Dear Monica, Arny, Mark and All -

Thanks so much again for all of your kind hospitality and interest. I truly enjoyed my visit to the UK (exchange rate included!) and—to Arny, Mark, and Monica—in addition, many thanks to you for your thoughtful replies—btw, may I hire you to write reviews?!

Ziliak and McCloskey (2008, p. 2) do not claim originality for divining the main point of *The Cult of Statistical Significance*—that Precision is not the same thing as Oomph and vice versa. Here is how we put it in the book (p. 2):

"[O]ne part of mathematical statistics has gone terribly wrong, though mostly unnoticed. The part we are worrying about here seems to have all the quantitative solidity and mathematical shine of the rest. But it also seems—unless we and some other observers of mathematical statistics such as Edgeworth, Gosset, Egon Pearson, Jeffreys, Borel, Neyman, Wald, Wolfowitz, Yule, Deming, Yates, Savage, de Finetti, Good, Lindley, Feynman, Lehmann, DeGroot, Chernoff, Raiffa, Arrow, Blackwell, Friedman, Mosteller, Kruskal, Mandelbrot, Wallis, Roberts, Granger, Press, Berger, and Zellner are quite mistaken—that reducing a scientific problem of measurement and interpretation to a narrow version of "fit" or of "statistical significance," as some sciences have done for more than 80 years, has been an exceptionally bad idea."

True.

We devote two whole chapters of our book to Meehl, Cohen, and others cited and not cited by Monica. (BTW, I'm sorry we didn't know about your very fine article, Monica—we can make good on it in the second edition!) Meehl, Cohen and many others in psychology-- from Edwin Boring in 1917 to Gerd Gigerenzer et al. in 1989 by and large agree with us---they preceded us—they are eminent in psychology--and, as Monica shows—Meehl and Cohen in particular are highly quotable as each wrote with immense wit and wisdom ("$p < .01$," in echo of Cohen's joking title of one of his dead serious articles, "The earth is round" ($p < .05$)"!)

Our originality comes in four ways, maybe more: 1) My historical work on the Gosset-Fisher-Pearson-Pearson-Neyman et al. debates which is based on five years of archival research on original letters, manuscripts, notebooks and published material at numerous archives; notably, at Guinness Archives, Dublin and at the University College London Special Collections Library; (2) Z's and M's systematic study of how the cult operates in journals in most areas of research in

the life and human sciences; we show for the first time comparative results from empirical studies (including our own) of thousands and thousands of articles, in fields from the economics of insect studies to breast cancer epidemiology, 1885 to the present (cf. Meehl and Cohen, whose brilliant work stopped at psychology and medicine with maybe a few forays into education---for the mid and late 20th century). We take on every field, including the law; (3) We give a sociological explanation for the rise of the cult and the decline of the economic approach to the logic of uncertainty; (4) We emphasize the cost of observations (how big must *n* be such that the odds of seeing X in range Y are 8 to 1 or better? 19 to 1 or better? How might a significance tester place a bet at Keeneland??) and (5) We give lots of jokes , as Fisher did not . . . but to hear them you have to buy the book!

Anyway, back to Oomph. What follows are chunks of chapters from our book on the significance problem in psychology and related fields. Eight or nine of every ten articles neglects oomph. That's our finding. The University of Kentucky has a chance to make a positive difference. I suspect you will!

Sincerely,

Steve Ziliak
Professor of Economics
Roosevelt University
http://faculty.roosevelt.edu/Ziliak
email: sziliak@roosevelt.edu

* * *

Have a look for example at a study from 2003, funded by the CDC, on the social psychology of recreational drug use. You can see immediately that the study is oomphless. Count the instances of oomphless claims:

> When examined in bivariable analyses, 15 of the 16 temptations-to-use drugs items were found to be associated [that is, statistically significantly related] with actual drug use. These were: while with friends at a party ($p < .001$), while talking and relaxing ($p < .001$), while with a partner or close friend who is using drugs ($p < .001$), while hanging around the neighborhood ($p < .001$), when happy and celebrating ($p < .001$), when seeing someone using and enjoying drugs ($p < .05$), when waking up and facing a tough day ($p < .001$), when extremely anxious and stressed ($p < .001$), when bored ($p < .001$), when frustrated because

> things are not going one's way ($p < .001$), when there are arguments in one's family ($p < .05$), when in a place where everyone is using drugs ($p < .001$), when one lets down concerns about one's health ($p < .05$), when really missing the drug habit and everything that goes with it ($p < .010$), and while experiencing withdrawal symptoms ($p < .01$) . . . . The only item that was not associated with the amount of drugs women used was "when one realized that stopping drugs was extremely difficult"
>
> *Journal of Drug Issues* 2003, pp. 171-172.

We count 16 instances of precision-only considerations in this one paragraph—and zero instances of oomph. It yields an oomph ratio of 0 percent. The authors believe they are doing serious scientific research, and we suppose that in many ways they are. We find it strange that they used "bivariable" instead of multiple regression techniques. But, still, their research was funded by a grant from the Centers for Disease Control, with implications for the War on Drugs, and surely that certifies their science as serious. Yet their use of statistical significance is nonsense. The whole of their world is significant ($p < .05$). They are in the grips of their own addiction to recreational statistics.

So, yeah, as Monica says, the point is old – as old, we find, as Edgeworth (1885) and especially Gosset (1905)—who was first to raise a flag (and beer) in support of what de Finetti (1976) called "the economic approach to the logic of uncertainty." But the cult of statistical significance, divined by Fisher, is winning.

On Arny's point about low p-values being 'necessary' in some cases – I know about Berger's simulations on false positives but I suspect Arny is referring to studies I don't know about. Still I agree with Harold Jeffreys. Since the p-value asserts no standard for a minimum quantitative difference that would matter it is mostly useless. As Jeffreys put it in his beautiful *Theory of Probability* (1961):

> If P is small, that means that there have been unexpectedly large departures from prediction [under the null hypothesis]. But why should these be stated in terms of P? The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value [of the test statistic] is nearly always negligible. ***What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.*** This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law [or null hypothesis], not against

> it. The same applies to all the current significance tests based on P integrals
>
> Jeffreys 1961, p. 385; editorial insertions by Arnold Zellner 1984, Chap. 3, p. 288 and Ziliak and McCloskey 2008, p. 167; emphasis in original.

The CDC drug use study, pimpled with meaningless p values, is a good example: "a remarkable procedure," as Jeffreys put it. Our examples are not a biased selection of the worst. You can test this by taking down the journals yourself in fields like quantitative economics, epidemiology, animal science, education, social psychology and looking for the place—we call it the "crescendo" of the article—at which statistical significance is said to imply the result. Judging by the prestige of the journals and the seriousness of the articles in other ways, the articles we mention here attain the average standard for the leading contributors to the science.

In other words:

(1) The average sentence of the average article in the leading journals of the statistical sciences claims that size doesn't matter.

(2) To put it another way, the average statistical proposition fails to provide the oomph-relevant quantitative information on the phenomena it claims to study.

(3) The average statistical article in the leading journals of science commits the *fallacy of the transposed condition*, equating "the probability of the data, given the hypothesis" with "the probability of the hypothesis, given the data."

(4) To put it another way, false hypotheses are being accepted and true hypotheses are being rejected.

We find the results strange. The part of civilization claiming to set empirical standards for science and policy has decided to use illogical instruments and irrelevant empirical standards for science and policy. In journals such as *The New England Journal of Medicine*, *The Journal of Clinical Psychiatry*, *Annals of Internal Medicine, Animal Behaviour, Educational and Psychological Measurement, Epidemiology and Infection, Decision Sciences*, and the *American Economic Review* the oomph, the size of effects, the risks and kinds and sizes of loss, the personal and social meanings of relationships do not seem to matter much to the scientists. An arbitrary level of statistical significance is the only standard in force—regardless of size, of loss, of cost, of ethics. That is, regardless of oomph.

Economists, we have noted repeatedly, are not the only scientists to fall short of significance (see Steve Ziliak's faculty website at Roosevelt University for

additional articles and replies: http://faculty.roosevelt.edu/Ziliak; "The Standard Error of Regressions" [McCloskey and Ziliak 1996] and "Size Matters" [Ziliak and McCloskey 2004] are good places to start, followed by "Significance Redux" [Ziliak and McCloskey 2004].) Psychologists have done so for many decades now. An addiction to transforms of categorical data, a dependence on absolute criteria of Type I error, and a fetish for asterisk psychometrics have been bad for psychology, as similar mistakes have been for economics.

