Lady Justice v. Cult of Statistical Significance:
Oomph-less Science and the New Rule of Law

By Stephen T. Ziliak and Deirdre N. McCloskey

Roosevelt University and the University of Illinois-Chicago

Email: sziliak@roosevelt.edu; deirdre2@uic.edu

July 2014

Forthcoming, *Oxford Handbook on Professional Economic Ethics* (Oxford University Press, 2014), edited by G. DeMartino and D. N. McCloskey

Abstract: We have an ethical problem in economics and other sciences using null hypothesis statistical significance testing without a loss function—a test that avoids asking, How Big is a Big Loss or Gain? Statistical significance is not equivalent to economic significance, nor to medical, clinical, nor any other kind of scientific significance—those functions of gain and loss. The mistake in the falsely made equation is evident when one reflects that the estimated payoff from a lottery is not the same object as the odds of winning that lottery. Yet a widespread failure to make the distinction between an estimate of human consequence and an estimate of its probability—between the meaning of an estimated average and the random variance around it—is killing people in medicine and impoverishing people in economics (Ziliak and McCloskey 2008). The ethical problem created by a test of statistical significance is made worse by the method's blatant illogic at the foundational level, a fact that is unacknowledged by the bulk of decision makers depending upon it. Several changes to the scientific paper—and a recent decision by the Supreme Court of the United States—could help.

Keywords: ethics, statistical significance, oomph, Matrixx v. Siracusano, Fisher, Gosset
JEL codes: C10, C12, C13, B23, A13

Lady Justice v. Cult of Statistical Significance:
Oomph-less Science and the New Rule of Law

By Stephen T. Ziliak and Deirdre N. McCloskey

July 2014

Forthcoming, *Oxford Handbook on Professional Economic Ethics* (Oxford University Press, 2014), edited by G. DeMartino and D. N. McCloskey

This statistical significance always works and always doesn't work.

- Stephen Breyer, Associate Justice, U.S. Supreme Court

We have an ethical problem in economics and other sciences using null hypothesis statistical significance testing without a loss function—a test that avoids asking, How Big is a Big Loss or Gain from the null?

Statistical significance is not equivalent to economic significance, nor to medical, clinical, biological, psychometric, pharmacological, legal, physical, nor any other kind of scientific significance—those functions of gain and loss. The mistake in the falsely made equation is evident when one reflects that the estimated payoff from a lottery (of, say, one million U.S. dollars) is not the same object as the odds of winning that lottery (odds of, say, one in two million chances). Yet a widespread failure to make the distinction between an estimate of human consequence and an estimate of its probability—between the meaning of an estimated average and the random variance around it—is killing people in medicine and impoverishing people in economics (Ziliak and McCloskey

2008). The ethical problem created by a test of statistical significance is made worse by the method's blatant illogic at the foundational level, a fact that is unacknowledged by the bulk of decision makers depending upon it.

The Statistically Significant Median Is Not the Message

In sciences from accounting to zoology the errors from null hypothesis significance testing without a loss function—without some quantitative standard of meaningful gain or loss, separate from the probability of its occurrence—pile up daily. Adam Smith noted that "whatever praise or blame can be due to any action, must belong either . . . to the . . . affection of the heart. . . or . . . to the external action . . . which this affection gives occasion to. . . or to the . . . consequences which. . . proceed from it" (Smith 1790, Pt II, Sect iii). How *much* unemployment, or inflation, or toxic asset bailout, is too much? That is the scientific, and ethical, question. We are not doubting "the affection of the heart" of the average econometrician. We find him instead neglecting Smith's second and third elements of an ethical judgment. Most significance testers—for instance, 80% in economics and 90% in breast cancer epidemiology, we have found—fail to take the correct "external action" with their evidence. And they neglect the ethical, economic, and other "consequences." They are testing by an ethically irrelevant criterion.

