The Significance of Statistical Significance: Two Authors Restate An Incontrovertible Caution. Why a Book?

Richard O. Lempert
University of Michigan Law School, rlempert@umich.edu

THE SIGNIFICANCE OF STATISTICAL SIGNIFICANCE

Two Authors Restate
An Incontrovertible
Caution. Why a Book?

by: Richard Lempert

The Caution

The authors are: Stephen T. Ziliak and Deirdre N. McCloskey. Their book is: *THE CULT OF STATISTICAL SIGNIFICANCE How the Standard Error Costs Us Jobs, Justice and Lives.*[1] The caution: *Statistical significance is not the same as substantive significance. Statistically significant relationships may, and often do, tell us nothing that matters, while relationships that do not achieve conventional levels of statistical significance can be important, and we may neglect them at our peril.*

There is nothing new or controversial about this caution. Ziliak and McCloskey name many scholars, more than 100 I would guess, who have said as much. Moreover, alerts of this sort may be found in numerous statistical texts and on the pages of journals in many fields, including law, sociology, economics, political science, psychology, medicine, public health, and epidemiology. So why do the authors -- who in a number of published articles have made a point of making this point -- feel compelled to write a book that reiterates what

[1] University of Michigan Press (2008)

they have several times said and what anyone who has taken an elementary course in statistics should know? Much of their book seeks to answer this question.[2]

Evidence of a Problem

Their short answer is that even if everyone who studies statistics should have heard the message, "Significance (statistical) ≠ Significance (substantive)," the matter is so skimpily treated in most statistical tests, if it is treated at all, that many who study statistics do not get the message, and among those who do understand it, many neglect what they have learned in reporting research results. Moreover, editors of and reviewers for even the most prestigious journals routinely give a pass to articles that emphasize statistical significance over the strength of relationships, thus subordinating issues of practical and theoretical importance to the question of whether it is safe to say that relationships reliably exist.[3] As troublesome, if not more so, authors, editors and reviewers typically accept without question the dismissal of a relationship if the association between variables does not reach some conventional level of statistical significance -- .05, for example. Indeed these gatekeepers

[2] About a quarter of the text, from p.187 through p. 244, is something else. It is a fascinating history of the relations, contributions and views of many those who pioneered the development of statistical science, most of whose names linger on because of tests associated with them. (e.g. Karl Pearson, Egon Pearson, William Gosset ("Student"), Jerzey Neyman, Ronald Fisher, Harold Hotelling and G. Udny Yule.) Much of this is a partisan discussion of the relationship between, and relative contributions, of Fisher and Gosset, with Fisher being the villain not just in his relationship with his friend Gosset but also of the entire story Ziliak and McCloskey tell. Gosset is characterized as "the bee" because of his concern with the weight that ought to be accorded statistical results (think beta weights or "b"s in a regression) and Fisher is designated the wasp because of his apparently waspish personality. For me this historical discussion with its focus on personalities, contributions and relationships was the most interesting portion of the book, not just because of the inherent interest of even pallid celebrity gossip but also because much in this portion was new to me. However, bearing in mind the authors' admonition that in examining data we should focus on what matters most, I shall only reference these matters briefly in the remainder of this review.

[3] Ziliak and McCloskey know full well that a relationship, no matter how significant statistically, may be illusory, for with better data or a more adequately specified model an apparent relationship may disappear. But to better highlight the contrast between too prevalent practice and how they think statistical results should be presented they often assume significance tests tell us whether relationships exist and argue that rather than focusing on the "whether" question we should attend to how much a relationship is likely to matter to economic, health or other values.

sometimes insist that if a relationship is not significant, its non-existence is the only permissible conclusion.

The centerpiece of the authors' attempt to establish the need for their book is a summary of two previously published studies that examined articles published in the 1980s and 1990s in the *American Economics Review* (*AER*).[4] The *AER* is academic economics most prestigious journal, economics is the most quantitatively oriented of the social sciences, and academic economists are, on average, the social sciences' most "sophisticated" users of mathematics and statistics. Indeed, the most mathematically able graduate students in fields like sociology and political science are often advised to take the graduate econometrics sequence rather than their own field's statistical offerings. Thus, if there are any social scientists who should understand the difference between statistical and substantive significance and who should know that if the goal is to make sense of what is going on it is substantive significance that matters, it should be those economists who publish in their field's most prestigious journal.

Economists may do better than others in these respects. Without comparative data, which Ziliak and McCloskey do not provide, we cannot tell. But it is clear that Ziliak and McCloskey think that by their criteria serious problems exist. Not only did McCloskey and Ziliak find that in the 1980s economists publishing in the *AER* were often naïve, misguided

[4] The author's intent was to examine all *AER* articles in the 1980s and 1990s. However, in their study of *AER* articles published in the 1990s Ziliak and McCloskey overlooked a not insubstantial portion of what had appeared. They checked the omitted articles after learning of their mistake, and report in this book that including them changes nothing that matters.

or mistaken in their reliance on significance tests[5], but their article showing this, although it received considerable attention, seemed to have little practical impact.

Assessing articles with the aid of a checklist of best practices, McCloskey and Ziliak found, among other failings, that only 4.4% of 182 articles that appeared in the 1980s reported on the power of the tests they were using,[6] only about 30% considered more than statistical significance decisive in constructing an argument or distinguished between statistical and economic significance, and 20% did not discuss the size of their regression coefficients. In the articles published in the1990s, some coded errors were less frequent, while others were more common; overall the picture was pretty much unchanged.[7]

These conclusions depend, however, on accepting Ziliak and McClosky's judgment as to whether errors have been made, and I would be surprised if the check list they used does not substantially overstate the degree to which what they found or didn't find was problematic.[8] . In addition, the authors define as error choices that are not always erroneous.[9] Finally, I also question whether Ziliak and McCloskey were being fair to those

[5] *The Cult* p. 75 reproducing a table from Dierdre McCloskey and Stephen T. Ziliak, (1996) The Standard Error of Regressions, *Journal of Economic Literature,* 34 (March) 97-114

[6] Statistical significance, loosely speaking, gives the probability that an association as great or greater than the one found in the data would arise by chance. Hence, if one is testing a "null" or "no difference" hypothesis, it is the likelihood that one is mistakenly claiming a relationship exists in the data when an association is pure coincidence. The power of a test speaks to the likelihood of making the opposite mistake; that is, claiming there is no relationship between two variables beyond that attributable to chance when there is in fact a true (non-chance) relationship.

[7] *The Cult* p. 81, reproducing a table from Stephen T. Ziliak and Dierdre McCloskey (2004) Size Matters: The Standard Error of Regressions in the *American Economics Review, Journal of Socio-Economics* 33 (5) 527-46

[8] For example, depending on sample size and results it is often not necessary to report explicitly on the power of tests either because adequate power is obvious or because key null hypotheses are rejected. Yet the proportion of articles that they list as attending to power issues seems to have as a denominator all the *AER* articles in their sample

[9] For example, they disparage what they call "sign econometrics," that is noting the sign but not the size of coefficients. Yet if signs do not carry the analysis but are used as a check on the plausibility of a model, it may be good practice to discuss them. Consider a study reporting deterrent effects of executions and including control variables, like arrest rates and unemployment levels that are thought to relate directly to homicide rates and certainly should not diminish them. If these and similar variables are insignificant but their signs are unexpectedly negative, there is reason to suspect the study's data or the model specification and to be cautious about accepting any findings regarding the deterrent effect of executions. Unexpected directionality, even if not significantly different from zero,

whose work[10] they reviewed when they faulted 73% of the 1990s articles where the issue arose and 32% of the 1980s articles for choosing variables to include in regressions solely on the basis of statistical significance.[11] Those acquainted with articles specifically criticized could, no doubt, offer further objections.