Since Boring warned in 1919 against mixing up statistical and substantive significance the quantitative psychologists have been told by their own people again and again about the sizeless stare. Still they yawn—such a *boring* point, ha, ha. Since 1962, when Jacob Cohen published his pioneering survey of statistical power in the field, psychologists have been shown in more than *thirty* additional studies that most of their estimates lack it (Rossi 1991). Between snores, few psychologists cared.

"Significance is Low on My Ordering"

Norman Bradburn, a psychologist and past-president of the National Opinion Research Center, a member and former chair of the Committee on National Statistics of the National Academy of Sciences, told a story about *p*-values in psychology at the Memorial Service for William Kruskal. Bradburn spoke of Kruskal's gentle demeanor. But "sometimes his irritation at some persistent misuse of statistics would boil over, . . . as with the author of an article that used *p*-values to assess the importance of differences" (Bradburn, 2005, p. 3). Bradburn himself was the author in question. "I'm sorry," wrote Kruskal, "that this ubiquitous practice received the accolade of use by you and your distinguished coauthors. I am thinking these days about the many senses in which relative importance gets considered. Of these senses, some seem reasonable and others not so. Statistical significance is low on my ordering. Do forgive my bluntness."

Despite impressive attempts by such insiders to effect editorial and other institutional change—impressive at any rate by the standards of an economics burdened with cynicism in its worldview—educators and psychologists have produced "significant" results in volume. Kruskal, though a past-president of the American Statistical Association and a consummate insider, was too gentle to stop it. "Do forgive my bluntness" was as forceful as he got. Neither could Paul Meehl stop or even much slow down the beat of the Five Percenters, though a famous academic psychologist. Meehl, by the way, was also a clinical psychologist, and was able to help the difficult Saul Bellow—which astonished Bellow himself.[i] Meehl was Bellow's model for Dr. Edvig, in *Herzog*: Edvig was "calm Protestant Nordic Anglo-Celtic." Changing the psychology of significance testing seems in psychology too much even for a calm Protestant Nordic Anglo-Celtic.

The Melton Manual

The history of the *Publication Manual of the American Psychological*

Association exhibits the problem. The *Manual* sets the editorial standards for over a thousand journals in psychology, education, and related disciplines, including forensics, social work, and parts of psychiatry. Its history gives a half-century of evidence that reform of statistical practice won't succeed if attempted by one science alone. It's embedded like a tax code in the bureaucracy of science. The failure of the Kruskals, Cohens, Meehls, and others contradicts our own optimistic hope that a change of editorial practices in the *American Economic Review* or the *Journal of Political Economy* would do the trick in economics. In psychology a large number of useful-sounding manifestos and rewritten editorial policies have not built a rhetoric or culture of size mattering.

In the 1952 first edition of the *Manual* the thinking was thoroughly pro-Fisher and anti-Gosset, obsessed with significance: "Extensive tables of non-significant results are seldom required," it says. "For example, if only 2 of 20 correlations are significantly different from zero, the two significant correlations may be mentioned in the text, and the rest dismissed with a few words"[ii] The *Manual* was conveying what Fisher and Hotelling and others, such as Klein in economics and A. W. Melton in psychology, were preaching at the time. In the second edition—twenty years on—the obsession became compulsion:

> Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return. Take what's coming to you, but no more
>
> APA 1974, p. 19; quoted in Gigerenzer 2004, p. 589.

Recent editions of the *Manual*—as both critics and defenders of the Establishment observe— do at last recommend that the authors report "effect size."[iii] The fifth edition, published in 2001, added exact levels of significance to analysis of variance. But as Gerd Gigerenzer, a leading student of such matters, observes, the *Manual* retained also the magical incantations of $p < .05$ and $p < .01$. Bruce Thompson, a voice for oomph in education, psychology, and medicine, commends the fifth edition for suggesting that confidence intervals are the "best reporting strategy."[iv] Yet, as Thompson and Gigerenzer and Fiona Fidler's team of researchers have noted, in Gigerenzer's words, "The [fifth] manual offers no explanation as to why both [confidence intervals for effect size and asterisk-superscripted *p*-values] are necessary . . . and what they mean" (Gigerenzer, p. 594). The *Manual* offers no explanation for the significance rituals—no justification, just a rule of washing one's hands of the matter if $p < .05$ or $t > 2.00$.

In psychology and related fields the reforms of the 1990s were nice sounding but in practice ineffectual. The 2001 edition of the *Manual* appears to reflect pressure exerted by editors and scientists intent on keeping their machine for article-producing well oiled. Some 23 journals in psychology and education now warn readers and authors against the sizeless stare.[v] It is about 2% of the journals. The other 98% are sizeless. Despite the oomph-admiring language in recent

editions of the *Manual*, published practice in the psychological fields is no better than in economics.

In 1950 A. W. Melton assumed editorship of the trend-setting *Journal of Experimental Psychology*. In 1962 Melton described what had been his policy for accepting manuscripts at the journal (Melton 1962, pp. 553-7). An article was unlikely to be published in his journal, Melton said, if it did not provide a test of significance and in particular if it did not show statistically significant results of the Fisher type. Significance at the 5% level was "barely acceptable"; significance at the 1% or "better" level was considered "highly acceptable," and definitely worthy of publication (p. 544). Melton justified the rule by claiming that it assured that "the results of the experiment would be repeatable under the conditions described." The statisticians Freedman, Pisani, and Purves have observed sarcastically that "many statisticians would advise Melton that there is a better way to make sure results are repeatable: namely, to insist that important experiments be replicated."[vi] The 5%/1% statistician ruled, and the scientific standard of replication fell away in psychology as it had in economics. Gigerenzer et al. notes that after Melton's editorship it became virtually impossible to publish articles on empirical psychology in any subfield without "highly" statistically significant results. Some parts of psychology were spared: literary and humanistic psychology, for example. But we do not regard this as good news. The quantitative parts of a science should not be notable mainly for their lack of common sense.

In a penetrating article of 1959, "Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa," the psychologist Thomas D. Sterling surveyed 362 articles published in four leading journals of psychology: *Experimental Psychology* (Melton's journal), *Comparative and Physical Psychology*, *Clinical Psychology*, and *Social Psychology* (Sterling 1959). Table 11.1 shows his results:

{***insert Table 11.2}

Everyone knows---but no one corrects their significance levels for it---that "significant" results are the only ones that see the printed page. The fact undermines the claim of significance, since the so-called random sample is selected out of the numerous samples collected exactly for statistical significance. People in various statistical sciences complained about the publication bias often in the 1950s and 1960s, as the Fisherian machinery took hold and as academic publishing expanded (in economics, for example, Tullock 1959). "The problem simply," Sterling explained, "is that a Type I error (rejecting the null hypothesis when it is true) has a fair opportunity to end up in print when the correct decision is the acceptance of H_0 . . . . The risk stated by the author cannot be accepted at its face value once the author's conclusions appear in print" (p. 34).

Sterling was therefore not surprised when he found that only 8 of 294 articles published in the journals and using a test of significance failed to reject the null. Nearly 80% of the papers relied on significance tests of the Fisherian type to

make a decision (286 of 362 published articles). And, though Sterling does not say so, every article using a test of significance—that is, those 80% of all the articles—employed Fisher's 5% philosophy exclusively. (Melton's stricter rule of 1% was adopted by some of the journals.) The result "shows that for psychological journals a policy exists under which the vast majority of published articles satisfy a minimum criterion of significance" (Sterling 1959, p. 31).

Sterling observed further that despite a rhetoric of validation through replication of experiments---to which Gosset gave much of his scientific life, by the way, quite unlike Fisher, who preferred to do more statistical calculations on existing data---not one of the 362 research articles was a replication of previously published research.[vii] From his data Sterling derived two propositions:

> A1: Experimental results will be printed with a greater probability if the relevant test of significance rejects H_0 for the major hypothesis with Pr $(E \mid H_0) \leq .05$ than if they fail to reject at that level.
>
> A2: The probability that an experimental design will be replicated becomes very small once such an experiment appears in print.
>
> Sterling 1959, p. 33.

He understated. Nearly certainly an experimental result that "fails to reject" will not be printed, and by A. W. Melton with probability 1.0. And why actually replicate when the logic of Fisherian procedures gives you a virtually replication without the bother and expense? Why not go ahead and use the alloys F1 and F2 in airplanes? After all, $p < .05$.