By the 1930s the null hypothesis significance test at the 5 percent level was considered so sophisticated in economics, psychology, and medicine "that these studies might be raised to the ranks of sciences," according to Ronald A. Fisher, the inventor

and steady advocate of the test of two standard deviations (Fisher 1925, Preface). He wrote, "Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and *ignore entirely* all results which fail to reach this level" (Fisher 1926). Fisher kept saying so up to his death in 1962, exerting a great influence on others. The sociologist Robert Merton would have considered the cult he initiated, based on merely statistical significance, to be a social pathology of science—a "bureaucratization of knowledge" (quoted in Ziliak and McCloskey 2008, p. 243). "Adherence to the rules," Merton wrote, "originally conceived as a means, becomes transformed into an end-in-itself" (Merton 1949, p. 199).

The core problem is that statistical significance is neither necessary nor sufficient for testing an ethical, scientific, commercial, or material fact in a court of law or of scientific and business opinion. An insignificant coefficient can be substantively significant, important to real people or eco-systems, if for example the possible loss from ignoring it (as in "ignore entirely", as Fisher urged) is large enough. And a statistically significant coefficient might be economically speaking irrelevant to the choice problem, if for example the significance is caused by merely having a very large sample size, or if the variable to which the coefficient is attached is not in any case a policy tool.

What people want from the analyst is a demonstration of oomph, of meaningful effect size, whether large or small. Probability comes in only as an inquiry into the odds of various levels of oomph. What we have in statistical sciences today, by contrast, is a table of Student's *t*-statistic distributed on a scale of Fisher's *p*-value, ranging in

probable error from .05 in economics and medicine to, in some parts of particle physics, $p < .00000001$ (but see Ziliak 2013—most physics does *not* use such tests, but rather interocular trauma: does the result hit you between the eyes?). A reported finding of "significant" or "not significant" does not answer the scientific question, which is always How Big along a scale agreed to by other scientists or clinicians of How Big *Is* Big.

The need to consider the human meaning of effect sizes, our "oomph," throughout the distribution was well described by Stephen Jay Gould in "The Median Isn't the Message" (Gould 1985). In July 1982 Gould had learned that he "was suffering from abdominal mesothelioma, a rare and serious cancer usually associated with exposure to asbestos." He was told by his doctor that the *median* survival time after discovery of the rare cancer is eight months.

> What does "median mortality of eight months" signify in our vernacular? I suspect that most people, without training in statistics, would read such a statement as "I will probably be dead in eight months" . . . . [But] it isn't so, and . . . attitude matters [to recovery]. The distribution was. . . strongly right skewed. . . . I saw no reason why I shouldn't be in that small tail, and I breathed a very long sigh of relief. . . . I had read the graph correctly. . . . I didn't have to stop and immediately following Isaiah's injunction to Hezekiah –set thine house in order: for thou shalt die, and not live.

He recovered and survived for another 20 years—more than 30 times the median survival time. Oomphful, for sure.

Unhappily, for the past 90 years in fields infected by Fisher's rule the merely "statistically significant" median or mean is the sole message in economics and most other branches of science, ignoring the oomph. Since the first large scale survey of best practice significance testing in economics, covering the 1980s in the *American Economic Review*, the significance mistake has gotten worse, not better (Ziliak and McCloskey 2008; cf. McCloskey and Ziliak 1996).

Matrixx Inverted: High Court Rejects Significance at the 0% Level

A new rule of law, handed down in 2011 by the Supreme Court of the United States, might help (Supreme Court 2011b). The high court ruled that companies can no longer conceal from investors relevantly bad news about their products by claiming that the adverse effects are not "statistically" significant at the 0.05 or any other level. Companies must consider the human meaning of over- or under-estimation. Statistical significance without a loss function is no longer the rule of securities law. Substance, effect size, oomph, is. On March 22, 2011, in *Matrixx Initiatives, Inc. v. Siracusano*, No. 09-1156, the Supreme Court rejected Fisher's rule by a 9-0 vote.

The case involved a homeopathic medicine called Zicam, a zinc-based cold remedy produced by Matrixx Initiatives. When swabbed or sprayed in the nose the drug is expected to reduce incipient colds. But it also causes some users to lose permanently their sense of smell (and thus of taste), a condition called anosmia. The loss function here is a function, then, of a high probability of stopping a cold balanced

against a low probability of losing all taste of food and not smelling the flowers or your lover, ever again.