These complaints are, however, mere quibbles. Whether or not one accepts Ziliak and McCloskey's scores and scorecard as presented, they are almost certainly correct that even in the academic world's most prestigious economics journal numerous articles rely too heavily on statistical significance and pay too little attention to economic significance. Focusing more on statistical than substantive significance not only leads readers to infer that variables that matter little are in fact important, but it can also lead readers to ignore the likely consequential impact of variables that are not significant at conventionally acceptable levels. Ziliak and McCloskey drive this point home in a chapter discussing several clinical trials that dismissed helpful drugs as ineffective because the low power of the trial designs meant that conventional levels of statistical significance were not attained; I will say more about this example later in the essay.

---

should not be ignored. I would be surprised if Ziliak and McCloskey, who make their sympathies with Bayesian approaches clear throughout, think otherwise.

[10] Ziliak and McCloskey do not just identify the prevalence of what they consider error in the aggregate but they identify by names those authors from the 1990s who score well or poorly on their checklist. *The Cult,* pp 90-91. I think their willingness to provide names adds value to their presentation, but it does make fairness to the authors whose work they criticize of considerable importance.

[11] There are good reasons for wanting to pare a regression model following a pre-test, exploratory or other preliminary analysis, including a desire to preserve degrees of freedom or to simplify output so as to better communicate findings. If there are no strong theoretical reasons to expect a variable to have a particular effect, if the variable's omission has little effect on the overall explained variance or on the coefficients on the remaining variables, and if it is statistically insignificant, it is a good candidate for omission. Omitting variables for these reasons is, of course, not to omit them *solely* on the basis of statistical significance, but a researcher who has carefully considered the status of a variable may nevertheless mention only statistical insignificance in justifying the omission, perhaps assuming that his readers will take his or her care in deciding on variables to exclude for granted. (Ziliak and McCloskey's criteria speak in terms of *including* variables solely on the basis of their statistical significance, but their discussion of what this means in practice indicates that they are thinking more of dropping variables on this basis.)

Overreliance on statistical significance is a concern not just because it leads to the acceptance of substantively unimportant results or to the dismissal of real effects because some arbitrary threshold of significance is not reached. It also means that information on what data show is not effectively communicated. As Ziliak and McCloskey repeatedly remind us, significance tests are designed to answer only the question of whether something matters. They neither tell us how much something matters nor specify a plausible range of mattering. Other ways of presenting data, like confidence intervals,[12] error bars[13] or the amount of variance a variable uniquely explains,[14] address these latter concerns, but often all

[12] A confidence interval may be thought of as indicating, with a specific probability, the range of values a variable may plausibly take in a population. More precisely if one were to take a series of random samples from a population and compute a confidence interval of N% for each sample, N% of the sample confidence intervals would encompass the true population parameter. Suppose, for example, we randomly sampled a group of starting large firm associates and asked men and women their salaries, finding that the men in our sample earned on average $4500 more than the women, and that the confidence interval at the 95% level ranged between $1100 and $7900. This indicates that if we could know the salaries of all starting large firm associates there is a 95% chance that the calculated mean difference in the starting salaries of all male and female associates would be between $1100 and $7900.

If one is testing a null hypothesis (e.g. there is no difference in the starting salaries paid male and female large firm associates) a confidence interval that does not encompass 0 is statistically significant at the level indicated by 1 – N% where N% is the confidence level. Thus in our starting salary example the $4500 difference that we found would be significant at the .05 level. If the 95% confidence interval had spanned a range of possibilities from women earning $100 more than men to men earning $9100 more than women, (-$100 - +$9100) the difference between the starting salaries of male and female associates would not be significant at the .05 level. Note, however, how much more informative the confidence level information is than a simple statement that the difference between the starting salaries of male and female associates is not statistically significant. If all we have is the latter information, we leave the study thinking that male and female associates are paid more or less the same. If we are given the confidence interval, we know that although there was no statistically significant difference between the starting salaries of male and female associates, more likely than not there is a difference, and it may well be so large that it is a matter of considerable concern. Thus the same data may be presented so as to suggest that any discrimination between men and women in large firms does not involve starting salaries and we should focus our research on other areas, or to suggest that there is a potentially serious problem of starting salary discrimination that cries out for further investigation.

[13] Error bars are a graphical way of presenting much the same information found in confidence intervals. Typically they extend one standard deviation above and below a sample mean, corresponding to the 95% confidence interval. If the error bars around the means for two samples overlap, then the difference between the means is not significant at the .05 level. If the error bars do not overlap, then the difference is statistically significant.

[14] In regression equations independent variable are often correlated and it may be impossible to untangle the relative importance of correlated variables. A stringent test of a variable's importance, which has been called "usefulness," (Darlington, Richard B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 71, 161-182.) is how much a model's explanatory power suffers if the variable is dropped from the model. Suppose for example, that following up on the findings hypothesized in note 15 above, we decided to use regression analysis to ascertain the determinants of large firm associate starting salaries and, in particular, whether associate gender had an important effect. We might have constructed a model that included such variables as law school attended, rank in class, law review membership, firm size, firm location, clerkship experience, age, marital status, firm billable hour

an author reports (I, like Ziliak and McCloskey, have been guilty of this) are tables of associations or point estimates with asterisks indicating whether certain "magical" significance levels (.1, .05, .01, .001) have been reached.

Although it is natural to conclude that a relationship significant at the .001 level is stronger and more important than one that is significant at the .05 or .1 level, this is simply not true. One reason is that for any given strength of effect, the larger the N the higher the significance level. With a large enough N virtually all associations in a sample will be statistically significant, for as size increases random effects are more likely to cancel out and even weak signals will emerge through the real world's noise.

Why Test for Significance?

The reader may have noted that I have gotten quite a way into this discussion without ever clearly saying what significance tests are good for. Ziliak and McCloskey do likewise. They do not deny the importance of appropriately used significance tests, but they are not as clear as they might be in explaining what the appropriate uses are. This may be because their primary imagined audience is economists, who presumably need no instruction on this score. This book may, however, be profitably read by statistical novices. The authors' prose is

---

norms, practice area and gender. With this information we could, no doubt, do quite well, in explaining starting salaries; let us assume our model explained 74% of the variance. Then suppose we reran the model but did not include our gender variable. If the model then explained 73% of the variance, the result would not necessarily exclude gender as an important cause of salary differences because other variables in the model might be proxying for it, but it would make the claim that gender was by itself not a very important factor in determining starting associate salaries far more plausible, and it would suggest that the $4500 starting salary difference in our sample was not the result of lower offers being extended to women because they were women. On the other hand, if eliminating gender from our model reduced the explained variance in starting salaries from 74% to 59%, the claim that women were paid less because of the types of jobs they chose or aspects of their resumes would be far less plausible, and the conclusion that large firms discriminated against women in setting their starting salaries would be substantially strengthened. These differences might well not be obvious, and we would have little or no idea of their magnitude, if all we asked was whether the coefficient on the gender variable in our model was statistically significant.

clear (more about their "poetry" later) and no technical knowledge is needed to understand the book's core points. Thus, a word about why social scientists do significance testing is in order.