"A picture emerges," wrote Sterling with gentle irony, "for which the number of possible replications of a test between experimental variates is related inversely to the actual magnitude of the differences between their effects. The smaller the difference the larger may be the likelihood of repetition" (p. 33). Sterling concluded that "when a fixed level of significance is used as a critical criterion for selecting reports for dissemination in professional journals it may result in embarrassing and unanticipated results" (p. 31). In a recent study similar to Sterling's, Hubbard and Ryan (2000) found that in twelve APA-affiliated journals between 1955 and 1959 fully 86% of all empirical articles published had employed the 5% accept/reject ritual.[viii] Educational psychology and other subfields of education had meantime taken the same turn.[ix] They continue therefore to yield embarrassing and unanticipated---in plain words, wrong---results.

Some Psychologists Tried to Ban the Test

Joined by a few academic students of education, some psychologists, alarmed by the oil-slick of the standard error, tried to ban it outright. Startlingly, the American Psychological Association arranged in the 1990s symposia to discuss the banishment of statistical significance testing from psychology journals. In 1996 an APA Task Force on Statistical Inference was appointed to investigate the matter.

In 1997 *Psychological Science* published the proceedings of the first symposium. Some of the main critics of statistical significance, such as Jacob Cohen, Robert Rosenthal, Harold Wainer, and Bruce Thompson, served on its 12-member jury.

Such a selection would be highly unlikely in economics, where such committees become sites for exercise of power in aid of established ideas. A similar committee of the American Economic Association formed to investigate the over-formalization of graduate education in economics (for example the excessive econometrics, without training in other means of investigating economic phenomena) was torpedoed by some of the barons appointed to it. The Task Force of psychologists, in contrast, was not a whitewash.

It speaks well for the intellectual seriousness of psychology. McCloskey's colleague in psychology at UIC, Chris Fraley, gives a detailed graduate course on null hypothesis significance testing that would be very hard to match for statistical and philosophical sophistication in economics. A section of the reading list entitled "Instructor Bias" quotes Meehl: "Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." Says Fraley to his graduate students, "I echo Meehl's sentiment."

Nonetheless the Task Force decided in short order that they would *not* be recommending the banning of significance testing. Reasons varied. For some it was to maintain freedom of inquiry and of method in science. A fine idea, but where, one might ask, were those voices for freedom and for rational method when the 1%/5% rule was codified in the APA Manual, or when Melton was imposing his reign of 1% terror on *Experimental Psychology*? The Task Force did at least urge reporting effect size, and in a meaningful context: "Reporting and interpreting effect sizes in the context of previously reported effects is essential to good research."[x] One step forward.

No Change

Fiona Fidler is a researcher in the department of history and philosophy of science at the University of Melbourne. She and her coauthors have registered the significance error in psychology and medicine in the way we have for economics. They were not surprised that the recommendations of the Task Force have essentially led to "no change" in practice, remarking bitterly that "for a discipline that claims to be empirical, psychology has been strangely uninterested in evidence relating to statistical reform."[xi] In a 2004 article entitled, "Editors can lead researchers to confidence intervals, but they can't make them think," Fidler and another team of co-authors show that in psychology from the mid 1990s to the present only 38% "discussed clinical significance, distinguished from statistical significance" (p. 120). It was better, in other words, than the 1990s practice of the economists. But in light of the large investment made by the APA in changing the

rhetoric of significance in psychology, the payoff was slight. Sixty-two percent said size *doesn't* matter.

In a major survey in 2000 of psychology and education journals by Vacha-Haase, Nilsson, Reetz, Lance, and Thompson the picture is worse. "Effect sizes have been found to be reported in between roughly 10 percent . . . and 50 percent of articles . . . notwithstanding either historical admonitions or the 1994 *Manual*'s 'encouragement' [to report effect sizes]."[xii] The main exception would have been *Educational and Psychological Measurement*, edited by Bruce Thompson from 1995 to 2003. Thompson tried to attract articles to his journal that were devoted from start to finish to substantive significance. But he too saw little permanent progress. Recent issues of the journal, after Thompson, resemble on average the *American Economic Review* at its worst.

Typical is the experience of Philip Kendall in editing the *Journal of Consulting and Clinical Psychology*. In 1997 Kendall began to encourage authors to report on "clinical significance" and not merely the statistical significance of their results. In 1996 only about a third of the articles in the journal (59 in total) made *some* mention of clinical significance. The other two-thirds relied exclusively on statistical significance, similar to the *American Economic Review*. By 2000 and 2001 the situation had not much improved. Only 40% of the articles—4 percentage points more—drew a distinction between clinical and statistical significance (Fidler et al, p. 619).

Author Sloth

Around the same time an experiment in reporting strategy in the journal *Memory and Cognition* brought sad results. Despite the "requirement" by the its editor, Geoffrey Loftus, that authors use error bars showing the confidence around their point estimates, less than half actually did so. Loftus was willing to impose the burden of the "new" reporting on himself. It has been said that he computed confidence intervals for more than a hundred of the articles submitted to the journal during his editorship. Though authors were asked officially in the back matter of the journal to do the work, and sometimes again in correspondence or in phone communication with the editor, hundreds reverted in their articles to the null-testing ritual and the sizeless stare. Maybe they didn't know what a "confidence interval" is.

Chapter 12
Psychometrics Lacks Power

Professor Savin: "What do you *want*, students?"
Iowa Graduate Students: "Power!"
Professor Savin: "What do you *lack*, students?"
Iowa Graduate Students: "*Power*!"

Eugene Savin's econometrics course,
University of Iowa, 1993

The cost of the psychological addiction to statistical significance can be measured by the "power function." 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 the other hypothesis--some other effect size--is true. A power *function* graphs the probability of rejecting the null hypothesis as a function of various assumed-true effect sizes. Obviously 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.

Suppose a pill does in fact work to the patient's benefit. And suppose this efficacy is what the experiment reveals, though with sampling uncertainty. What you want to know—and are able in almost any testing situation to discover—is with how much power you can reject the null of "*no* efficacy" when the pill (or whatever it is you are studying) is in truth efficacious to such-and-such a 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.

There are reasons to quibble about this notion of power, as descended intuitively from Gosset and formally from Neyman and Pearson. Sophisticates in the foundations of probability such as Savage and now Edward Leamer at UCLA have complained about its alleged objective certitude. Said Leamer to a 2004 assembly of economists, "hypotheses and models are neither true nor false; they are useful or not, depending." But this and other sophisticated complaints aside, power is considered by most statisticians—including Savage and Leamer—to provide a useful protection against unexamined null-hypothesis testing.

Power is so to speak "powerful" because hypotheses are plural and the plurality of hypotheses yield overlapping probability distributions. In a random sample the sleeping pill Napper may on average induce 3 extra hours of sleep, plus or minus 3. But in another sample the same scientist may find that the same sleeping pill, Napper, 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. "With what probability can I expect to get the additional 2 or 3 hours of rest?" she reasonably wants to know. "And with what probability might I actually get *less* rest?"

Without a calculation of power, to be provided by the psychometricians, she can't say. Calculators of Type I error pretend otherwise: following the practice of R. A. Fisher, they act as if the null hypothesis of "no, zero, *nada* additional rest" is the only hypothesis that is worthy of probabilistic assessment. They ignore the other hypotheses. They tell the business traveler and other patients: "Pill Napper is

statistically significantly different from zero at the 5% level." To which their better judgment—their Gosset judgment—should say, "So What?"

Power is mathematically speaking a number between 0 and 1. It is the difference between 1.0 (an extremely high amount of power, a good thing) and the probability of an Error of the Second Kind (a bad thing). The error of the second kind is the error of accepting the null hypothesis of (say) zero effect when the null is in fact false, that is, when (say) such-and-such a positive effect is true. Typically the power of psychological research is called "high" if it attains a level of .85 or better. (This too is arbitrary, of course. A full treatment with a loss function would not entail such rules of thumb.) High power is one element of a good rejection. If the power of a test is low, say ,.33, then the scientist will a third of the time accept the null and mistakenly conclude that another hypothesis is false. If on the other hand the power of a test is high, say, .85 or higher, then the scientist can be reasonably confident that at minimum the null hypothesis (of, again, zero effect, if that is the null chosen) is false, and that therefore his rejection of it is correct.

If the "null" is "no efficacy at all, when I would rather find a positive effect of my beautiful sleeping-pill theory," too-often-rejecting-the-null without consideration of the plurality of alternatives is the same thing as doing bad science and giving bad advice. It is the history of Fisher significance testing. One erects little "significance" hurdles, six inches tall, and makes a great show of leaping over them, concluding from a test of statistical significance that the data are "consistent with" ones own very charming hypothesis.

