Journal of Economic Methodology,
Vol. 15, No. 1, March 2008, 57–68

The rhetoric of 'Signifying nothing': a rejoinder to Ziliak and McCloskey

Kevin D. Hoover[a]* and Mark V. Siegler[b]

[a]*Departments of Economics and Philosophy, Duke University, Durham, NC, USA;* [b]*Department of Economics, California State University, Sacramento, CA, USA*

For a quarter century, Deirdre McCloskey has pushed the idea that economists should take rhetoric seriously. So, here we would like to take her advice and consider the rhetoric of Ziliak and McCloskey's reply to 'Sound and fury.'

If that old rhetorical device, repetition, were a logically compelling argument, then we would be obligated to concede and withdraw our paper. Alas, it is nothing but the fallacy of *argumentum ad nauseam*; and, indeed, once the repetition is stripped away it is clear that they give no adequate answer to our key points and that our argument emerges unscathed. Much of their oft repeated points also commit the fallacy of *ignoratio elenchi* – that is, arguing against a point that was never maintained in the first place. Ziliak and McCloskey write as if we had offered a general defense of existing econometric practice. In fact, we are quite critical of many aspects of econometric practice, but we think that Ziliak and McCloskey are wrong on particular points. They also write as if we were simply channeling their *bête noir*, R.A. Fisher. We are happy to stipulate, along with one of their heroes, that

> If only … if only … RAF had been a *nicer* man, of [*sic*] he had taken pains to be clearer and less enigmatic, if only he had not been obsessed with ambition and personal bitternesses. If only. But then we might not have had Fisher's magnificent achievements. (Kruskal 1980, p. 1029; emphasis in original)

Nevertheless, there are some things on which we agree with Fisher and others on which we do not, and we have not rested our argument on Fisher's authority: our argument stands or falls on its own merits, whatever Ziliak and McCloskey think of Fisher.

Our argument has two parts, which we consider here in the reverse order from our original paper: First, we disagree with Ziliak and McCloskey's strong claim that the areas in which tests of statistical significance can be usefully and validly applied are so constricted that 'all the econometric findings since the 1930s need to be done over again' (p. 47). Second, although we agree that it is a mistake to confuse economic (or substantive) significance with statistical significance, we find that Ziliak and McCloskey's evidence that this confusion is widespread in the economics profession to be unconvincing.

The most important example of *ignoratio elenchi* arises because Ziliak and McCloskey believe that, since we admit that economic and statistical significance are distinct, we must, therefore, concede that they are right to conclude that the use of tests of statistical significance is generally a mistake. But that is a *non sequitur*: Their

*Corresponding author. Email: kd.hoover@duke.edu

ISSN 1350-178X print/ISSN 1469-9427 online

DOI: 10.1080/13501780801913546
http://www.informaworld.com

argument takes the form: (A) apples are different from cherries; (B) cherries make poor pies; premise A is true; therefore, cherries make poor pies. The logical form is, of course, nonsense. That we acknowledge a difference between statistical and economic significance, and find it a mistake to confuse them, has no bearing on whether statistical significance is valid or useful. As cherries to a good pie, so the usefulness and validity of statistical significance requires independent arguments. But rather than supplying these arguments, Ziliak and McCloskey keep hammering on our admission that economic and statistical significance are different, which is simply off the point.

When the repetition is cut away, our arguments stand largely unchallenged and completely unscathed. Let us make an inventory, beginning with our first point: Ziliak and McCloskey do not make the case against statistical significance.

