolerated his brief lectures on significance over lunch, especially Stefan Hersh, a friend who combines oomph *and* precision and Lynn Weiner and Paul Green, for openhandedly helping. In London Ziliak was well cared for, too. Andrew Trigg and his sons provided an amusing diversion from Gower Street, and Sheila Trigg, an Oxford-trained political adviser and chef extraordinaire, was her usual goddess self. McCloskey thanks the College of Liberal Arts and Sciences at the University of Illinois at Chicago for continuing research moneys used for the project and her colleagues at UIC, such as Lawrence Officer, who have Gotten It.

We also gratefully acknowledge the University of Wisconsin Press, the *Journal of Economic Literature,* the *Journal of Socio-Economics,* and *Rethinking Marxism* for permission to use some of our earlier writings and statistics; University College London, Special Collections Library, for permission to quote from the Galton Papers and Pearson Papers; Guinness Archives (Diageo) for permission to reproduce images of Gosset; the *Journal of Socio-Economics* and Professor Erik Thorbecke for permission to print a version of Thorbecke's (2004) very illuminating figure on economic significance; *Biometrika,* for permission to reprint a page from Student 1908a; *Educational and Psychological Measurement* for permission to

reprint a table from Fidler et al. (2004b); the Johns Hopkins University School of Hygiene and Public Health and the *American Journal of Epidemiology* for permission to reproduce a table from Savitz, Tolo, and Poole (1994); the *New England Journal of Medicine* for allowing a version of a figure from Freiman et al. (1978); and Professor Kenneth Rothman for supplying an unpublished graph of a *p*-value function. James F. Reische, our editor at the University of Michigan Press, carried our book over the hurdles.

Ziliak would be lost without Flora, Jude, and Suzette, whose love is all about oomph. And he dedicates the book to his parents, Barbara and Lawrence Ziliak, real world examples of unconditional love. McCloskey dedicates the book to her grandchildren, Connor and Lily. May they someday read this and understand a part of love.

Love comes in more academic forms, too: together we dedicate the book to the memory of William H. Kruskal for his many kindnesses extended from the 1970s to the 2000s and for a long life in theoretical and applied statistics of substantive significance.