When a doctor appeared on the *Good Morning America* television show in 2004 explaining the dangers of zinc-based treatments, Matrixx stock price plummeted. The company replied, though, that the adverse effect reports were not statistically significant. The company assured investors that revenue from Zicam, a hundred million dollar a year seller, was expected to grow vastly—by "50 and then 80 percent" (Supreme Court 2011b, p. 3).

In the January 10, 2011 oral arguments before the Supreme Court, Justice Sotomayor chastised counsel for the petitioners (petitioning, that is, to have an appeals-court ruling against Matrixx reversed [Supreme Court 2011a]). "Mr. Hacker" was chastised for neglecting to respond to technical briefs on the subject that had been authored and filed by *amici* of the court. Many of the friends of the court, the Justice said, "did a wonderful job." (Full disclosure: we were two of the *amici* [McCloskey and Ziliak 2010]. As is common in such matters, though, the "wonderful job" was mostly done by Allan Ingraham, an economist who drafted the brief for a law firm on the basis of our writings.)

Investors in Matrixx stock had filed suit against the company in a federal district court. They told the court that the company had failed to disclose the bad news it had received from expert nose doctors. But the district court dismissed the suit on the basis that investors did not prove "materiality," which meant, under then-existing precedents, statistical significance. Statistical significance had long since become part

of securities law: if it is statistically "insignificant" then, however illogical, it is materially insignificant, too. The Court of Appeals for the Ninth Circuit then reversed the district court's decision, reasoning in a narrow fashion "that whether facts are statistically significant, and thus [under the then-existing rule of law] material, is a question of fact that should ordinarily be left to the trier of fact—usually the jury."

The Justices went deeper. They disagreed with the definition of materiality invoked by the district court in the first place. The Justices said that the district court "erred when it took liberties in making that determination on its own." "Something more is needed," Justice Sotomayor wrote for the unanimous Court, and the something, she said, should address the "source, content, and context" of the bad news. Matrixx v. Siracusano presented the Court with the question whether plaintiffs can sustain a claim of securities fraud against a company neglecting to warn investors about bad news that is *not* statistically significant. Nine to zero it ruled that they can.

The Court is not well known for economic or statistical sophistication. But in this case it got it right. The precedent, now the law of the land, should be followed, we believe, for all statistical reporting, nine to zero, from climate change research to randomized field experiments in developing nations. In other words, loss functions matter. Oomph is what we seek. And oomph, not the level of Student's *t*, is the new rule of law.

What a Reasonable Investor or Impartial Spectator Might Say

The Court examined the expectations of a "reasonable investor." Would

undisclosed bad news be likely to negatively affect the "total mix" of information considered by a reasonable investor? If yes, then the report must be disclosed, regardless of statistical significance or insignificance. Sotomayor wrote for the Court (Supreme Court 2011b),

> medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence [unhappily the movement for "evidence-based medicine" may falsify her claim]. . . . The FDA similarly does not limit the evidence it considers for purposes of assessing causation and taking regulatory action to statistically significant data. In assessing the safety risk posed by a product, the FDA considers factors such as "strength of the association," "temporal relationship of product use and the event," "consistency of findings across available data sources," "evidence of a dose-response for the effect," "biologic plausibility," "seriousness of the event relative to the disease being treated," "potential to mitigate the risk in the population," "feasibility of further study using observational or controlled clinical study designs," and "degree of benefit the product provides, including availability of other therapies.". . . [The FDA] does not apply any single metric for determining when additional inquiry or action is necessary.

To the theory of the attorneys for Matrixx that statistical significance set the standard for disclosure, over and above "background noise," Justice Breyer (Supreme Court 2011a, p. 22) replied to a Mr. Hacker, "Oh, no, it can't be. I mean, all right –I'm

sorry. I don't mean to take a position yet." [Laughter.]

JUSTICE BREYER. But, look—I mean, Albert Einstein had the theory of relativity without any empirical evidence, okay? So we could get the greatest doctor in the world, and he has dozens of theories, and the theories are very sound, and all that fits in here is an allegation he now has learned that it's the free zinc ion that counts.