Social science investigates relationships between variables. Sometimes the aim is simply description, but the more usual goal is to infer cause, which is to say finding associations that make causal claims more credible. To illustrate, imagine a study that seeks to determine whether preschool attendance affects the likelihood that children from disadvantaged backgrounds will stay in school through high school graduation and, if so, whether the gains are such as to make social investments in pre-schooling worthwhile. Suppose we were able to identify a group of disadvantaged students who participated in a well-designed pre-school program and a group without any pre-school experience and, following them through their educational careers, we found that 70% of the former graduated high school as compared to 40% of the latter. This difference makes the claim that pre-schooling has a positive effect on educational persistence appear more likely than it appeared before we examined the data.

Can we be completely confident that pre-schooling has such a large, positive effect? No. Many *rival hypotheses*; that is, reasons for our findings that have nothing to do with the efficacy of pre-school, might explain the data. [15] Perhaps those parents most interested in their children's education enrolled them in pre-school, and educational persistence reflects degrees of parental support rather than pre-schooling. Or maybe only English speaking

[15] My first published article, applying ideas from the writings of Donald T. Campbell and originally written for a seminar he taught, had to do with these issues. *See* Richard Lempert, Strategies of Research Design in The Legal Impact Study: The Control of Plausible Rival Hypotheses, 1 *Law & Society Review* 111 (Fall, 1966)

families learned of the pre-school opportunity, and English language ability explains why some students were more successful through high school than others. To strengthen our claim that pre-schooling is a key causal variable, we would want to rule out these and other rival hypotheses.[16]

One such rival hypothesis is "chance." It may be that neither pre-schooling nor parental support nor English language ability nor any other factor that might have systematically advantaged the pre-schooled group explains their greater success. Instead, by sheer happenstance more members of the schooled than unschooled group may have succeeded. By the luck of the draw students who had been pre-schooled may have been assigned to more capable teachers in kindergarten and first grade, or the pre-schooled girls as a group may have been no less sexually active than their unpre-schooled counterparts but more fortunate in escaping pregnancy.[17] Significance tests assess the plausibility of the rival hypothesis "chance" and nothing more. A significance level tells us the likelihood that we would find an association or difference between two variables as or more extreme than the one observed if the association or difference only existed by chance.[18] As Ziliak and

[16] For an outstanding, accessible treatment of the logic of social science inquiry see Arthur L. Stinchcombe, *Constructing Social Theories*, Harcourt, Brace and World, Inc. 1968 (Now available from the University of Chicago Press).

[17] Sometimes what appears to be random luck unrelated to the variable of interest is not. For example, pre-schooled youth might have more capable first grade teachers because the better teachers try to arrange things so that their classes will contain a high proportion of children who have been pre-schooled, and it may be that pre-school has its effect *only* because it channels its students to the better first grade teachers and these teachers in turn prepare their charges to do better second grade work, etc. Pre-school would still be a cause of the pre-schooled students' later success, but the reason it leads to success (*i.e.,* the *mechanism* by which it worked) would not be what most would suppose. This matters. If we knew the mechanism, we would not choose to invest more in pre-schooling but would instead spend our money to increase the quality of first grade teaching. Mechanisms are often acknowledged to be unknown or are assumed to be the most obvious ones in the circumstances. Understanding the mechanism by which independent variables have their effects can be as or more important than knowing whether an effect exists or how large it is.

[18] A different way of thinking about this is to imagine that we could do the same study time after time with repeated random samples from the same population. If we would seldom get results like those observed in the original study in subsequent studies, then our conclusion that we had found a real relationship is suspect. If we would often get

McCloskey point out, the greater number of other plausible rival hypotheses that exist and the more plausible they are, the less one gains from being able to rule out chance as an explanation for data. Nevertheless, the culture of social science calls for ruling out the chance explanation as a starting point for analysis.

Ruling out chance as an explanation is most valuable and most necessary when a study has been designed so that chance is the most plausible rival explanation. This is the case with studies that use random assignment to treatments or surveys that randomly sample members of a population. Suppose that to determine the effects of pre-school, we had first identified a large group of disadvantaged children and then randomly assigned some but not others to attend pre-schools. If the number of students in each group was moderately large, it would be only by extremely bad luck that on average the pre-schooled group would turn out to have had parents who were more concerned with their children's education than the parents of those in the unschooled group or that the pre-schooled children were better English speakers at the time assignments were made to treatment. [19] If the pre-schooled group did better at persisting to high school graduation than the group without pre-schooling, a significance test would tell us how likely it was that we had been unlucky; *i.e.*, that chance factors might have resulted in a relationship between pre-schooling and high school graduation at least as strong as the one we observed.

Purists have argued that the only appropriate use of significance tests is when there has been random sampling or random assignment to treatments, but most social scientists

---

results like those we observed, then the association is unlikely to be accidental. It could, of course, still exist for reasons other than the one we posit.

[19] In our imaginary experiment all those assigned to pre-schooling and only those assigned to pre-schooling would have a pre-school experience. In real world experiments of this sort matters get more complex as some parents invited to send their children to pre-school do not and some parents whose children are not in the pre-school group find pre-schooling for their children on their own. Similar problems occur in many field experiments.

believe that significance tests are properly employed whenever "*it happened by chance*" is a plausible rival hypothesis[20]. To illustrate, suppose we had not randomly assigned three and four year olds to pre-school, but instead had data on a group of 18 year olds, some but not all of whom had been pre-schooled. If we found that students with pre-schooling were more likely to have graduated high school than those without pre-schooling, we would then have to decide whether the pre-school experience figured in the pre-schooled sample's greater success. This task would be more difficult than it is in the randomized experiment because there would be more plausible rival hypotheses for us to dispose of.[21]

To dispose of these hypotheses, we would most likely construct a regression model that included not just pre-school attendance but also other variables in our data set that might plausibly explain school persistence, such as parents' education, birth order, or whether a student had made a sports team as a high school freshman. If, after controlling for these and other plausible causal variables we found, that pre-school attendance remained associated with educational persistence (i.e., had a positive beta coefficient in our regression equation) we would have greater confidence that pre-schooling had the effects we posit. Before proclaiming our greater confidence, however, we would want to be relatively certain that factors we could not control for (good or bad luck in getting pregnant after unprotected sex, for example) were not by chance associated with our pre-schooled sample in ways that might

[20] For discussion of this issue see Denton E. Morrison and Ramon E. Henkel, (1970) *The Significance Test Controversy: A Reader* Chicago: Aldine

[21] In either case it could be that the pre-schooled group contained children whose parents valued education more highly. In the version of the study we are now discussing this possibility would have to be specifically addressed. In a fully randomized experiment, this hypothesis would be assimilated into the rival hypothesis that the association between pre-schooling and high school graduation exists by chance, for it would only be by chance that those students who enjoyed pre-schooling initially had parents who were more concerned with their children's education than the parents of children in the control group. Randomization is valuable because it converts many otherwise plausible rival hypotheses into the single rival hypothesis of chance, and the likelihood of this hypothesis can be addressed by significance tests.

explain their greater apparent success. If the effect (beta coefficient) we found for pre-school attendance was statistically significant, meaning the coefficient differed from zero to a degree that was very unlikely if only chance mattered, we would have that assurance, or, more precisely, know the degree to which we might feel reassured.[22]

Typically significance tests measure deviation from a null (or no difference) hypothesis.[23] This means that researchers are usually testing hypotheses they don't believe and hope to dismiss. In our example the researcher is most likely looking at pre-school attendance because she believes it increases later learning, but she is positing for testing purposes that pre-school attendance has no (or a *null*) effect on school persistence and that any apparent association between the two variables is plausibly attributable to chance. If it turns out that significance tests justify rejection of the null hypothesis,[24] the researcher will have increased confidence that pre-schooling promotes educational persistence -- and she will also find it easier to get her results published.