A good and sensible rejection of the null is, among other things, a rejection *with high power*. If a test does a good job of uncovering efficacy, then the test has high power, and the hurdles are high, not low. The skeptic—the student of R. A. Fisher—then is properly silenced. The proper skeptic is a useful fellow to have around, as Descartes observed. But in the Fisherian way of testing a null as if absolutely, by the 5% criterion, the skepticism is often enough turned on its head. It is in fact gullibility posturing as skepticism. That is, in denying the plurality of overlapping hypotheses, the Fisherian tester asks very little of the data. She sees the world through the lens of one hypothesis—the null.

To put it another way, power puts a check on the naivety of the gullible. He too, a faithful fellow, can be useful, as Cardinal Newman observed. But the failure to detect a significant difference between two sleeping pills, say, Somnus and its market competitor, Mors, does not mean that a difference is not there in God's eyes. (Mors sometimes puts you to sleep for eternity.) A Fisher test of significance asks what the probability is of claiming a result when it is *not* really there, that is, when the null hypothesis is true: no efficacy. Power protects against undue gullibility, then, an excess of faith. It is of course a legitimate worry.

Gosset discovered the legitimate worry, we have said, in his letter of May 1926, pointing out to Egon Pearson that the significance level trades off against power, still to be named. The confidence we place in Student's *t* depends, Student said, other things equal, on the probability of one or more relevant "alternative

hypotheses" perhaps more true. Naively accepting the singular null hypothesis involves a loss—"but *how much do we lose*?".[xiii] In 1928, and then more formally in 1933, Neyman and Pearson famously operationalized Gosset's improvement of his test.

Yet power is usually ignored in psychometric practice. It is wrong to be too gullible, granted. But it is also wrong to be too skeptical. If you protect yourself from gullibility in thinking a cancer cure is efficacious, you will avoid the embarrassment and cost of recommending peach pits when they don't work. But if you don't *also* protect yourself from excessive skepticism, by getting sufficient power, you will *not* avoid the other cost—of dead patients who might have been saved by a pill that does work. You will have set the hurdles too high rather than too low.

Setting the height of the statistical hurdles involves a scarcity, just as the setting of real hurdles does. Holding sample size constant, seeking low (mistaken) skepticism—high statistical significance—has the inevitable opportunity cost of higher (mistaken) gullibility. For a given sample size, power is a declining function of significance levels (Figure 12.1). This makes sense: the more area under the bell curve you want to yield to your null experiment (making rejection of the null more difficult by lowering the level of Type I α-error), the more you encroach into the probability distributions—the bell curves—of adjoining hypotheses.

***{insert Figure 12.1}

But high power is no perma-shield against other kinds of oomph-ignoring errors rife in the statistical sciences. 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 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 though the power of the test is. "A sample size of 2780 patients per treatment group," the authors of the infamous study said, "was expected to provide 90% power to detect a difference of 2 percentage points between treatments for the priMonica safety variable" (Lisse et al., p. 541). But as we have seen 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 a 8-to-1 cardiac damage or death ratio, a magnitude or "safety variable" of some importance.

How to Get Powerful

Mosteller and Bush (1954) seem to be 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 of it in psychology proper (1962). Cohen surveyed all 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.

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 you should be uncovering. (In the face of such language one can sympathize with the humble pragmatism of a Leamer or a Savage.) 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, or the power, of an effect is not itself much of an accomplishment. If you are a Fisherian, the fact of a large sample becomes your problem. You're deluded, thinking you've proved oomph before you've considered what it is.

In psychology, Cohen noted, as often and more alarmingly in medicine, few in 1960 reported the effect size they had found. A reader could not therefore, even if she had wanted to, estimate the power of their tests against the alternative effect sizes. Power estimation requires effect sizes. So in his large-scale survey of power Cohen had to *stipulate* the effect sizes, assigning what seemed to him small, medium, and large magnitudes for each case. It was quite a task. Having done a little of this sort of thing ourselves (and on desktop PCs in 2005, not on old Frieden mechanical calculators in 1962) we stand amazed at his scientific energy. For articles using t tests Cohen assigned .25, .50, and 1.00 standard deviation units to stand for small, medium, and large effect. For articles using Pearson correlation coefficients he used .20, .40, and .50 standard deviation units to stand for small, medium, and large effect.

Cohen's standard, alas, is a merely statistical one. On such heterogeneous subjects as one finds in the *Journal of Abnormal and Social Psychology*—from the relation of a medical treatment to paranoid schizophrenia to the relation of mother and son to sexual fetishism—a different investigator might well have divided up the regions of effect size in a different way. Cohen would have had to be expert in every sub-literature in psychology to judge for each the relevant standard of large and small effect. For some phenomena a 0.20 standard deviation change may produce a *large* clinically or ethically important effect on the dependent variable, such as anxiety or crime rate reduction. Cohen himself, fully aware of the issue, suggested in 1969 a downward revision of his 1962 effect sizes (Cohen 1969).

Still, Cohen's standard of effect size is a good deal better than nothing, with the advantage of being easily replicable. And for our current point it suffices: Cohen established a measure of the largeness of effect, which allows calculations of power. The authors of the original articles did not. As Gosset told Karl Pearson long ago, each investigator has to answer the question of what he means by an importantly large effect. The question must be answered in a scientific study, somehow.

Cohen's three assumptions about effect size gave him three levels of power for each of the 2,088 tests of significance in the 70 articles—notice that even in 1960, long before electronic computers, the average article in the *Journal of Abnormal and*

Social Psychology was exhibiting 30 significance tests. Thirty tests per article. The price per test dramatically fell in the next few decades, and as an economist would expect the number of tests per article correspondingly ballooned into the hundreds.

From the large-scale survey Cohen reckoned that the power in detecting "large" effects was about .83. So the probability of mistakenly rejecting a treatment having a "large" effect is of course 1.00 minus the .83 power, or 17 percent. That seems satisfactory, at any rate for a moderate loss from excessive skepticism. On the other hand, if you were dying of cancer you might not view a 17 percent chance of dying needlessly as "satisfactory," not at all. You might well opt for peach pits. It always depends on the loss, measured in side effects, treatment cost, death-rates. The loss to a cool, scientific, impartial spectator will not be the same as the loss to the patient in question. In 1933 Neyman and Pearson said, "how the balance should be struck" between Type I and Type II errors "must be left to the investigator"[xiv] That formulation is progress over the sizeless stare. But it would seem that a better formulation would be that it "must be left to the patient, friends, and family."

At smaller effect sizes, Cohen found, power faded fast. For the effects assumed to be in God's eyes "medium" or "small" the power Cohen found was derisory. It was, averaging one article with another, about .48 for medium effects and only .18 for small. That is, for a small, 0.25-standard-deviation-unit departures from the null hypothesis, the wrong decision was made 1.00 minus 0.18, or *92 percent of the time*. Cohen in 1969 re-did the power calculations at lower effect sizes and got about the same results.

The result was the same in similarly large-scale studies conducted by Sterling (1959), Kruskal (1968), and Gigerenzer *et al*. (1989). In fact dozens of additional surveys of power in psychology have been performed on the model of Cohen's original article. Rossi summarizes the findings: "The average statistical power for all 25 power surveys [including Cohen's] was .26 for small effects, .64 for medium effects, and .85 for large effects and was based on 40,000 statistical tests published in over 1,500 journals" (Rossi, p. 647). For example, Sedlmeir and Gigerenzer (1989) surveyed the power of research reported in the *Journal of Abnormal Psychology*. Using Cohen's original definitions of effect size, they found mean power values of .21, .50, and .84—in other words, nearly the same as Cohen found for small, medium, and large effect sizes decades earlier.

The Power of Rossi

A power study by Joseph Rossi (1990) was crushingly persuasive. Rossi calculated power for an astonishing 6,155 statistical tests in 221 articles. The articles had been published in the year 1982 in three psychology journals, the *Journal of Abnormal Psychology*, *Journal of Consulting and Clinical Psychology*, and *Journal of Personality and Social Psychology*. We again stand in awe: would that critics of the idiocy of null-hypothesis significance testing in economics—not excepting Ziliak and McCloskey—had such scientific energy. Using Cohen's effect sizes, Rossi found power to detect large, medium, and small effects of .83, .57, and .17. He

calculated power for 1,289 tests in the *Journal of Abnormal Psychology*, 2,231 in the *Journal of Consulting and Clinical Psychology*, and 2,635 in the *Journal of Personality and Social Psychology*. The conclusion: "20 years after Cohen conducted the first power survey, the power of psychological research is still low" (p. 646). And 20 years after Rossi it is still equally low.