1 Noise matters

Our most important point is that it is wrong to base scientific conclusions or practical actions on the means of noisy estimates as if those estimates were not noisy, and that tests of statistical significance provide useful information on the degree of noise. Ziliak and McCloskey's only reply to our argument commits the fallacy of *ad misericordiam* (the appeal to pity). They repeat the example of the woman who cries 'help, help,' but whom we ignore on the off chance that she is yelling 'kelp, kelp,' and add more hypothetical examples of misery: 'killing cancer patients and misdiagnosing schizophrenia' (p. 44). But they fail to address the argument that we illustrate with the investment equation. Taxes may be highly economically significant for investment; but, if the noise is high relative to the signal (e.g. if the variation in tax rates is low), then the estimate of that effect may be wildly different from the real value; indeed, it may be quite opposite in direction. Statistical significance is a measure of the strength of a signal relative to background noise. Ziliak and McCloskey's slogan 'size matters' advertises the fallacious conclusion that we ought to be guided by noisy measurements as if there were no noise at all. Experience shows, for example, that, when pilots act as decisively on noisy signals as on clear ones, airplanes sometimes crash with spectacular loss of life. Pity the passengers. Pity the migraine sufferers given licorice jelly beans. Pity the cancer sufferers given laetrile. And pity the victims of assault who are neglected when the police are chasing will o' the wisps of what might have been 'help' or 'kelp' or just the wind in the trees.

Ziliak and McCloskey cannot make up their minds about whether our mistake is to insist that we need to consider the noise or that statistical significance provides a useful way of assessing that noise. We admit that the 5% critical value is a conventional threshold, and that one could choose other conventions. Ziliak and McCloskey object: 'The criterion for ''measurable'' cannot be handed over to ''a conventional threshold.'' Scientific judgment is not like that' (p. 47). Actually, scientific judgment is often like that; conventions are employed widely for a variety of purposes. To be conventional is not necessarily to be arbitrary nor ungrounded nor to be mechanical, which is a suspension of *judgment* that neither we nor serious applied economists would endorse.

Beyond that, Ziliak and McCloskey confuse 'measurable' with a standard of assessment. A *p*-value (the probability of type I error) is a well-defined measurement; the critical value is a convention with respect to its size. Such conventions play

context-dependent parts in both science and everyday life: small, large, extra-large in everything from women's clothing to soft drinks; the classification of weather systems from tropical depression to tropical storm to hurricane (with five subcategories, defined – conventionally, but not arbitrarily – by wind speed and air pressure); and so on. That these classifications are not absolute makes them no less useful. The 5% rule (or the 1% or 10% – context matters) is a convention aimed at trading off size versus power against diffuse alternatives. In applied work, the rationale is typically implicit; while among econometricians developing testing procedures, it is frequently explicit (e.g. see Hoover and Perez 1999, pp. 11–22, and 2004, pp. 776–778). The issue with respect to any convention is whether it is effective. Ziliak and McCloskey are content simply to point out that a convention is involved; but that is no objection at all.

Ziliak and McCloskey's specious objection to convention *per se* is a key factor in their advocacy of the Neyman-Pearson framework as an all purpose statistical method for science. We do not object to the Neyman-Pearson framework for those problems to which it can be effectively applied.[1] But 'no science without a loss function' is a slogan, not an argument. Originally, Ziliak and McCloskey offered no argument, beyond the fallacy of *ad verecundiam* (appeal to authority), for the Neyman-Pearson framework as a *generally applicable* methodology of science; nor did they provide an example of how any large-scale scientific advance was achieved through an appeal to loss functions or other formal decision-theoretic devices. And in their reply, they offer only one argument against our contrary case: to wit, that we 'indignantly assert … that every scientific question must have a vulgar application to a world of money' adding *sotto voce* '[t]hough many do [have such vulgar applications]' (p. 52). But, say Ziliak and McCloskey, in fact it is the loss function of the scientists *qua* scientists that matters.

We confess to having stressed the practical, although we never characterized this as either vulgar or monetary. And we believe that any fair reading of Ziliak and McCloskey's voluminous writings would reinforce this interpretation of the supposed advantages of the Neyman-Pearson framework. But Ziliak and McCloskey provide a highly partial account of our stated position, since we explicitly acknowledge that costs and benefits in science are 'generally defined in terms of their import to the enterprise of pure science' and that the Neyman-Pearson approach is a useful tool of scientific investigation 'where the alternatives are clearly articulable, the probability models tractable, and the loss functions relevant to the scientific problem' (HS, p. 19).