MR. HACKER. But . . . .

JUSTICE BREYER. And that could be devastating to a drug even though there isn't one person yet who has been hurt.

To Hacker's argument that statistical "significance" is the way to truth, Breyer snorted, "This statistical significance always works and always doesn't work." In the same session Sotomayor (citing *amici*) said that what counts as "statistical importance can't be a measure because it depends on the nature of the study." Justices Kagan and Ginsberg argued that small numbers of humanly meaningfully large effects can be materially relevant, independent of the level of statistical significance. Thus the loss function. Loss of smell is bad enough, but suppose (a small number of) people died? Kagan referred to a situation in which a small number of instances of blindness were known to be associated with the use of a contact lens solution. The FDA, she noted, would not wait around for statistical significance to make a determination or to investigate further into the facts of such black swans. Chief Justice Roberts sympathized with the test of expectations of a "reasonable investor," concluding that statistical significance was not necessary for establishing causation or belief in association. Sotomayor in the Court's

decision again (Supreme Court 2011b, pp. 1-2, 11):

> We conclude that the materiality of adverse event reports cannot be reduced to a bright-line rule. Although in many cases reasonable investors would not consider reports of adverse events to be material information, respondents have alleged facts plausibly suggesting that reasonable investors would have viewed these particular reports as material". . . . Matrixx's argument rests on the premise that statistical significance is the only reliable indication of causation. This premise is flawed"

Significance is not Material

The Matrixx decision is consistent with the high court's prior rejection of a bright-line rule in a fact-finding and economically important situation. Citing Basic v. Levinson (1976), a case involving a bright-line definition for what is meant by "merger negotiations," Justice Sotomayor argued (Supreme Court 2011b) that "we observed [in Basic] that 'any approach that designates a single fact or occurrence as always determinative of an inherently fact-specific finding such as materiality, must necessarily be overinclusive or underinclusive'."

Consider a pill that is thought to be effective at relieving pain, but at the cost of an increased risk of heart attack. Suppose a well-designed experiment is conducted on a sample of adult humans, half taking the drug, the other half taking another and competing drug. The significance tester—in search of a single, determinative fact—then

poses the question: "Assuming there is no real difference between the two pills, what is the chance that the data—showing some amount of difference—will be observed?" If the chance of seeing a difference in adverse effect larger than the one observed is less than or equal to 5 percent it is declared to be statistically significantly different from the null hypothesis of "no difference," without saying how much that difference is, or how one should view it. But it is an ethically flawed procedure, and before the Supremes spoke it was accepted by American law.

In the early 2000s, around the time that Matrixx and Zicam were getting into trouble, a much larger producer, Merck pharmaceutical, got into billions of dollars of trouble with their Vioxx pill. Vioxx-takers began to die from heart disease and attacks. In a clinical trial the Merck scientists reported that Vioxx takers risk a big adverse effect—death. Yet the *p*-value came in at 0.20, meaning that a 4-to-1 or higher odds of experiencing a major cost (such as death) is not worthy of policy consideration, because not "statistically" significant at p=0.05 or higher than 19-to-1 odds (see Ziliak and McCloskey [2008], Chapter 3). Therefore the company neglected the adverse outcomes. Therefore they committed the error of under-inclusiveness, a deathblow to science and lives, an error caused by unnecessary adherence to a bright line rule of statistical significance.

We Need Something More

What the Supreme Court did not say is that the test of significance gives us the *wrong* information, period. The test gives a probability of finding a larger difference

than that observed in the sample on offer, assuming that treatment and control drugs are actually the same. But that is "the fallacy of the transposed conditional" (Ziliak and McCloskey 2008). What we really want to know is the probability of a hypothesis being true (or at least practically useful), given all the data we've got, not the other way around. We want to know the probability that the two drugs are *different*, and by how much, given the available evidence. The significance test—based as it is on Fisher's fallacy of the transposed conditional—does not and cannot tell us that probability. The power function, the expected loss function, and many other decision-theoretic and Bayesian methods descending from William Sealy Gosset (known in the trade as "Student," as in Student's *t*) and Harold Jeffreys, now widely available, do (Ziliak 2008).