Usually, as we have seen, the statistical test is not of an efficacy of treatment so much as *in*efficacy, that is, a null of No Effect, from which the psychologist wants to conclude that there *is* an effect. Either way, low power is a scientific mistake. As Rossi writes, "if power was low, then it is reasonable to suggest that, a priori, there was not a fair chance of rejecting the null hypothesis and that the failure to reject the null should not weigh so heavily against the alternative hypothesis." That's to put it mildly: the six-inch hurdles are lined up and the scientist courageously leaps over them. The scandal of low power in the social sciences should bring their practitioners to some humility. Yet Fisherian testers are *very* proud of their rejections of the null, and very willing to impose conformity in leaping stylishly over them. By contrast, Rossi recommends Gosset-like expressions of "probable upper bounds on effect sizes." We would add only that "probable *lower* bounds on effect sizes" are also needed (cf. Würtz 2003).

A "real" index of Type I error might help: scientists can express "real" Type I error as the ratio of the *p*-value to the power of the test:[xv]

real Type I error = empirical *p*-value / empirical power of the test

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. Recall in the power studies how often this was indeed the power for small effect sizes. 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.

According to Rossi, the real rate of false rejections in psychology is grim: "More than 90% [of over 6,000] of the surveyed studies had less than one chance in three of detecting a small effect"—very far above Fisher's 5% error claimed. Psychologists need to know that the real rate of false rejection is for small effect sizes at best .05/.17, or about 29%, and for medium-sized effects .05/.57, or about 9%. The same is true of economics and its imitative younger brother, political science. In other words, a Five Percent significance test is actually a vaguer Nine or Twenty-nine Percent significance test. So much for precision.

The problem of making the dull tool appear sharp is rife. Robert Shiller, a leading financial economist, wrote in his "The Volatility of Stock Market Prices" in *Science* (1987), "the widespread impression that there is strong evidence for market efficiency may be due just to a lack of appreciation of the low power of many statistical tests." Can other scientists claim to know as much about the statistical power of their field as the psychologists do? Very few. In economics we think of Zellner, Horowitz, Eugene Savin, and Allan Würtz. In heart and cancer medical science, Jennie Freiman, et al. know their power. A few biologists and ecologists

can claim to know their power, too, for example, Anderson et al. (2000). Most statistical scientists do not.

Designing experiments to find the maximal and minimal effect size is a better way to get powerful results and to keep the focus where it should be: on effect size itself. As Gosset argued back at University College, so Rossi says:

> Increasing the magnitude of effects may be the only practical alternative to expensive increases in sample size as a means for increasing the statistical power of psychological research. We tend to think of effect size (when we think of it at all) as a fixed and immutable quantity that we attempt to detect. It may be more useful to think of effect size as a manipulable parameter that can, in a sense, be made larger through greater measurement accuracy. This can be done through the use of more effective measurement models, more sensitive research designs, and more powerful statistical techniques. Examples might include more reliable psychometric tests; better control of extraneous sources of variance through the use of blocking, covariates, factorial designs, and repeated measurement designs; and, in general, through the use of any procedures that effectively reduce the "noise" in the system
>
> Rossi 1990, p. 654.

Those sound like good ideas for a science.

Chapter 13
The Psychology of Psychological Significance Testing

> A significance test is more likely to suggest a difference than is Jeffreys' [1939] method. This may partly account for the popularity of tests with scientists, since they often want to demonstrate differences. It would be interesting to know how many significant results correspond to real differences. It is also interesting that many experimentalists, when asked what 5% significance means, often say that the probability of the null hypothesis is 0.05. This is not true, save exceptionally. In saying this they are thinking like Jeffreys but not acting like him
>
> D. V. Lindley 1991, p. 12.

Why have psychologists been unwilling to listen? One reason seems to be insecurity in a so-called "soft" or "subjective" field. Recall even the learned Paul Meehl, a scientist as well as a philosopher, speaking of his own field as “soft.” The “hard/soft” dichotomy is surely a poor one for any science. It does not acknowledge the hardness of Greek contrary-to-fact conditionals in a "soft" field like classics or the softness of linked index numbers in a "hard" field like economics. Like soft and hard, subjective and objective are laymen's not philosopher's terms, and no better for it. The metaphysics and epistemology implied are dubious. Deciding what is hard and what soft, objective and subject, in a chain-weighted

adjustment of the GDP Price Deflator is an irrelevant diversion from the central scientific question: what in the current state of the science persuades?

One can see in the dichotomy of hard and soft a gendered worry, too. The worry may induce some men to cling to Significance Only. Barbara Laslett, for example, has written persuasively of the masculine appeal of quantification in American sociology after 1920 (Laslett 2005). By the early 1920s the percentage of articles published in sociology journals and employing statistical methods exceeded 30% (D. Ross 1991, p. 429). "Statistical methods are essential to social studies," Fisher wrote in the first edition of *Statistical Methods for Research Workers* (1925). "It is principally by the aid of such methods that these studies may be raised to the rank of sciences" (p. 2). Hardboiled-dom was the rule these sciences used to raise themselves to a 5% science.

Around 1950, at the peak of gender anxiety among middle-class men in the United States, nothing could be worse than to call a man "soft." "For this"—the era of war of depression, observed Saul Bellow—"is an era of hardboiled-dom," and was, too, the era of Fisher and Yates, Hotelling and McNemar. The "code . . . of the tough boy—an American inheritance, I believe, " said Bellow, "from the English gentleman—that curious mixture of striving, asceticism, and rigor, the origins of which some trace back to Alexander the Great—is [in the 1940s] stronger than ever. . . They [for instance, the Fisherian hardboiled] are unpracticed in introspection, and therefore badly equipped to deal with opponents whom they cannot shoot like big game or outdo in daring" (Bellow 1944, p. 9).

The Hardness of the Soft Sciences

Psychology is anyway nothing like as "soft" as is sometimes believed. Psychologists have long employed macho statistics, and in the beginning at a level akin to the commonly used techniques in English biometrics. In Germany the experimental psychologists began to use statistics as early as the middle of the nineteenth century.[xvi] Wilhelm Wundt and especially Gustav Fechner, the father of "psychophysics," were the first to walk philosophy of mind down the wooden stairs of German metaphysics and into the counting room of empirical perception.

Wundt's and Fechner's laboratories worked on "applications" only. Theory would remain, to the new experimentalists, deterministic. Yet their demarcation of theory and practice—of ideas and applications—was friendly toward statistical testing and classical inference. "A frequency interpretation grounded this work," writes the historian of statistics David Howie, "since the categories of perception were defined on a scale marked by the ratio of an individual's repeated assessments of the physical stimulus under consideration."[xvii] To the experimentalists a statistic could measure the degree of the grounding. Their tools were in the 1870s understandably primitive, but in no sense "soft." Fechner himself employed means, standard deviations, and the occasional coefficient of variation.

By the early twentieth century Americans and Europeans alike began to learn, especially from the Gower Street statisticians, the ways and means of

hardboiled "testing." American psychologists, like many of America's human and life scientists, were especially open to it. Measurement felt hard. The biometric methods of Karl Pearson were introduced to psychologists by, it seems, the Columbia professors F. S. Chapin and Franklin Giddings. *Psychometrica* was founded by L. L. Thurstone, a University of Chicago "psychophysicist" as he called himself, who began in the late 1920s to use inferential techniques.[xviii] Fisher had given Thurstone and others the scientific legitimacy they sought. Significance testing and quantitative methods generally were promoted early and late by the Social Science Research Council. Chapin, Goldings, Thurstone, and others wanted to free themselves from what they took to be that mere "opinion and crankery" of non-quantified fields.[xix]

A wider cultural "trust in numbers" had triumphed, and the life and human sciences, including psychology, would trumpet their new trust.[xx] In 1910 the Flexner Report on medical education advocated a scientific medicine, and a monopoly of a small group of medical schools. In 1915 Flexner told a gathering of social workers that if they wanted a formula for professional success—standardization of procedures, consensus in decision-making, monopolization of goods and services, higher salaries—they should follow the model he had designed for medicine.[xxi] They did. Flexnerian professionalization—inspired by the positivism of Karl Pearson—is one reason that early 20th-century social workers, drawn initially to problems of human rights, ended in the service of bourgeois values and spying for the state A rather similar pattern is found in nursing, with a lag.