The key objection to the Neyman-Pearson framework as the general method of science – and the question that Ziliak and McCloskey never address – is whose loss function? Scientists, like people in ordinary life, have disparate goals, benefits, and costs – even when nothing but the scientific motivates them. Every loss function is somebody's. Yet the goal of scientific theories, with its appeals to empirical evidence, is to establish understandings that are common among scientists and not idiosyncratic or radically subjective. There is a distinction between what is true (or what we believe to be true), which holds for all, and the uses (even non-vulgar ones) to which truth is put. The latter is the natural domain of the loss function. Whose loss function should Newton have consulted? Not his own; for that would not account for the universal import of his conclusions. Nor scientists in general; for not

only is every loss function somebody's, every loss function operates in a specific context. Ziliak and McCloskey simply do not address this central question.

Yes, scientists must make decisions. Sometimes small problems can be cast in a tight Neyman-Pearson decision-theoretic framework. But at all times, scientists require observations that rise above the noise, and statistical significance tests are a tool for assessing such observations.

Economics is mainly an observational, as opposed to an experimental, science. In reiterating their objection to viewing economic data as samples from larger (perhaps hypothetical) populations, Ziliak and McCloskey would deny the tools of probability to economists. While professing to be friends of counterfactual analysis, the only argument that they can muster is to say that Fogel knows that his sample size is N=1. (Another fallacious appeal to authority when unaccompanied by an argument. And where does Fogel make this point?) Ziliak and McCloskey offer the example of crime rates in Philadelphia in the 1980s as not an acceptable random sample, even under the hypothetical interpretation, '[i]n view of autocorrelation and of structure and path-dependence over time' (p. 50). They do not address – except by assertion – the insistence in modern econometrics (starting with Haavelmo (1944)) on specification and specification testing, the aim of which is precisely to determine appropriate transformations such that non-experimental data can be cast into a framework in which the error terms conform to tractable probability models.

Ziliak and McCloskey write as if the typical empirical economist gathered data and ran regressions, judging their outcome, once and for all, from the *t*-tests on the individual regression coefficients, with no attention to specification or to the processes that might cause the *residuals* to deviate from properties that could be usefully modeled as random. This does not correspond to our own practice or to that of virtually all of the applied economists we know, who are generally concerned about choosing appropriate economic and statistical models. They are concerned about endogeneity, sample-selection bias, omitted variables, data structure (e.g. truncation, missing observations, appropriate transformations), unobserved heterogeneity, dynamic structure, to name but a few. The goal of accounting for these concerns is to deliver models of the data for which the assumptions of randomness hold well enough for reliable inference. Specification testing is essential in such model development, and it relies on significance tests that are not easily cast into a Neyman-Pearson framework.

Ziliak and McCloskey prefer estimation to testing. Estimates are quite useless without the assumption that they can be used counterfactually. We typically go from what the data *did* to what the data *might do*. Ziliak and McCloskey accept counterfactuals and advocate simulation. Once these are accepted, they have in fact undercut their own objection to regarding non-experimental data – properly specified and tested (and we cannot emphasize that qualification enough) – as legitimate samples. Haavelmo's (1944) interpretation of probability relies on exactly the counterfactual basis that Ziliak and McCloskey license. But Ziliak and McCloskey do not attempt to address the argument. Simply repeating 'sample size … One' (p. 50) is not a valid argument, but another example of the fallacy *argumentum ad nauseam*.

Ziliak and McCloskey's response to our points about confidence intervals illustrates yet another case of not wanting to accept the implication of one element of one's beliefs for another element. Apart from their specious calculations intended to

minimize the use of confidence intervals in economics, they concede our main contention that the confidence interval and the test of statistical significance rest on the same logical foundation and that their advocacy of the confidence interval is, as Elliott and Granger (2004, p. 548) put it, 'more of a ''literary critique'' than a deep point about whether or not the paper is misleading science.'[2] They cannot have it both ways: either confidence intervals are wrong or their point is the relatively minor complaint that there is a more effective means of communicating. Even if the minor point were true, it would hardly provide a basis for jettisoning all of empirical economics.