A "significant" result does not in any way answer the How Much question, the question of how much or how valuable the difference in magnitude is (such as loss of smell or sight, or relief from pain, or nipping a cold in the bud). The significant result cannot demonstrate economic, medical, or any other importance, for the obvious reason that it does not address it. In other words, we want to know the probability of detecting a *large and practically important difference* when the difference is truthfully there. We need exploratory methods, a power function, an expected loss function, and ideally speaking a series of independently repeated experiments controlling for random and real error.

Here are four small but important improvements suggested by Gosset and a lengthy list of first-rate statisticians after him:

1.) *Visualize the data, the whole distribution*.

Visualizing uncertainty leads most people—highly trained econometricians included—to make better science, approaching the standard of most physics, biology, geology, and the rest of the non-Fisherian sciences in examining interocular trauma (Soyer and Hogarth 2012). For key variables a scientific article would do better to present graphical and other representations of the whole distribution *on axes of effect size and their likelihood*. That way other scientists can judge by eye whether Big is Big or Small is Small in a meaningful metric of gain and loss. At present it is nearly impossible to extract such information from the usual computer dump of six regression specifications, complete with asterisks on "significant" variables.

Gosset himself was a visualizer. He conveyed to fellow statisticians the then-novel idea of "kurtosis" (for example, the highly non-normal variation of Gould's cancer-patient survival data) with ink drawings of two kangaroos face-to-face illustrating long tails and a platypus with a short, asymmetric tail (Student 1927, p. 160). As Gould learned first-hand, the median is not the message. An ethical economist will graph it out.

2.) *Report Power and Real Type I Error, not the Nominal*

Power is protection against merely statistical significance. Power asks, "What in the proffered experiment is the probability of *correctly* rejecting the null hypothesis, concluding that the null hypothesis is indeed false when it *is* false?" If the null hypothesis is false perhaps another hypothesis—at some other effect size than literally and exactly zero—is true. The power *function* graphs the probability of rejecting the null hypothesis as a function of various assumed-true effect sizes, 0 or 3 or 2034. The further the actually true

effect size is away from the null, the easier it is going to be in an irritatingly random world to reject the null, and the higher is going to be the power.

To calculate a power function one needs a random sample, a fixed level of significance (Type I error of, say, .05), and one or more measures of effect size different from the null and from the result obtained. The effect size is the assumed efficacy in God's eyes, so to speak, which we should be in the business of uncovering. If you have a very large sample there is no problem of power. With n = 10,000 even weak effects will show through a cloud of skepticism. Everything will be significant and with high power—though in that case the significance of an effect, or for that matter its power, is not itself much of an accomplishment. If you are a Fisherian, though, a large sample *becomes* your problem. You will be deluded into thinking you've proved oomph before you've considered what it is.

Suppose a pill does in fact work to the patient's benefit, though with sampling uncertainty. What you want to know—and are able in most testing situations to discover—is with how much power you can reject the null of "*no* efficacy" when the pill is in truth efficacious *to such-and-such a clinically interesting degree*. In general, the more power you have the better. You do not want by the vagaries of sampling to be led to reject what is actually a *good* pill. That way lies the ethical disaster of death or discomfort for the patients (a death rate which in fact the best medical statisticians have calculated, and have found large, in criticizing the reliance on "significance" alone).

Power is powerful because hypotheses are plural and the plurality yields overlapping probability distributions. In a random sample a sleeping pill Napper may on

average induce 3 extra hours of sleep on the airplane, plus or minus 3. But in another sample the same scientist may find that the same sleeping pill induces 2 extra hours of sleep, plus or minus 4 (after all, some sleeping pills contain stimulants, causing negative sleep). The traveler would like to know from her doctor before she takes the pill exactly how much confidence she should have in it. She reasonably wants to know, "With what probability can I expect to get the additional 2 or 3 hours of rest on my overnight to Heathrow? And with what probability might I actually get *less* rest?"