The 5% science was promoted by the new leaders of quantitative psychology and education. European humanists can score themselves by how many generations they are removed from Hegel—that is, in being taught by a teacher who was taught by a teacher who was taught by a teacher who was taught by Hegel at the University of Berlin. Likewise, statisticians can score themselves by how many generations they are from Fisher. Quinn McNemar, for example, of Stanford University, was an important teacher of psychologists, who had himself studied statistical methods at Stanford with Harold Hotelling, the chief American disciple of Fisher. Hotelling had worked directly with Fisher. McNemar then taught L. G. Humphreys, Allen Edwards, David Grant, and scores of others. As early as 1935 all graduate students in psychology at Stanford, following the model of Iowa State, were required to master Fisher's crowning achievement, analysis of variance. Already by 1950, Gigerenzer et al. reckon, about half of the leading departments of psychology offered training in Fisherian methods (1989, p. 207).

Even rebels against Fisher were close to him, starting with Gosset himself. Palmer Johnson of the University of Minnesota studied with Fisher in England, though he later had the bad taste to write articles with Fisher's erstwhile colleague Jerzy Neyman, whom Fisher had cast into outer darkness. George Snedecor, an agricultural scientist at Iowa State University at Ames, was a co-founder of the first department of statistics in the United States. His important book on *Statistical*

Methods was influenced directly by Fisher himself, who somewhat surprisingly was in the 1930s a visiting professor of statistics at Iowa State. One can think of the Iowa schools then as one thinks of Gower Street in the 1920s and 1930s—a crossroads of statistical methods. E. F. Lindquist of the University of Iowa, the American leader of standardized testing for educators, was deeply influenced by Snedecor. Lindquist invented the Iowa Test of Basic Skills for school children. He too spent time with the great man himself.

Some psychologists knew about the work of Neyman and Pearson, and some even about that of Harold Jeffreys. But textbook authors, editors, and teachers—inspirited by Fisher's promise of raising their fields to the level of science—helped Fisher win the day. Statistical education narrowed at the same time as it spread. Decision theory and inverse probability, and Gosset's views on substantive significance, alternative hypotheses, and power, were pushed aside. Bayes' Theorem was, and is, as well. Too introspective for the hardboiled.

And power was altogether too confusing to keep in mind. J. P. Guilford's influential *Fundamental Statistics in Psychology and Education* (1942) decided that power was, in his words, "too complicated to discuss."[xxii] In 2004 an influential textbook writer, a psychologist, told Gigerenzer that he regretted leaving out power. He noted that in the first edition of his successful textbook he had discussed Bayesian and decision-theoretic methods, including power. "Deep down," he confessed, "I am a Bayesian." But in the second and subsequent editions the notions of power and decision theory and the costs of decisions, both Bayesian and Neyman-Pearson, vanished. The "culprits," Gigerenzer believes, were "his fellow researchers, the university administration, and his publisher."[xxiii]

Fisher Denies the "Cost" of Observations

During the 1940s and '50s and '60s among a tiny group of sophisticates the Neyman-Pearson approach was the gold standard. But at lower levels of statistical education Bayes and Neyman-Pearson, as we have said, were seldom presented as alternative and competing approaches to Fisher. Gosset's economic approach to statistics, picked up later by Savage an nbd Zellner, was at mid-century invisible. Fisher realized that acknowledging power and loss functions would kill the unadorned significance testing he advocated, and fought to the end, and successfully, against them (Fisher 1955, 1956). "It is important that the scientific worker *introduces no cost functions for faulty decisions*," Fisher wrote in 1956, "as it is reasonable and often necessary to do with an Acceptance Procedure" in something so vulgar as manufacturing.

> To do so would be to imply that the purposes to which new knowledge was to be put were known and capable of evaluation. If, however, scientific findings are communicated for the enlightenment of other free minds, they may be put sooner or later to the service of a number of purposes, of which we can know nothing. The contribution to the

> improvement of Natural Knowledge, which research may accomplish, is disseminated in the hope and faith that, as more becomes known, or more surely known, a great variety of purposes by a great variety of men, and groups of men, will be facilitated. No one, happily, is in a position to censor these in advance. As workers in Science we aim, in fact, at methods of inference which shall be equally convincing to all freely reasoning minds, entirely independently of any intentions that might be furthered by utililizing the knowledge inferred.
>
> Fisher 1956, pp. 102-3.

Fearful of the growing attraction of decision theory among at least the mathematical sophisticates, Fisher tried to identify Deming-type and Neyman-Pearson-type decisions with, as Savage put it mockingly, "the slaves of Wall Street and the Kremlin."[xxiv] With Deming and Shewhart in mind, Fisher wrote in 1955: "In the U. S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions [by 'correct' he means of course his own sizeless stare at Type I error and his logically flawed *modus tollens*], with those aimed rather at, let us say, speeding production, or saving money" (Fisher 1955, p. 70). Thus "Wall Street." Notice the sneer of the new aristocracy of merit, as the clerisy fancied themselves. Bourgeois production and money-making, Fisher avers, are *not* the appropriate currencies of science.

And the Kremlin. Neyman was a Polish Catholic, but *raised in Russia*. A progressive, *he was active in the American civil rights movement*, and tried unsuccessfully for years to get his Berkeley colleagues to hire his friend the mathematician David Blackwell (b. 1919), whose chief defect as a mathematician was the color of his skin.[xxv] Ah-hah. (In 1949 Blackwell published an anti-Fisher article with Arrow and Girshick, on optimal stopping rules in sequential sampling, a "loss function" idea inspired by Wald and Bayes.[xxvi] Blackwell credited Savage, with whom Blackwell had worked at the Institute for Advanced Study, for showing him the power of the Bayesian approach. "Jimmy convinced me that the Bayes approach is absolutely the right way to do statistical inference," he said.[xxvii])

Fisher viewed science as something quite distinct from "organized technology" and the questionable social purposes of the left wing:

> I am casting no contempt on acceptance procedures [he continues], and I am thankful, whenever I travel by air, that the high level of precision and reliability required can really be achieved by such means. But the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so wide that the analogy between them is not helpful, and the indentification of the two sorts of operation is decidedly misleading
>
> Fisher 1955, pp. 69-70.

To which Savage replied mildly: "in the view of a personalistic Bayesian like me, the contrast between behavior and inference is less vivid than in other views. For in this view, all uncertainties are measured by means of probabilities, and these

probabilities, together with utilities [or natural selection or whatever your currency is measured by], guide economic behavior, but the probability of an event for a person (in this theory) does not depend on the economic [or psychological] opportunities of the person" (Savage 1971a, p. 465). And yet, "almost in the same breath with criticism of . . . decision functions, Fisher warns that if his [5%] methods are ignored and their methods used a lot of guided missiles and other valuable things will come to grief" (Fisher 1958, p. 274; Savage, p. 465).

Textbook writers have concocted since the 1950s a hodgepodge of Fisher and his enemies. Early on in an elementary statistics or a psychometrics or an econometrics book there might appear a loss function—what if it rains on the company picnic? But it disappears when the book gets down to producing a formula for science. In the more advanced texts the discussion of power and decision theory comes late. In both cases it is marginalized. The names in the competition have long since dropped out: Gosset, Fisher, Neyman, Egon Pearson, Jeffreys, Wald, de Finetti, Savage, Blackwell, Lehmann, Good, Lindley. The hodgepodge, as Gigerenzer et al. (1989) have noted, is introduced anonymously, as though there was only one way of testing and estimation, and Fisher's is it (compare Efron 1996). There is no god of profitable scientific action, but Ronald Fisher is his prophet.

A philosopher of science, Deborah Mayo (1999), has recently entered the debate in favor of a hybrid of Neyman-Pearson decisions and Fisher's rule of significance for accepting an experimental result. Her *Error and the Growth of Experimental Knowledge* attempts to steer scientists toward a more systematic analysis of "errors," a Popperian direction we salute. Like Gosset (of whom Mayo is apparently unaware) and the Bayesian Harold Jeffreys, she "came to see" how "statistical methods . . . enable us, quite literally, to learn from error."[xxviii]

But Mayo places too much faith in the ability of tests of statistical significance to guide error-based, experimental learning. Our point here is that such a guide is useless, and therefore will not direct science to the *right* correction of its errors. Central to her "radically different" (p. 13) notion is "The falsification of statistical claims . . . by standard [for her, the standard Neyman-Pearson-Fisher hybrid] statistical tests" (p. 13). A strength of her project is, she says, "fundamental use of this [Neyman-Pearson] approach, albeit reinterpreted, as well as of cognate methods [e.g. Fisherian tests]. My use of these methods," she says, "reflects their actual uses in science" (p. 16).[xxix] Alas. If one turns to Mayo's discussion on what constitutes a "severe test" of an experiment, one finds only sizeless propositions, with loss or error expressed in no currency beyond a scale-free probability. Her notion of a "severe test" would seem beside the point to her muse Egon Pearson, and especially to Egon's muse, William Sealy Gosset (Mayo, pp. 178-87, 460).