The Dark Side of Significance Testing

Ziliak and McCloskey's complaints about significance testing are complaints about how they are used. Their primary complaint is that, although for most policy and theoretical purposes it is the *strength* of an association and not its existence that matters, often the only reported measures of association are significance tests which, without more, tell us nothing

[22] The reader may correctly intuit that a far greater threat to our confidence in claiming that pre-schooling promotes school persistence comes from absent variables that are systematically rather than randomly associated with pre-school attendance. I'll say a bit more about this below.

[23] Significance tests can also measure deviation from a value other than zero. A researcher may hypothesize that a population parameter has a particular value and use significance tests to answer the question of whether sample deviations from that specified value exceed what might be plausibly attributed to chance.

[24] Rejecting a null hypothesis is, by the most common social science convention, justified if an association as strong or stronger than that which is observed would, on average, occur no more than five times in 100 repeated random samples of size *n* drawn from a population, or, as it is more commonly phrased, has a no greater than 5% probability of appearing by chance. Depending on the issue as well as sample characteristics, the use of significance levels greater or less than .05 may, however, be justified.

about the likely importance of the association observed. Thus, in our example a significant relationship might exist if pre-schooling raised the apparent probability of graduating high school from 40% to 80%, but if the sample were large enough a significant relationship might also exist if experiencing pre-school raised the probability of graduation from 40% to 40.4%.[25] In the former instance major societal investments in pre-schooling might well be justified. But if the gain were only .4%, most would feel that if our goal was to increase educational persistence there would be far better uses for our money than funding pre-schools. Perhaps beguiled by the cult of significance, researchers too often simply tell us that a variable matters, and ignore the question of *how much* it matters (which Ziliak and McCloskey refer to as a variable's *oomph*)..[26] Ziliak and McCloskey's response to studies that emphasize significance and ignore oomph is "Who cares?" In their view neither the theorist nor policy maker should care about the "whether question" when it is separated from the question "How much?".

Another of Ziliak and McCloskey's complaints about significance tests is that they too often lead researchers and their readers to read too much into a study's results. Rejecting a null hypothesis is not the same as proving a favored one. Indeed, disposing of the null hypothesis may have almost no effect on the likelihood a favored hypothesis is true. In many studies the null hypothesis is only one of many rival hypotheses that might explain research results, and it is often not the most plausible competitor. Consider the variant of our pre-school study in which only retrospective assessment was possible. Unless we could control

[25] As Ziliak and McCloskey more than once point out and as I have already noted above, with large enough samples weak and unimportant relationships can be statistically significant.

[26] Ziliak and McCloskey also object to simple significance testing because it tells us nothing about the range of plausible effects. Hence they advocate the routine presentation of *confidence intervals*. See note 15, *supra.*

for the range of educationally relevant factors that might distinguish those children who had attended pre-schools from those who hadn't, knowing that the better performance of those pre-schooled was unlikely to be attributable to chance would do almost nothing to make more plausible our hypothesis that the pre-school experience was a crucial cause, nor would it tell us how important the pre-school experience might be.

Where inadequately controlled potential confounds are known and obvious, as in the preceding example, there may be little danger that too much will be read into a finding of significance or, indeed, that the work will even see print,[27] but where the need for or the inadequacy of controls is less obvious, studies are often published -- and occasionally become influential -- in part because researchers write as if disproving a null hypothesis means their favored hypothesis has been proven.[28]

Significance tests also do harm when they discourage authors from submitting work where the results are not statistically significant, or when they lead reviewers and editors to reject submissions because key results fall short of (conventionally defined) statistical significance. For starters there is what has been called the *file drawer problem*. If enough different researchers examine the same relationship in enough different samples, chances are that one researcher will draw a sample which suggests a significant association even if there is no relationship in the larger population, for by definition we can expect that one study in twenty will be significant at the .05 level when there is no relationship between the variables

[27] This is not to say that such work is never published. For example, until about 30 years ago the problem of sample selection bias was seldom recognized and even today there are often no adequate ways to account for it.

[28] I do not mean to suggest that studies that do not adequately control for all plausible rival hypotheses should not be published. If this were the rule, almost no social science research would be done. In the social sciences as in the other sciences knowledge generally accretes as studies build on each other and the strengths of one study compensate for the weaknesses of another. All I, and I presume Ziliak and McCloskey, are objecting to are those who write as if disproving a null hypothesis establishes the truth of the hypothesis they offer. At best, and then usually only in the context of other research, disproving a null makes a favored hypothesis substantially more likely.

studied. If the one study "with results" is published while the many studies that found no relationship rest unread in file drawers, consumers of the research will be misled by the published results.

An even greater threat to science-based understandings is the problem of low power. Particularly when samples are small, even strong relationships may not be statistically significant. This may lead researchers to report finding no relationship in the data when a relationship not only exists, but is also substantively important. Ziliak and McCloskey use data from a clinical trial of Vioxx as an example.[29] An important paper assessing this trial indicated that during the trial there were 5 deaths from heart attacks in the Vioxx group compared to 1 death in the control condition.[30] The sample size and the number of deaths were, however, such that the difference in heart attack deaths was too small to achieve statistical significance. This led Merck, the manufacturer of Vioxx, to claim that there was *no* difference in the adverse cardiovascular effects associated with Vioxx and naproxen (*e.g.*, Aleve) the drug taken by the control group. But a 5 to 1 difference in deaths is a difference that might concern us if we were prescribed Vioxx, regardless of whether the difference reached the .05 level of significance.

In the case of Vioxx the numbers of deaths (or one version of them) were published for all to see. But the cultural demand for significance means that many studies with low power will not be published, even some that suggest the possibility of reasonably strong relationships. This is particularly unfortunate today because advances in meta-analytic

[29] *The Cult*, Ch. 1.
[30] These are the figures given in the original report on the trial. Later it turned out that 8 people in the Vioxx condition had died, but three cases had been dropped from the data for reasons that are not clear. Ziliak and McCloskey suggest that this may have been done to get the number of Vioxx deaths below what was needed to make the fatality difference statistically significance. *Id* at 29.

techniques allow researchers to aggregate studies and find relationships that do not emerge in studies taken individually. If work is not published because results are not statistically significant, important information will often be unavailable for meta-analysis.

The Law Gets It (almost) Right

Although Ziliak and McCloskey refer in several places to the law as one of the fields affected by the problems that concern them, only once do they offer a legal example. When they do, they go wrong in an interesting way.

The authors criticize the United States Supreme Court for embracing in *Castenada v. Partida* [31] significance criteria as grounds for a deciding whether a jury was chosen in a discriminatory manner. But *Castenada* is that rare situation where a significance test may be appropriate grounds on which to rest a decision.[32] This is because juries are supposed to be chosen at random from identifiable populations, and significance tests measure the likelihood that deviations from an expected distribution might have occurred by chance.[33] Moreover, legal norms are violated by discrimination, however slight.[34] If *Castenada* can be faulted, it is more for suggesting that conventional measures of statistical significance are, *as a general matter,* appropriate measures of legal significance.[35]

---

[31] 430 U.S. 482 (1977)

[32] Ziliak and McCloskey seem confused about what the legal issue is. Their language suggests that they view the relevant null hypothesis as the defendant's innocence of a charged crime.