2 Who's behaving badly?

Let us take stock with regard to the first part of our argument (that is, that Ziliak and McCloskey fail to make the case against the logic or utility of statistical significance). Ziliak and McCloskey have marshaled only a desultory and fallacy-ridden response, one that leaves each of our points intact. Before turning to the second part of our argument, we wish to address some points, secondary to the main argument about econometric theory and the practices of economists, to which, nonetheless, Ziliak and McCloskey devote considerable space in their reply.

Of course, the use of statistical significance in the physical sciences is secondary to any discussion of economics – physics envy is pointless, as there are material differences between physical and economic sciences that naturally suggest systematic differences between their methods. However, it was Ziliak and McCloskey, not we, who raised the question. We show in considerable detail that physics does, when the situations warrant, employ statistical methods, including tests of statistical significance, familiar to economists. We dwelt on this point, first, because Ziliak and McCloskey are so clearly wrong when they repeatedly assert that 'physicists approximately never use tests of statistical significance' (Ziliak and McCloskey 2004, p. 533) or that a specific issue of the *Physical Review* contains 'not a single use of … statistical significance' (McCloskey 1999, p. 357). Falsehoods should not be allowed to stand – especially when they are so easily checked. Second, we find Ziliak and McCloskey's bold, but false, claims to illustrate a general feature of their work: they frequently make strong claims that evaporate on even cursory examination of the evidence.

Their response to our evidence? They duck and weave – asserting now their 'never' really means 'sometimes' and that their 'not a single use' really means 'quite a few times.' To diminish the force of the evidence, they simply cite what we had already noted: proportionately fewer physics articles employ statistical significance than do economics articles. But they do not address our point that differences in the empirical problems faced by physical scientists account for differences in relative use of statistical significance. (And indeed, since our data are scaled by the number of articles in the journals and not by the nature of those articles, part of the differences in proportion could be, for example, the differences in the number of non-empirical articles among the journals, for which statistical significance or any other empirical techniques would be quite irrelevant.)

Ziliak and McCloskey introduce completely hypothetical calculations to try to further diminish the relative importance of our evidence. This is rather like the scholastic philosophers who argued from *a priori* principles over how many teeth a

horse must have. The horse and the journals are there: so, why not just go and look? The advantage of making the numbers up is, of course, that then they say what you want them to – in this case, that proportions of 7% to 34% indicate rareness rather than ubiquity. The reader can judge for himself how compelling that is.

Recall that the point of Ziliak and McCloskey's false claims about the sciences was to bolster the case that statistical significance is a suspect tool – scientists presumably being wiser than economists (again a fallacious appeal to authority). But again the argument is specious. My toolbox contains a large pipe wrench, which I hardly ever use, as well as a screwdriver, which I use frequently. This fact says nothing about the appropriateness of the pipe wrench when I need to turn a large pipe. Anyone who doubts that scientists find tests of statistical significance to be an appropriate 'wrench' should simply follow up the precise references supplied in Table 3 of our original paper. Our facts are on the table; Ziliak and McCloskey's assertions on this point evaporate on inspection.

Ziliak and McCloskey claim that we intended to impugn their scientific integrity (p. 53, note 5). Although we were careful not to make such explicit judgments in our paper, we did report – and felt obliged to report, because they bear on the foundations of their argument – a number of instances in which Ziliak and McCloskey's bold claims did not stand up when one checks the supposed source. We do regard these instances as serious lapses in scholarship.

We have already dealt with the bold claims about the science journals. But there are others: First, McCloskey repeatedly used the example of aspirin and heart attacks to show that real scientists make decisions solely on 'oomph,' without regard to statistical significance. But her characterization of the example is false. In their reply, Ziliak and McCloskey do not controvert that judgment.

A second example is Ziliak and McCloskey's claim to have included all the relevant papers from the *American Economic Review* in their surveys; we show that they omit 8% of the articles in the survey from the 1980s and 29% from the 1990s. They reply that the additional articles do not change their findings. But, of course, that never was our point, since we doubt the usefulness of the surveys on independent grounds. The omissions simply illustrate the frequent mismatch between their claims and ascertainable facts.