A better approach to error-based learning that keeps both statistical hypothesis testing and experimental control at its center, as Mayo desires, would put Gosset's notion of *net pecuniary value* in its center. A notion of a "severe test" without a notion of a loss function is a diversion from the main job of science, and

the cause, we have shown, of error. As Gosset said to Karl Pearson in 1905, "it would appear that in such work as ours the degree of certainty to be aimed at must depend on the *pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment."* If human life, not money, is what is at stake in the experiment, then loss of life should define the "error" function (cf. Freiman et al.). And so forth, across the sciences.

Considering the size of the intellectual investment, many psychologists worry that psychology has not learned much from its statistical revolution. We believe the same can be said of econometrics. In both places the problem can be put as the low statistical power of the tests, coming from the one-sided devotion to Fisherian procedures. As Eugene Savin asks of his students in econometrics at the University of Iowa: "What do you *want* students?" "Power!" That way of putting it makes consideration of power into a full solution. It is not, though the sign of change is right.

The big ideas of psychological theory, one can argue, have not been shaped by Fisherianism. As in economics, they come chiefly from non-statistical sources. "Wolfgang Kohler derived his Gestalt laws," Gigerenzer et al. note of the classic period of psychological theory, "Jean Piaget his theory of intellectual development, B. F. Skinner his principles of operant conditioning, and Sir Frederic Bartlett his theory of memory, all without a statistical method of inference."[xxx] William James, John Dewey and, to give a recent example, Howard Gardner, did not depend on Student's *t*. And yet IQ testers and the students of personality still base their claims to science on statistical measures of correlation and fit. Educational psychology was therefore particularly vulnerable. Its flagship journal, the *Journal of Educational Psychology*, has been saturated by significance testing of the Fisher type since the early 1940s.

I'm O.K., You're a Bayesian

There's a contradiction here, well known to Mayo and other philosophers of knowledge, and particularly evident in the literature concerning the psychology of learning. Anyone modeling human learning adopts something like a Bayesian framework. Cognitive psychologists and philosophers of mind ascribe a "testing propensity" to the human mind itself, claiming that it is Bayesian or decision-theoretic in nature. Piaget, for example, spoke of the child as if she were a "scientist." The child learns as scientists do, conditioning present actions on the most recent information. Max Frisch characterized humans as "gamers", *Homo ludens,* as the Dutch historian Jan Huizinga famously called us. It is an idea a Bayesian game theorist could readily agree with. By the 1960s the cognitive psychologists, operations researchers, and a few economists had begun to argue that "rational" economies were tacitly or explicitly Bayesian.[xxxi] Others believed the mind to be a "Neyman-Pearson detector" or "observer." Human action, they said, in

notable contrast with Fisherian docility, was necessary, and possible—independent of "belief." The mind's internal "detector" simply

> adopts the decision goal of maximizing the hit rate (i.e. the correct detection of a signal) for a given false alarm rate (i.e. the erroneous "detection" of a signal). This view of perceptual processing was in analogy to Neyman and Pearson's concept of "optimal tests," where the power (hit rate) of a test is maximized for a given type-I error (false alarm rate).
>
> Gigerenzer *et al*., p. 214

Human subjects, they said, know things already, prior to the experiment, of course. They learn from acting and by being, by learning and by doing, and update their beliefs about the world accordingly. Of course.

But---here's the contradiction---when such Bayesians or quasi-Bayesians went to do the statistical analysis on their experiments, they used Fisher tests of significance, ignoring the framework of beliefs and learning by doing. Some few adjusted psychology to Fisher, such as Harold H. Kelley (1967), proposing that "the mind attributes a cause to an effect in the same way as behavioral scientists do, namely by performing an analysis of variance and testing null hypotheses" (p. 214). Kelley's ANOVA model of causal attribution generated for twenty years a large amount of research in experimental social psychology.). One doubts it. But most were theoretical Bayesians and Fisherian testers. Gigerenzer et al. conclude that "the reasons for this double standard seem to be mainly rhetorical and institutional rather than logical" (p. 233).

But Fisherians Survive on the Illogic of Neopositivism

The rhetoric is that of neopositivism. One reason for the success of the Fisherian program against more logical alternatives, such as Bayesianism or Neyman-Pearson decision theory or Gosset-Savage economism, is that the Fisherian program emerged just as neo-positivism and then falsificationism emerged in the philosophy of science. It would have fallen flat in philosophically more subtle times, such as those of Mill's *System of Logic Ratiocinative and Inductive* (1843) or of Imre Lakatos' *Proofs and Refutations* (1963-64). No serious philosopher nowadays is a positivist, no serious philosopher of science a simple falsificationist. But the philosophical atmosphere 1922-1962 was perfect for the fastening of Fisher's grip on the sizeless sciences.

Fisher recommended that the investigator focus on falsifying the null. If you suppose that correlation of infant habituation to novel stimuli, measured by eye gaze, predicts childhood IQ "remarkably well (0.7, corrected for unreliability)," you "test" it by asking what is the probability of a result so high as 0.7 if the null were true, namely, if zero correlation were true (Gottfredson 1996, p. 21). If the probability is suitably small you have falsified the null hypothesis, and conclude triumphantly that the maintained hypothesis, namely, 0.7, must therefore be true.

This is the method recommended by Popperian philosophy of science since the 1930s, Popper's self-declared killing of the confirmationism of the Vienna logical positivism since the 1920s. Both versions of positivism---Popper modified it rather than killed it, in truth---hang on the hypothetico-deductive view

of science. The notion is that your theory, *T*, can be stated as axioms, *A*, which imply hypotheses, *H*, which in turn imply observations, *O*. The theory is tested, says the hypothetico-deductive philosopher, by measuring *O* and then working back up the chain of implication through *H* and *A* to the grand *T*.

In other words, science is supposed to be summarizable as a *logical* system—hence "logical" positivism and the hypothetico-*deductive* view. Suppose H_0, the hypothesis of no connection whatever between an infant's ability to shift her gaze and the IQ of the little girl she grows into, implies some observations in the world, *O*. Symbolically, $H_0 \Rightarrow O$. Then by what is known in logic as the *modus tollens,* it follows strictly that not-*O* => not-H_0. If ignition applied to gasoline implies combustion, then a *lack* of combustion strictly implies a lack of ignition. If independence between gaze shift and later IQ implies the presence of a low correlation between the two, then a *lack* of a low correlation strictly implies a lack of independence.

So if you observe a correlation between gaze shift and later IQ (of, say, 0.7) very unlikely to occur in a world in which the null is true, then you have observed the probabilistic version of not-*O*. Therefore by *modus tollens* you know that the null is (probabilistically) falsified, that is, probably not-H_0 is true. Consequently—*this step is not valid*—your alternative to the null is true. Consequently, *H* is true and so is *A* and so is *T*. Consequently, you can with assurance assign children to streams in school and university and life on the basis of their gaze shift as infants.

You Can Abandon Falsification and Still Be Scientific

Falsificationism has retained a grip on scientists with a little philosophical learning ever since it was first articulated by Karl Popper. The notion was not in English firmly tied to Popper's name until his Ph. D. dissertation, published in 1935 as *Logik der Forschung*, was translated into English as *The Logic of Scientific Discovery*, in 1959. Most scientists nowadays, if they have philosophical ideas, reckon that they are Popperians. No wonder, since Popper portrayed the scientist as a hero, courageously facing up to his own refutation within a system of strict logic.