[33] For example, if black citizens constituted 20% of a county's voting age population, we would expect that unbiased selection from that population would yield a venire that was about 20% black. If only 10% of those called for jury duty were black, we would want to know how likely it was that random draws from the county's population would produce a venire that had 10% or fewer blacks. Significance tests could answer this question.

[34] The degree of violation may be relevant to harmless error analysis, but the first issue for a court to decide is *whether* a violation occurred.

[35] The Court may also be criticized for writing as if it didn't matter whether a court required a significance level corresponding to a difference of two or three standard deviations before deciding that a null hypothesis should be rejected. These levels differ in their stringency, with the former indicating a relationship significant at the .05 level

The Law Gets It (very) Wrong

Because Ziliak and McCloskey – with the exception of *Castenada* -- don't examine the use of significance tests in legal cases, the legal literature or the law and social science literature, we need to ask whether the issues they identify should concern lawyers, legal academics and socio-legal scholars. The answer, unsurprisingly, is "yes."

It is easy to find in the legal and law-related literature anecdotal evidence of the problems and misunderstandings that concern Ziliak and McCloskey, sometimes in work that has had significant impact. For example, Richard Sander, who holds both a Ph.D. in economics and a law degree, published an article in the *Stanford Law Review* in 2005 that purports to show that affirmative action harms black law students more than it helps them.[36] Results from this study have been highlighted in newspapers and blogs, discussed on NPR and presented to the United States Commission on Civil Rights.[37] His article could, however, stand as a poster child for the issues Ziliak and McCloskey discuss. Not only does Sander's article rely largely on significance tests rather than measures of strength to carry its statistical argument forward, but, even worse, it explicitly instructs its readers, many of whom will be statistically naïve, that statistical significance indicates substantive significance:

---

and the latter a relationship significant at the .01 level. However, when jury discrimination is the issue, if there has been intentional discrimination and samples are not small, differences between the expected and observed values of minority or female jury membership are likely to be greater than even three standard deviations.

[36] Richard H. Sander, A Systemic Analysis of Affirmative Action in American Law Schools, *57 STAN. L. REV. 367* (2004) I have been a critic of Sander's research methods and conclusions. *See e.g.*, David L. Chambers, Timothy T. Clydesdale, William C. Kidder, and Richard O. Lempert, The Real Impact of Affirmative Action in American Law Schools: An Empirical Critique of Richard Sander's Study, 57 *Stan. L. Rev.* 1855 (2005)

[37] It is one of those rare studies that received prominent mention in the *New York Times*. Adam Liptak, For Blacks in Law School, Can Less be More? New York Times, News of the Week in Review, February 13, 2005. It was also the subject of a story in *The Chronicle of Higher Education*. Katherine S. Mangan, "Does Affirmative Action Hurt Black Law Students," Nov. 12, 2004.

> The "t-statistic" tells us how consistent or reliable a relationship is, with a higher t-statistic indicating a stronger, more reliable association. T-statistics generally increase as a function of the standardized coefficient and the size of the sample. T-statistics above 2.0 are usually taken to signify that the independent variable is *genuinely helpful* in predicting the dependent variable. A t-statistic of less than 2.0 indicates *a weak, inconsistent relationship*—one that might well be due to random fluctuations in the data. (emphasis added)[38]

Thus Sander encourages his readers to draw conclusions that Ziliak and McCloskey, numerous econometricians, statisticians and others whose names they mention, and many whose names they do not mention, caution against. These experts correctly advise that the fact that a significance test on an independent variable exceeds the .05 level[39] tells us nothing about whether that variable is "genuinely helpful" in predicting a dependent variable,[40] nor does an insignificant t-statistic necessarily indicate a "weak, inconsistent relationship."

To move from the level of anecdote and gain a more general picture of how problematic Ziliak and McCloskey's concerns are in the context of socio-legal scholarship and legal decision making, I examined small samples of judicial decisions and socio-legal articles that featured quantitative data. I identified relevant decisions by searching the Westlaw federal courts data base using the terms "statistical significance" or "statistically significant" or ".05 level" or ".01 level." To identify articles I examined 14 articles in the *Harvard Law Review* retrieved through Westlaw by searching with the term "statistical

[38] *Id* at 428-29. The *Stanford Law Review* is a student edited rather than a peer reviewed journal. One would like to think that a misstatement like this would not pass the screen of peer review.
[39] A *t* statistics of 2 (1.96 precisely) is significant at the .05 level.
[40] Professor Sander inadvertently illustrates this point in a logistic regression of 21,425 cases whose implications he calls "profound." In this regression, which aims at predicting whether a law student will pass the bar, law school GPA, LSAT score, law school tier and undergraduate GPA are all significant beyond .0001, and being male is significant beyond .01. Moreover, the equation as a whole improves our ability to predict who will pass the bar to a statistically significant degree (beyond .001). This sounds good, but 95% of the law students in this sample pass the bar. If in each case we predicted the student would pass the bar, we would be right 95% of the time. If we used Sander's model to refine our prediction, we would be right in an additional 29 of the 21,425 cases or 95.1% of the time. In short, we have highly significant results, but what they tell us is of virtually no use, and the implications of the regression are hardly profound. Sander, *supra* note 28 at 444-45.

significance," I similarly identified and examined about ten additional articles retrieved from Hein on Line, and I skimmed all the quantitative articles in two issues each of the *Law and Society Review* and the *Journal of Empirical Legal Studies* that I happened to have at hand.

I do not want to oversell this effort. It is impressionistic and casual rather than systematic and scientific. I did not attempt to apply the Ziliak and McCloskey scorecard to the writing I examined, but focused only on their core concerns; namely, whether significance tests were inappropriately relied on and whether information about the strength and importance of associations and differences was presented and appropriately used.

Courts, it appears, too often focus only on statistical significance, perhaps because this is what expert witnesses emphasize in their reports and testimony. I had originally intended to examine 25 recent decisions, but after looking at the first ten I retrieved and finding only 2 that paid any attention to the impact of findings or the weight they deserved in deciding the dispute, I decided it would be a waste of time to continue the investigation. Not only did courts in the cases I looked at seldom refer to more than statistical significance in discussing studies offered in evidence, but they often did so in ways that indicated little understanding of statistical concepts.

For example, in a case arising from an allegation that a drug company, GlaxoSmithKline, had committed securities fraud by failing to disclose adverse drug trial results, a panel of the Second Circuit wrote:

> We have held that reports of harmful drug effects are immaterial - and thus need not be disclosed - unless those reports … show statistically significant evidence of an adverse effect… The complaint does not explain how the results of the research trials at issue could be deemed statistically significant in light of the test results from another trial that GSK did disclose.[41]

[41] Masters v. GlaxoSmithKline, Slip Copy, 2008 WL 833085, C.A.2,2008. March 26, 2008

If the panel is using the word "immaterial" in its legal sense, this language tells us that studies that don't meet its significance criterion are *completely unrelated* to any fact in issue in the litigation. This is the same as saying that a study showing a drug to be more harmful than an alternative or placebo has no bearing on the drug's likely safety so long as the study did not find the difference in ill effects to be statistically significant. Then, since statistically insignificant results do not, in the court's view, convey any information about a drug's likely safety, the court concludes that disclosing the results of the study in this case would not have affected the price of a company's stock. The court is wrong on both counts. After receipt of the adverse trial results, confidence in the safety of the drug in question should diminish, and if the drug were important enough to GSK's balance sheet, the price of its stock should decline.