Related to the omission of the papers, Ziliak and McCloskey (p. 46) get in a dig at us that misfires on close inspection. They write: 'But give credit where credit is due. Hoover and Siegler roused themselves at least to defend the papers we discussed by actual quotation or reference. Well, at any rate they defend five of the dozens of papers we discussed.' Dozens? The references in McCloskey and Ziliak (1996) list only 13 papers from the *American Economic Review*, while Ziliak and McCloskey (2004) lists nine more. Not only are these fewer than 'dozens,' some do not count toward their point, as several are praised for 'good' practices (e.g. Romer 1986) and some are discussed so briefly that there is no analysis to evaluate. In addition, they refer to a few additional papers from the *American Economic Review* without providing precise citations. We did not examine just the five papers that we discuss in detail, but every one of the 22 (=13+9) papers that Ziliak and McCloskey cite, finding the same sorts of interpretive problems throughout. And we also examined every paper in the *American Economic Review* for two decades closely enough to identify the 81 papers that Ziliak and McCloskey overlooked, despite their having met the criteria for inclusion in the surveys. If Ziliak and McCloskey had been as

thorough as they claim to have been, what is the likelihood that would have missed so many papers and what is the likelihood that any group of 5 of the 22 cited would illustrate such clear misreading?

A third example is Ziliak and McCloskey's claim that Edgeworth supports their view of statistical significance. But we showed in detail in note 10 of our original paper that Edgeworth in fact advocates the use of statistical significance and that Ziliak and McCloskey miss Edgeworth's point in his discussion of Jevons. In their reply, Ziliak and McCloskey do not confront this head on. Instead, they now say that we misunderstood and that it is not the case 'that when we recommend "Edgeworth's standard" we mean his conventional level of significance. That's actually what we *don't* like about Edgeworth (1885) in contrast to, say, Edgeworth (1907)' (p. 40). But this is disingenuous; they misrepresent their own article. Ziliak and McCloskey (2004, p. 531) advise that the economics 'profession adopt the standards set forth 120 years ago by Edgeworth ….' The year 1885 is 119 years before Ziliak and McCloskey's 2004 publication date, whereas 1907 is only 97 years before publication. And Edgeworth (1885) is cited in Ziliak and McCloskey (2004); while no other work by Edgeworth, including Edgeworth (1907), is cited there.

But even if they really did mean Edgeworth (1907), it does not help their case, since even a casual look at the text shows that Edgeworth maintains essentially the same view in 1907 as he held in 1885. The 1907 paper is an attempt to infer statistically the average time that wasps and bees spend away from their nests based on their entrances and exits. Edgeworth (1907, pp. 379–380) proposes a test based on a comparison of the observed modulus to the modulus under what we would now call the null hypothesis, citing his 1885 paper (pp. 209–210), both with reference to earlier observations of wasps and to his testing procedure. (Recall from our original paper (sec. 3.1) that Edgeworth's *modulus* is a rescaling of the standard deviation.)[3] After providing a statistical analysis, Edgeworth writes:

> If in an insect republic there existed theories about trade as well as an industrial class, I could imagine some Protectionist expressing his views about 12 o'clock that 4th day of September [the day of Edgeworth's observations], and pointing triumphantly to the decline in trade of 2½ per cent. as indicated by the latest returns. Nor would it have been easy off hand to refute him, except by showing that whereas the observed difference between the Means is only 2, the modulus of comparison is $\sqrt{\frac{70}{5} + \frac{70}{13}}$, or 4 at least; and that therefore the difference is insignificant. (Edgeworth 1885, p. 209)

Ziliak and McCloskey's suggestion that we follow 'oomph,' whatever the uncertainty of measurement is precisely the position of Edgeworth's Protectionist wasp – a position that he rejects in 1885 and cites to support significance testing in 1907.