Scientists express "real" Type I error as the ratio of the *p*-value to the power of the test. Thus an alleged $p = .05$ will turn out actually to be an alarming "real" p of .20 if the power of the test is only .25. An alleged $p = .10$ is really .33 if the power of the test is .30. Dozens of power studies have shown how often this is indeed the case for small effect sizes. Mosteller and Bush (1954) were the first to have assessed the amount of statistical power in the social sciences. The psychologist Jacob Cohen was the first to conduct a large-scale, systematic survey in psychology (Cohen 1962, cited in Ziliak and McCloskey 2008). He surveyed 70 articles published in the *Journal of Abnormal and Social Psychology* for the year 1960, excluding minor case reports, factor-analytic studies, and other contributions for which the calculation was impossible. Little power in the tests, he found, and little use for psychologists.

According to a large scale study by Joseph Rossi (1990) of top psychology journals the real rate of false rejections of null hypotheses is equally grim: "More than 90% [of over 6,000] of the surveyed studies had less than one chance in three of detecting a small effect" — as against the Fisherians' false claim of 5% error. Decision makers need to know

that the real rate of false rejection is for small effect sizes at best .05 divided by.17, or about 29%, and for medium-sized effects .05 divided by.57, or about 9%. Reporting the *real* level of Type I error has the advantage of allowing the reader to approximate how many "false" rejections of the null will occur for every "true" or correct rejection. Real Type I error adjusts the *p* value upward, by basing the value on the rate of true rejections—which is certainly more relevant than a baseline of nothing at all. Most Fisher-tests reject with a power of less than 50 percent, with the predictive accuracy of your local psychic.

Freiman et al. (1978) found that if the authors of the 71 original medical studies they examined had considered the power of their tests, the studies would not have ended "negatively." That is, they would have found that the therapy was capable of producing "important therapeutic improvement." Freiman et al. reckoned that if 50 of the 71 studies published in the *New England Journal of Medicine* had paid attention to power and effect size and not merely to a one-sided, qualitative, yes/no interpretation of "significance," they would have *reversed* their conclusions. Astonishingly, they would have found up to "50 per cent improvement" in "therapeutic effect." The premature negative results were published in *Lancet*, the *British Medical Journal*, the *New England Journal of Medicine*, the *Journal of the American Medical Association*, and other elite journals. Effective treatments for cardiovascular and cancer and gastrointestinal patients were abandoned because they did not attain statistical significance at the 5% level.

The Fisherian tests of significance—the only tests employed by the original authors of the 71 medical studies—literally could not see the beneficial effects of the

therapies under study, though staring at them. The point about hearts and cancer is the same as the Gosset point about Guinness beer (Ziliak 2008), which is the same as the Neyman and Pearson point about justice, which is the same as the McCloskey point about purchasing power parity (McCloskey 1985a and 1985b), which is the same as the Ziliak point about black unemployment rates (Ziliak and McCloskey 2008), which is the same as the Jeffreys point about *p*-values.

Yet high power is no perma-shield against other kinds of oomph-ignoring errors rife in the sciences obsessed with significance testing. To estimate the power function one needs to define among other things a domain of relevant effect sizes different from the null. And that decision, as Gould recognized when he was diagnosed with cancer, is about oomph. The 2003 article on Vioxx is proof of what can go wrong when oomph of the test is not attended to, even if the power of the test is. In the study on which Vioxx was marketed, "A sample size of 2780 patients per treatment group was expected to provide 90% power to detect a difference of 2 percentage points between treatments for the primary safety variable" (Lisse et al., p. 541). But the authors did not estimate the power of their test to reject the hypothesis of no harmful cardiac effect between Vioxx and naproxen. Pretending to be excessively gullible, they ignored an 8-to-1 cardiac damage or death ratio, a magnitude or "safety variable" of some importance.