Moreover, the court fails to perceive how the results of one study might affect the evaluation of another. From the court's language it seems likely that GSK conducted two studies of its drug's safety, both with small N's and thus low power, each of which yielded statistically insignificant results. Taken together, however, in a mini meta-analysis, the evidence of adverse effects may have been, as the court's allusion to the complaint suggests, statistically significant. The court's language dismissing the claim suggests that the plaintiff's attorney did not explain why this might happen and that no judge was able to grasp the matter on his own.

In other cases "statistical significance" appears to be used more as an incantation than as a meaningful description of testimony or evidence. Consider a ruling by a panel of the

10th Circuit, overturning a District Judge's decision to exclude expert evidence in an insider trading case:

> Professor Fischel's testimony was to include a discussion of the economic incentives that inside information would have given Mr. Nacchio, the statistical significance of the differences in his trading patterns, and the likelihood that economic diversification better explained the challenged sales than inside information.[42]

What is the statistical significance of the differences in trading patterns? I have no idea; nor, I think, does the court.[43]

A panel of the 9th Circuit exhibits, if anything, even greater confusion in reviewing a lower court's decision to admit 120 statements relating discrimination experienced by Wal-Mart employees:

> Wal-Mart contends that the district court erred because the 120 declarations cannot sufficiently represent a class of 1.5 million. However, we find no authority requiring or even suggesting that a plaintiff class submit a statistically significant number of declarations for such evidence to have any value.[44]

I expect the panel was correct in rejecting Wal-Mart's claim of error, but it is meaningless to talk about receiving *a statistically significant number of declarations*. Although the likelihood of obtaining statistical significance is affected by sample size, the two should not be confused, and there is no such thing as a statistically significant sample size.[45] The court seems unaware of this.

---

[42] U.S. v. Nacchio, 519 F.3d 1140, 1155, C.A.10 (Colo.),2008.

[43] Since Fischel's methods were apparently never described, the court seems not to have had any particular comparison condition implicitly in mind. Thus its language reminds me of the old joke, "What's the difference between a duck?" Answer: "One leg is both the same."

[44] Dukes v. Wal-Mart, Inc., 509 F.3d 1168, 1182, C.A.9 (Cal.) ,2007.

[45] Since the declarations were unanalyzed volunteered complaints rather than any sort of researcher acquired sample, the language of statistical significance would be inappropriate even if there were thousands of declarations. Perhaps the misuse of "statistically significant" in this case and the misuse of "statistical significance" in *Nacchio* reflects judicial adoption of language used by the lawyers arguing to them. If so, the original misunderstanding may have been an attorney's, perhaps as a rhetorical flourish. Regardless, neither the *Wal-Mart* nor *Nacchio* panel had a

Having found three such statements in the first ten cases I looked at, I hope the reader will understand why I did not continue looking until I reached my target sample. After ten cases I had read enough to conclude that the law could only benefit if Ziliak and McCloskey's book were required reading for judges.

Scholars Do Better

Turning to socio-legal scholarship, a different picture emerged from my informal study. There were, to be sure, some articles whose authors would qualify for membership in Ziliak and McCloskey's cult of significance, and if I encountered a few such articles in the small group of articles I examined, there must be many more in the larger literature. What struck me most, however, was the degree to which articles reflected the caution that concerns us and focused on the magnitude and potential importance of effects. Articles frequently presented their raw data in detail, and when significance levels were noted, they were often accompanied in the text by attention to how much different variables mattered. Indeed the lawyer or legal academic who reads widely in the empirical law and social science literature is likely to encounter enough good examples that the difference between statistical and substantive significance might be unconsciously absorbed.

Ian Ayres, in a study of discrimination in the prices offered car buyers, is exemplary. He instructs his readers that they should "focus … not merely on statistical significance but also on the amount of the reported discrimination."[46] And, in case this isn't clear enough, he

---

member who understood the concept of statistical significance well enough to know when a reference did not make sense.

[46] Ian Ayres, Fair Driving: Gender and Race Discrimination in Retail Car Negotiations, 104 *Harv. L. Rev.* 817 (1990-1991)

points out in a footnote that in a large enough sample differences in prices offered of only $5.00 might be statistically significant. Perhaps because socio-legal scholars often write with an eye to the policy audience, many of them take care to follow their identification of significant variables with estimates of the importance of the variable's effects.[47]

Preponderance of the Evidence

Since Ziliak and McCloskey direct little attention to the law, they ignore an interesting problem. By the conventions of statistical significance, a scientist should not claim to have identified a true effect unless the effect is not likely to have arisen by chance.[48] In civil actions, however, a court is supposed to decide the matter before it by "a preponderance of the evidence," which is usually taken to mean that a plaintiff should prevail whenever the probability that her claim is justified exceeds 50%. How do we reconcile this burden of proof with the tendency of courts in civil cases to ignore, or even refuse to admit, study results that are not significant at the .05 level? Results need not be significant at the .05 or even the .1 level to make a plaintiff's claim appear more likely true than seemed to be the case before the statistical evidence was known.[49]

---

[47] Jodi L. Short and Michael W. Toffel, in an article looking at what leads firms to voluntarily disclose their environmental violations typify the attention to importance I found in a number of articles. They write: The statistically significant positive coefficients on inspections and enforcement actions support our hypothesis that specific deterrence measures encourage self-disclosure. The results suggest that an additional RCRA inspection increases the probability of self-disclosure the next year by 14% ($p = 0.020$) and that an additional CAA inspection increases this probability by 11% ($p = 0.053$). Being subject to at least one enforcement action—a much rarer event—had a much greater influence on disclosure, as our results suggest that this more than doubles the likelihood of self-disclosing the next year compared to the probability evaluated at the means of all variables ($p < 0.001$), Coerced Confessions: Self-policing in the Shadow of the Regulator, *Journal of Law, Economics, & Organization*, May, 2008

[48] Depending on the field and the topic, the conventional minimum probability is usually one time in ten or one time in twenty, with some arguing that these probabilities can be doubled if there is a strong prior hypothesis of directionality

[49] Law-trained readers will recognize that the last part of this sentence fits well with FRE 401"s definition of relevant evidence and will know that relevant evidence is admissible unless excluded by some other rule or principle.

To make the issue concrete, suppose a plaintiff attributes his kidney problems to a drug he took, and in a suit against the drug's manufacturer offers a study that shows that patients in a randomized clinical trial who took the drug were three times more likely to be later diagnosed with kidney ailments than patients in the placebo condition. Suppose also that the study had a relatively small number of participants and the 300% difference in the incidence of kidney ailments was significant at only the .2 level. Following dominant conventions, we would not call the difference in the incidence of kidney problems in the two conditions statistically significant. Noting the lack of statistical significance, many courts would refuse to admit the study's results, or, if they were admitted, would ignore them in a bench trial or on a motion for a directed verdict. Can excluding or ignoring findings consistent with the plaintiff's position but lacking statistical significance be justified when a plaintiff need only prove her claim by a preponderance of the evidence?

One argument supporting exclusion is that an inexpert court should not second guess an expert regulatory agency that judged the scientific evidence insufficient to show a link between kidney disease and the drug in question. This, however, is to ignore the evidence rather than evaluate it.