Falsificationism and the hypothetico-deductive view and logical positivism, and therefore Fisherianism, however, have had a flaw all along. It is that they are illogical. The flaw in logic was pointed out as early as 1906 by Pierre Duhem (1861-1916), a French physicist, mathematician, and philosopher of science, and later rediscovered by Willard Quine, the American philosopher. Duhem and Quine note that no hypothesis works so simply as $H_0 \Rightarrow O$. On the contrary, scientific hypotheses are accompanied by side conditions making the observation possible: H_0 and H_1 and H_2 and . . . $H_i \Rightarrow O$. To believe a hypothesis that those specks changing places on successive nights are the moons of Jupiter you need to believe also that the telescope does what it purports to do in seeing into the celestial sphere (Feyerabend 1975). To test the bending of starlight by the Sun you need to believe that the instruments are correctly calibrated. Dennis Lindley, who was in a position to know, writes that Harold Jeffreys "did not like Popper's views and tried to

prevent his election to the Royal Society on the grounds that Popper could not do probability calculations correctly. Certainly Popper did make a serious error," Lindley says, "which illustrates how difficult probability calculations can be" (Lindley 1991, p. 13).

The quite obvious richness of hypotheses H_0 and H_1 and H_2 and . . . H_i, is the death knell of *modus tollens* and the simple hypothetico-deductive/Fisherian view of science. If one needs ignition *and* a supply of gasoline *and* a supply of oxygen, then a lack of combustion implies *either* a lack of ignition . . . *or* a lack of gasoline *or* a lack of oxygen *or* a lack of any number of other necessary conditions. So much for the simple "falsification" of H_0. In the case of a regression equation with many tacit variables (for IQ: nutrition, family, community, social class), or an experiment with many side conditions (no trucks rumbling in the street close to the laboratory, and so forth), the falsification of a hypothesis, H_0, implies *either* that the hypothesis is wrong . . . *or* that one of the other variables or side conditions H_1, H_2, H_3, and so forth has intervened. *The Bell Curve* by Herrnstein and Murray (1994) makes the usual Fisherian mistake: "Psychometrics approaches a table of correlations with one or another of its methods for factor analysis . . . If they test traits in common, they are correlated, and if not, not. Factor analysis tells how many different underlying factors are necessary to account for the observed correlations between them" (p. 581). No. As Lindley put it:

> It would be interesting to know how many significant results correspond to real differences [we reply: in Economics, less than 20%; in Medicine and Epidemiology, between 10% and 30%; in Psychology, less than 10%]. It is also interesting that many experimentalists, when asked what 5% significance means, often say that the probability of the null hypothesis is 0.05. This is not true, save exceptionally. In saying this they are thinking like Jeffreys [or Gosset] but not acting like him.
>
> Lindley 1991, p. 12.

Precisely. *Modus tollens,* therefore, cannot be how science actually works, as the sons and daughters of Thomas Kuhn have been noting for decades. The children of Kuhn do not deny that a claimed falsification in the right circumstances is often persuasive. It is sometimes a sweet and useful argument. But falsification is nothing like the whole of scientific rhetoric. Near enough, the hypothetico-deductive model has been falsified. The sociologists and historians of science note that actual controversies in science are usually about whether this or that H_1, H_2, H_3 have intervened in the so-called crucial experiment. They note that the laws of science are metaphors and stories, and are persuasive often for reasons other than their implied *O*s. And their *O*s are processed by actual scientists in ways that have more to do with Peirce's "abduction" than hypothetico-deductive deduction.

What is relevant here for the statistical case is that refutations of the null are trivially easy to achieve if power is low enough, or if the sample is large enough. The heroism of the Popperian tester of null hypotheses is not very impressive. Remember the six-inch hurdles. What falsificationism strictly speaking replaced was *confirmationism*, namely, that if you have a hypothesis H_0 which implies observations *O,* then if you observe *O* you can have more confidence in H_0. This is

called in logic the fallacy of affirming the consequent. Because it was fallacious in simple logic—a logic that purposely *ignored* alternative hypotheses, power, and prior and posterior probabilities—the sons of the logical positivists such as Popper and the followers of Fisher sought a firmer ground in *modus tollens* for their deductive characterization of science.

But the so-called fallacy of affirming the consequent may not be substantively significant in a science that is serious about decisions and belief. It is after all how real scientists—such as Gosset, a lifelong Bayesian, and Egon Pearson, a lifelong decision-theorist and a late-in-life sympathizer with neo-Bayesianism, and Richard Feynman, a lifelong physicist and advocate of neo-Bayesianism—think. In his astonishing *Subjective and Objective Bayesian Statistics* the statistician James Press reports Feynman's view. Said Feynman, "The Bayesian [Jeffreys] approach is now the preferred method of comparing scientific theories . . . to compare contending theories (in physics) one should use the Bayesian approach."[xxxii]

Gosset and Feynman combine the confirmation approach with another, more common one seen more often in bench scientists than in philosophers of science. It is what the philosopher Paul Feyerabend called "counterinduction": "A scientist who wishes to maximize the empirical content of the views he holds and who wants to understand them as clearly as he possibly can must therefore introduce other views; that is, he must adopt a *pluralist methodology*."[xxxiii] And in the statistical terms relevant here, confirmationism and counterinduction are precisely the alternative to hardboiled Fisherianism. Power, simulation, variety of experiments, actual replication, and exploratory data analysis leading to inter-ocular trauma from the effect of magnitudes are different modes of affirming the consequent, and are more generally a reasonable program of Gosset or Bayesian and Feynman confirmationism than is the dogma of Fisherian or Popperian falsificationism.

[i] J. Atlas 2000, pp. 263-6.

[ii] APA *Manual* 1952, p. 414; quoted in Fidler et al. 2004, p. 619.

[iii] APA *Manual* 1994, p. 18; Thompson 2004, p. 608; Fidler et al., 2004, p. 619.

[iv] APA *Manual* 2001, p. 22; Thompson 2004, p. 609.

[v] Hill and Thompson 2004, in Fidler 2004, p. 619.

[vi] Freedman, Pisani, and Purves 1978, p. A-23.

[vii] Fisher 1955; cf. Student 1927 and Gosset 1936; Ziliak 2005a.

[viii] Hubbard and Ryan 2000, Figure 1; in Fidler et al., 2004, p. 618.

[ix] Thompson 2002, 2004; Schmidt 1996.

[x] Wilkinson and APA Task Force 1999, p. 599.

[xi] Fidler et al., 2004, p. 119.

[xii] Vacha-Haase, et al., 2000, p. 419, italics deleted; in Fidler et al., p. 120.

[xiii] Letter of W. S. Gosset to E. S. Pearson, May 11, 1926, Letter #1, in Pearson Papers, Egon file, Green Box, UCL; emphasis in original; partially reprinted in E. S. Pearson 1939, p. 243.

[xiv] Neyman and Pearson 1933, p. 296

[xv] cf. Bakan 1966 and Tversky and Kahneman 1971.

[xvi] Boring 1929 [1957]; Gigerenzer *et al*. 1989, pp. 205-214; Howie 2002, pp. 193-195.

[xvii] Howie 2002 p. 194.

[xviii] D. C. Adkins 1968, p. 22, in Sills, ed., 1968.

[xix] Gigerenzer, et al., 1989, pp. 205-214; Howie 2002, p. 194.

[xx] Porter 1995, pp. 3-8; I. B. Cohen 2005 (posthumous).

[xxi] Lowe and Reid 1999, pp. 15-17.

[xxii] Guilford 1942, p. 217, in Gigerenzer et al., p. 208.

[xxiii] Gigerenzer 2004, pp. 587-588.

[xxiv] Fisher 1955, p. 70; Savage 1971a, p. 464.

[xxv] Reid 1978, pp. 182-3, 264-6.

[xxvi] Arrow, Blackwell, and Girshick 1949. Blackwell and Girshick went on to publish a well-regarded book on game theory, *The Theory of Games and Statistical Decisions* (Wiley & Sons, 1954). Girshick and Savage collaborated, too. See, for example, their "Bayes and minimax estimates for quadratic loss functions." Pp. 53-74 in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley: University of California Press, 1951).

[xxvii] Blackwell, n.d. (www.amstat.org/statisticians/about). The American Statistical Association has done an admiral job on its profiles of the life and works of the great statisticians. Note that a Fisher test is *not* being employed by the Association as a criterion of inclusion.

[xxviii] Mayo, xii; cf. Gosset 1938 (posthumous), Student 1927, Jeffreys 1939, pp. 8-9, 13-14.

[xxix] Cf. Lehmann 1993, which comes to similar conclusions as Mayo while telling a similarly erroneous history of modern significance testing.

[xxx] Gigerenzer et al. 1989, p. 206.

[xxxi] For example, Arrow 1962; Cyert and DeGroot 1987 and references.

[xxxii] R. P. Feynman, in S. J. Press 2003, p. 230; Zellner 2005a, pp. 7-8.

[xxxiii] Feyerabend 1975 (1993), pp. 20-1, italics his.