3.) *Define, report, and examine the expected loss function*

Scientists and their clients wish to have relevance, not reports of amateur philosophy derived from positivism c. 1920, which is the method on offer in much of

the statistical sciences. They wish for standards that will help them, say, minimize the maximum loss of jobs, income, profit, health, or freedom in following this or that hypothesis as if true. The validity of the loss function is not sensitive to the degree of risk-aversion felt by an investigator. It need not be, to mention the automatic objection from Fisherian statisticians, "subjective." "Loss" is in the Gossetian tradition of statistics the value of a sacrifice, and may be positive or negative, properly addressing the *economic* meaning of significance. So even the gambler who sees nothing but blue sky—a real risk lover—will distinguish maximum losses from minimum, big wins from small. Adjusting the levels of Type I and Type II error is usually sufficient for handling differential attitudes toward risk. The problem is that today's statistical experts do not employ the loss function or Type II error at all. In his last year the great statistician and economist Leonard Savage asked, "When is one [statistical] expert, real or synthetic, to be preferred to another?" He replied: "Employ, until you have further experience, that expert whose past opinions, applied to your affairs, would have yielded you the highest average income" (Savage 1971b, pp. 145-6). Substitute "highest average income"—or rather add to it—other concerns, such as "highest average quality" or "highest rate of patient survival" or "lowest number of heart attacks" or "highest average rate of minority student graduations" or, in less practical studies, "highest acceptance rate of my results among my scientific colleagues," and you have what we are claiming here.

In any case, without a power and loss function a test of statistical significance is meaningless, no better than a table of random numbers. Pretending to afford a view from everywhere, statistical significance is in fact a view from nowhere. In its desire to

maximize precision in one kind of error from sampling (often enough, by the way, it is applied not to samples put to populations, the urn of nature spilled out on the floor), it turns away from the human purposes that motivated the research in the first place. Savage noted in his *Foundations* a part of the problem we are highlighting: "Many [scientists following in the footsteps of Karl Pearson and R. A. Fisher] have thought it natural to extend logic by setting up criteria for the extent to which one proposition tends to imply, or provide evidence for, another. . . . It seems to me obvious, however, that what is ultimately wanted is criteria for deciding among possible courses of action." Following Fisher's words precisely, significance testers do not think at all about "possible courses of action," holding themselves to be ethically harmless (they mistakenly believe) from possible courses of action. As Fisher declared late in his career, explicitly rejecting the loss functions introduced thirty years earlier by Gosset, Wald, Egon Pearson, and others:

> *Finally, in inductive inference we introduce no cost functions for faulty judgments*, for it is recognized in scientific research that the attainment of, or failure to attain to, a particular scientific advance this year rather than later, has consequences, both to the research programme, and to advantageous applications of scientific knowledge, which cannot be foreseen. . . . *We make no attempt to evaluate these consequences, and do not assume that they are capable of evaluation in any currency.*
>
> (Fisher 1955, p. 75, italics supplied)

4.) *Consider the choice of sample size, experimental design, and its implications for external validity.*

For example, is the experiment repeated once, twice, twelve, or thirty times, or is it not repeated at all? People need to know. The Chinese eyeglass experiment, described by Ziliak and Teather-Posadas (2014, this volume), did not need to grow to 19,000 schoolchildren before concluding that prescription eyeglasses help at school. We already knew it. As in the Tuskegee Syphilis Trial, a leave-out group is unethical if we already know. In the Vioxx clinical trial younger people with stronger hearts were overrepresented and elderly people with weaker hearts were underrepresented. Thus Gosset's recommendation of stratification, balance, repetition, and opportunity cost in experimental design (Ziliak 2014, 2011; Ziliak and Teather-Posadas 2014).

Oomphful Economics

The economic approach to the logic of uncertainty—pioneered by Gosset at Guinness—is a better way forward. Bruno de Finetti said that "The economic approach seems (if not rejected owing to aristocratic or puritanic taboos) the only device apt to distinguish neatly what is or is not contradictory in the logic of uncertainty." The new rule of law should help.

Works Cited

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

Fisher, R. A. 1925 [1928]. Statistical methods for research workers. G.E. Stechart, New York.

Fisher, R. A. 1926. Arrangement of field experiments. Journal of Ministry of Agriculture 33, 503-13.