A second argument, suggested to me by the eminent biostatistician Paul Meier, is that if data indeed reflect a causal relationship, demanding significance at the .05 level is not a serious hurdle.[50] A court might put the point this way:

> "There are many causes of kidney disease. Even if the plaintiff could show that we had considerable reason to be confident that the drug he took *could* cause kidney problems, he would still have difficulty showing the drug caused *his* kidney problems. If he can't even show that there is enough evidence to persuade scientists that the drug he took can cause kidney problems, I am not going to treat the study he

[50] Meier made this argument to me about twenty-five years ago, commenting on an article I had sent him.

> cites as evidence that the drug may have caused his kidney problems. If the plaintiff were right and the drug in fact caused kidney problems, I am confident the data would have made this abundantly clear."

As a practical matter, the judge will most often be right in his or her judgment, just as Meier's observation is most often correct, particularly in the case of the large scale randomized clinical trials which he long advocated and for which he is best known. But as a statistical matter there are problems here. As Ziliak and McCloskey point out, many real world studies, including some in the biomedical world, have low power,[51] so there is a good chance that real relationships will not pass the screen of conventional statistical significance. More importantly out hypothetical judge is using significance tests in exactly the wrong way, as indicators of the weight of evidence. Not only might the judge exclude evidence that would not pale into insignificance next to other potential causes of the plaintiff's kidney problems, but a judge who responds reflexively to the conventions of statistical significance is vulnerable to a greater danger. Had the link between the drug and kidney problems been statistically significant, the judge (or jury) might have accorded the study results far *too much* weight given other evidence in the case and the many possible causes of kidney problems.

Concerns Beyond Statistical Significance

Ziliak and McCloskey mention but do not describe in detail statistical approaches that researchers might adopt to aid judges in deciding whether to admit evidence and how to value it; these sometimes implicit and sometimes explicit suggestions are beyond to scope of

[51] This may either be because of small sample size or because they are searching for weak but important effects, such as small increases in risk, which despite their small size can lead to many excess deaths when spread across a large population.

this essay.[52] I do, however, want to elaborate on one point Ziliak and McCloskey mention but do not develop. The authors remind us at several places that a study's results may be suspect for many reasons apart from a lack of statistical significance. Indeed, one of their take-away messages is that we should not rely on a study's results simply because they attain some high level of statistical significance.

Recall what significance tests do. They deal with just one of the rival hypotheses that might explain a result, and except in studies involving randomized designs or random sampling, they do not usually address the most plausible of those rival hypotheses. Indeed, even where there has been randomization, other concerns may mean that reported results are not what they seem. We should thus add to "chance" other common plausible rival hypotheses that the reader of any quantitative -- and much qualitative -- empirical research should be alert to, especially when research results are offered as reasons for action.

> A v*ariable does not have the meaning the author gives it.* IQ for example, is not intelligence. It is a score on a test that imperfectly measures aspects of a broader concept, intelligence. Similarly, crimes reported to the police do not necessarily reflect the incidence of crime in an area. Apparent sharp changes in crime rates may be traceable to changes in how the police record complaints or efforts to get victims to report crimes. Variables may also not bear the meanings attributed to them because of coding conventions or coding errors. Readers of empirical work should

[52] For example, they make clear their sympathy with Bayesian approaches to the evaluation of evidence and the need to take a cost-benefit perspective when deciding how much to credit evidence.

always look behind the language in which results are presented for information about how a concept has been *operationalized*, which is to say instantiated in the data.

Sample biases favor a particular hypothesis. In our hypothesized pre-school study for example, results showing a strong effect of pre-schooling on high school graduation would be suspect if the only students pre-schooled were those whose parents had chosen to enroll them. Wherever subjects are selected for study, whether by the researcher, third parties, self-selection or by a process (*e.g.* only those whom the police arrest can be allowed or denied bail), readers should always ask whether the selection mechanism might potentially bias results one way or another and, if so, whether the biases have been controlled statistically or otherwise taken account of.

Crucial variables are omitted from the model. A study of racial discrimination in the administration of the death penalty would most likely conclude that race had no effect if it considered only the likelihood that blacks and whites who committed crimes of similar heinousness would receive the death penalty. But if the model also included information on the victim's race and the victim-defendant racial pairings, race would appear to play a substantial role in determining who is chosen to die.[53] Readers should look carefully at variables that are neither included nor proxied for in a study and ask whether their inclusion might plausibly have changed the study's results.

[53] David Baldus, George Woodworth and Charles Pulaski, *Equal Justice and the Death Penalty: A Legal and Empirical Analysis*, Northeastern University Press (1990); Sam Gross and Robert Mauro, *Death and Discrimination: Racial Disparities in Capital Sentencing*, Northeastern University Press, 1989.

The model has the wrong functional form or is unduly influenced by outliers. For example, a model may posit a linear relationship between variables when the relationship is curvilinear. This can result in a poor fit between model and data even though the model's independent variables are major determinants of the dependant variable when the relationship between the two is properly understood. Thus, a linear model associating income with age may find little relationship if a study follows people to age 90 because past a certain age incomes diminish, but it would be a mistake to conclude from the poor fit that income and age are unrelated. A model which showed income flat through about age 16, then rising with age through some age in the 60s and then falling off would portray the income-age relationship more accurately. Alternatively a model may fit misleadingly well, as when a few extreme cases have undue influence. Thus Isaac Ehrlich in a well known study claimed to have shown that over a period of several decades each execution deterred 8 homicides. But this result depended entirely on Ehrlich's decision to extend his time series into the 1960s when homicides were rising and executions had mostly stopped.[54] Richard Berk has similarly shown that in more recent research suggesting executions deter homicides, the association seems due entirely to the inclusion of a few outlying states, most notably Texas.[55] Readers should consider the reasonableness of the assumptions built into models for the problem at hand and the

[54] Richard Lempert, Desert and Deterrence: An Assessment of the Moral Bases of the Case for Capital Punishment, 79 MICH. L. REV. 1177 (1981)
[55] Richard Berk (2005) New Claims about Executions and General Deterrence: Déjà Vu All Over Again? Journal of Empirical Legal Studies 2 (2) , 303–330

degree to which results are sensitive to the inclusion or exclusion of particular variables or different ways of measuring them. (*e.g.*, Models including income often fit better when income is given not in actual dollars but by the log of the dollar amount. This dampens what would otherwise be the undue influence of a few cases with very large incomes.) Also, rather than looking just at summary statistics, readers should consider which cases a model predicts well and which it predicts poorly.