Fisher, R.A. 1933. The contributions of Rothamsted to the development of the science of statistics. In: Rothamsted Experimental Station, Annual report. Rothamsted, Rothamsted, pp. 43-50.

Fisher, R. A. 1935. The design of experiments. Oliver & Boyd, Edinburgh.

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

Freiman, J. A., T. Chalmers, H. Smith, R. R. Kuebler. 1978. "The Importance of Beta, the Type II Error and Sample Design in the Design and Interpretation of the Randomized Control Trial: Survey of 71 "Negative" Trials." *New England Journal of Medicine* 299: 690-4.

Gould, S. J. 1985. "The Median Isn't the Message," *Discover* 6 (June): 40-42.

Jeffreys, H. 1931 [1973]. *Scientific Inference*. Cambridge: Cambridge University Press.

Jeffreys, H. 1939a [1967]. *Theory of Probability*. Third revised edition. London: Oxford

University Press.

Lisse, J. R. 2003. And Monica Perlman, Gunnar Johansson, James R. Shoemaker, Joy Schectman, Carol S. Skalky, Mary E. Dixon, Adam B. Polis, Arthur J. Mollen, and Gregory P. Geba. "Gastrointestinal Tolerability and Effectiveness of Rofecoxib [Vioxx®] versus Naproxen in the Treatment of Osteoarthritis." *Annals of Internal Medicine* 139 (Oct.): 539-546.

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

McCloskey, D. N., S. T. Ziliak with Edward Labaton et al. Counsel of Record (Ed.), *Brief of Amici Curiae Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents* (vol. No. 09-1156, pp. 22). Washington DC: Supreme Court of the United States.

McCloskey, D.N., S.T. Ziliak. 1996. "The Standard Error of Regressions," Journal of Economic Literature 34 (March): 97-114.

Merton, R. M. 1949 [1957]. *Social Theory and Social Structure.* New York: The Free Press.

Rossi, J. 1990. "Statistical Power of Psychological Research: What Have We Gained in 20 Years? *Journal of Consulting and Clinical Psychology* 58: 646-56.

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

Savage, L. J. 1971b [1974 posthumous]. "Elicitation of Personal Probabilities." Pp. 111-56 in Fienberg and Zellner, eds., *Studies in Bayesian Econometrics.*

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

[Dover edition].

Smith, A. 1759 [1791, 2009]. The Theory of Moral Sentiments. New York: Penguin. Introduction by Amartya Sen.

Soyer, E., R. Hogarth. 2012. The illusion of predictability: How regression statistics mislead experts. International Journal of Forecasting 28 (3): 695-711.

Student. 1938 [posthumous]. Comparison between balanced and random arrangements of field plots. Biometrika 29, 363-78.

Student. 1927. Errors of routine analysis. Biometrika 19 (1/2): 151-164.

Supreme Court of the United States. 2011a. Oral arguments in Matrixx Inititiaves,, January 10 http://www.supremecourt.gov/oral_arguments/argument_transcripts/09-1156.pdf

Supreme Court of the United States. 2011b. Matrixx Inititiaves, Inc., et al., No. 09-1156, Petitioner v. James Siracusano et al., On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit, March 22, 2011. 25 pp., syllabus.

Ziliak, S.T. 2014. Balanced versus randomized field experiments in economics: Why W.S. Gosset aka 'Student' matters. Review of Behavioral Economics 1 (1): 167-208.

Ziliak, S.T. 2013. Junk science week: Unsignificant statistics. Financial Post, June 10.

Ziliak, S. T. 2008. Guinnessometrics: The economic foundation of 'Student's' *t*. Journal of Economic Perspectives 22 (Fall): 199-216.

Ziliak, S. T., D.N. McCloskey. 2008. The cult of statistical significance: How the standard

error costs us jobs, justice, and lives. University of Michigan Press, Ann Arbor.

Ziliak, S. T., E.R. Teather-Posadas. 2014. The unprincipled randomization principle in economics and medicine. Oxford Handbook on Professional Economic Ethics. New York: Oxford University Press, eds. G. DeMartino and D.N. McCloskey.