Rival hypotheses like those above often pose more severe threats to the conclusions an author asserts than the possibility that an asserted relationship exists by chance, but they often receive less attention than the chance hypothesis because without easy to perform tests, like significance tests, the degree to which other rival hypotheses are of concern is harder to nail down.[56]

[56] In many instances perceiving concerns like those I mention requires no technical statistical knowledge, although they may be obvious only to those who have considerable substantive knowledge of the area investigated. If, however, it is a mistake to ignore rival hypotheses like these, it is also easy to make too much of them. Only rival hypotheses that seem plausible are a serious concern, and the fact that a potential confound or other problem exists in a study does not mean the study's results are necessarily invalid. It is only a slight stretch, if it is a stretch at all, to say that no study will control for all plausible rival hypotheses, much less implausible ones. Yet lawyers motivated to attack research often try to convince a judge or jury that this is the case. For example, in a gender discrimination case brought against a university, a lawyer for the defendant might argue that a plaintiff's model is flawed because it measures productivity by the number of articles produced rather than by the selectivity of the journals in which they were published. This is a valid criticism since university salary setting is supposed to reflect not just the amount of scholarship produced but also its quality, and publication in selective peer-reviewed journals is regarded as a good proxy for article quality. Nevertheless, results from a model that measures scholarship without attending to peer review are not necessarily invalid, nor should they be inadmissible. Unless we know that the plaintiff had good data on publication quality and for no good statistical reasons chose not to use it in the model, there is no basis for assuming that had the data been included the model's finding of discrimination would vanish. If the defendant university thought the variable was important, it could have replicated the plaintiff's model, adding the publication quality variable. The same is true of other criticisms, like criticisms going to functional form, which could be tested if a party thought they really mattered but are instead often used only to suggest that the opposing party's evidence cannot be trusted. A similar point can be made about tendencies to criticize research for not controlling for factors that could not be controlled because of lack of data. The fact finder can consider the likely implications of being unable to control for plausible relevant variables but should not assume that controlling for them would have changed anything. Research can proceed only with the data that is available or which can be gathered at a reasonable cost relative to the knowledge or legal stakes. Courts that do not hesitate to credit a party's case even though a crucial witness has died or is otherwise unavailable, should not reject research results simply because information on a variable they would like to see included in a model cannot be acquired. If the data are

Virtues, Shortcomings and Peculiarities

The Cult of Statistical Significance has virtues that extend beyond its core message. It is clearly written and should be accessible to those who have neither formal training in statistics nor a desire to secure any. It is full of examples that illustrate why it is the strength of relationships and not their statistical significance that mainly matters. It mentions various ways that researchers can convey information about the substantive significance of what they have found. It highlights the need to view claims of no statistical significance in light of the power of the statistical tests employed. And, in a portion of the book I have not commented on except in a footnote, it paints fascinating portraits of many of the founders of modern statistics and their intellectual and personal relationships.

The book also has some shortcomings and peculiarities. For authors who rightly fault significance tests for the rigid, insensitive cutoff conventions they establish, Ziliak and McCloskey are surprisingly rigid in the standards they impose when evaluating the work of others. Reading their studies and critiques, one would think that the state of empirical economics and the other empirical social sciences is far worse than it actually is. Some might also be reluctant to credit any clinical trial that suggests a drug is no more dangerous than placebo unless the actual proportion of adverse incidents in the trial is the same for both conditions or favors the drug . A person might also mistakenly think that there was nothing substantive to be learned from articles that are more attentive to statistical significance and less attentive to impact than the authors would like. Yet often attention to the size of

available and are not included in a model, then there is the question of whether a spoliation inference is justified. But the better solution is to redo the model with the omitted data.

unstandardized coefficients, to the relative size of standardized ones, to explained variance and to the differential explanatory power of different models will allow moderately sophisticated readers to extract from articles information about importance that significance levels alone do not give, even if an article's authors do not emphasize all that can be learned.

Turning to peculiarities, one which helps make the story interesting is the authors' frequent attempts to characterize their endeavor as a struggle between the inheritors of the statistical philosophy of the good William Gosset (the "Student" of Student's *t* test) and the heirs to the statistical philosophy of the malevolent Ronald Fisher. In this they are, despite a few appropriate caveats, probably less kind to Fisher than Gosset would have been, for the book also discloses that Gosset admired Fisher as a mathematician and considered him a friend.

I could have done without another peculiarity. The authors feel compelled at times to wax poetic, and at various places they summarize an observation in what they consider a haiku. For example, their summary view of Fisher is,

> Here's a scientist
> Who sank the world with a *t*
> 5 percent per cup.[57]

It's hard to say whether the pun or the poetry is worse. Just as a significance level of .05 doesn't mean a relationship matters, so encapsulating a thought in 17 syllables in a 5-7-5 syllable per line arrangement does not a lovely Japanese haiku make. Doubters are referred to my subtitle.

Curing the Problem

[57] *The Cult* p.226

In a short final chapter, Ziliak and McCloskey address the question of how to cure the core problem they have identified. Part of their prescription is an exhortation to scientists to pay better attention to the message that runs throughout the book – in doing research and writing up results scientists should seek to understand what is important. At best, they tell us, significance levels make only a limited contribution to this determination, and they should certainly not be the sole focus of attention. To persuade the world of this truth, the authors call on Nobel Prize winners and other prominent econometricians to take up their cause and reiterate their message. They also encourage scientific gatekeepers, particularly journal editors, to insist that articles attend to the size and impact of effects using measures more revealing than tests of significance. They conclude by referencing the man who is the hero of their story, telling their readers to "repent" and urging them to "Embrace your inner Gosset."[58]

I would second both their advice and their desires, but my prescription for reform begins elsewhere. It is cognitively too easy to confuse statistical significance with actual (substantive) significance. This problem is a bad enough as it affects journal editors, reviewers and scientists. It is worse when students are reading studies that emphasize statistical significance over effect sizes. And correcting misunderstanding can be nearly impossible when statistically untrained people, including most judges and lawyers, are seeking to understand the implications of empirical research. This last is particularly unfortunate as statistically untrained politicians, judges and their ilk are most often in a position where they can or must act on their understanding of what research has to say.

[58] *Id,* at 251 The combination of revivalist and new age vernacular is striking. With tongue only slightly in cheek, I would put their bottom line thusly: We should all be new age social scientists and go back for the future.

Articles do not by themselves affect lives, no matter how much oomph exists in a model. Research affects lives only when people act on it. If those with power to act, whether by deciding a case, enacting a law or prescribing a medicine, cannot distinguish what actually matters from the likelihood that an association might exist by chance, actors with oomph may push in the wrong direction or refrain from pushing when it counts. Hence my prescription for reform begins with eliminating the term "statistical significance" -- or at least the "significance" part of it -- from our vocabulary. Other words with both technical and common sense connotations, like reliability, should be similarly avoided. We might, for example, rename "significance levels" "random possibility indexes" and refer to significant associations (or differences) as "non-chance associations (or differences)." But I have no brief for a particular term so long as some term or acronym unlikely to be confused with importance is used. Thus I suggest, in all seriousness, that the Committee on National Statistics of the National Research Council (the National Academy of Science's research arm) meet to settle on a term to replace statistical significance. With the Committee Members' personal prestige and the prestige of the National Academy behind them, perhaps a long overdue linguistic change might occur.

Why a Book?

This essay is almost at its end. All that remains is to answer the question my opening stanza poses: why a book? Do we need *The Cult of Statistical Significance*? We certainly shouldn't. As the authors acknowledge, their message is almost a century old, and it is not difficult to understand. Moreover, except for best sellers, a category into which few

academics works (surely not including this one) fall, more people read articles than books, and teachers seldom assign books rather than shorter works that make the same points.

Still I welcome the book's publication. It is well written and interesting. It collects far more information and arguments between its covers than any article could. Its publication could trigger a beneficent burst of attention in a variety of disciplines to the proper use of significance tests and the need to present measures that focus specifically on importance. Perhaps it will even give courage to editors who wish to deemphasize the role significance testing plays in deciding whether results are worth publishing or to reviewers who might otherwise hesitate to insist on revisions that make measures of impact front and center. So there are good reasons for this book; reasons that will increase with every lawyer, judge and law student who reads it. Legal academics too should not be left off this list. Those who do or read empirical research, which now includes most of us, can only benefit from taking to heart the lessons Ziliak and McCloskey teach.