A fourth example is Ziliak and McCloskey's claim 'to have examined scores of' econometrics textbooks. Perhaps. But they cite only seven econometrics textbooks (not counting statistics textbooks) in McCloskey and Ziliak (1996) and only two in Ziliak and McCloskey (2004). The charge that econometrics textbooks inculcate and perpetuate mistaking statistical for economic significance, is mostly based on a highly selective reading of the texts, focusing on what they do not say. None advocates the mistake as sound practice. The only evidence that McCloskey and Ziliak (1996, p. 100; also McCloskey 1998, p. 117) offer for positive error is based on three pages (misrepresented as 26 pages) of Johnston's (1972) once-popular econometrics text. We do not have space enough to demonstrate here that Ziliak

and McCloskey's charges of error in Johnston's text are utterly false, relying on a misinterpretation and selective reading that is obvious to anyone who reads the text. The interested reader will find the details in the earlier draft of 'Sound and fury' (pp. 40–44) available at http://econ.duke.edu/~kdh9/research.html.

Ziliak and McCloskey do not always give us the opportunity to check their claims. For example, they say 'Hoover and Siegler quote but misunderstand' William Kruskal (p. 40) without saying which of our quotations is in question, what they find misleading, or where we should look in Kruskal's work for support of their claim. Ziliak and McCloskey routinely assert support for their position of various statisticians and economists without pointing to the passages that warrant their claims. Case in point: Horowitz (2004) makes it clear that he strenuously disagrees with Ziliak and McCloskey's analysis. To take one of many examples: 'The distinction between statistical and substantive significance is important, but there are circumstances in which only the existence of a phenomenon, not its magnitude, is decisive' (Horowitz 2004, p. 551). Yet, they cite him as among their supporters (p. 45).

And Ziliak and McCloskey claim *secret* supporters as well: 'Clive Granger is one of the many econometricians who pretty much agree with McCloskey and Ziliak, in print and especially in private' (p. 47). 'In print' is impossible to reconcile with Elliott and Granger's (2004) highly critical assessment of Ziliak and McCloskey (2004), and it would make a mockery of Granger's actual practices as an econometrician. Again, to take one of many examples: '[Ziliak and McCloskey's claim that] 'signficance testing as used has no theoretical justification' is a case of throwing out the baby with the bathwater' (Elliott and Granger 2004, p. 548). If it were true that Granger said one thing in public and another in private, then he would be quite the hypocrite – but that is at odds with our experience of this distinguished scholar. McCloskey and Ziliak's charges of the hypocrisy of other economists are explicit on pp. 39–40, but the quotations are offered without attribution.

Ziliak and McCloskey complain about our 'hot tone.' We think that our tone was one of cool engagement and that the 'heat' in our paper mostly comes from quoting Ziliak and McCloskey. How can our calling a reading 'wooden' compare to the heat of McCloskey's words, quoted in our original paper, referring to most of the work in economics journals as 'unscientific rubbish' or suggesting that economists are beset by 'spreading, ramifying, hideous sins'? Or yet, how can it compare to the words in their reply: econometricians are 'scientific mice' or 'merely self-satisfied – after all, they control the journals and the appointments' (p. 39)? Ziliak and McCloskey are thrown into high dudgeon over our description of McCloskey (2002) as a 'tract.' The endpaper of that – let us say, to remain quite neutral – 'work' describes the series in which it appears as publishing 'challenging and sometimes outrageous pamphlets.' *The Concise Oxford Dictionary* defines *tract*: 'short treatise or discourse or pamphlet esp. on religious subject.' *Roget's Thesaurus* concurs, listing 'tract' as among the various synonyms for 'pamphlet.' Famously, Keynes titled one of his books *A Tract on Monetary Reform*. Who could have guessed that such venerable reference works provided such inflammatory advice or that the sainted Lord Keynes stood such a poor example of measured writing?

But in truth, we do not really object to a certain piquant prose. We do object to Ziliak and McCloskey conjuring insults and suggesting lapses of appropriate scholarly behavior where there are none. Ziliak and McCloskey wonder at the source

of our 'McCloskey-itis' (p. 40). Easily answered: We cite 17 works of McCloskey; only three (18%) are co-authored with Ziliak; eight (47%) predate Ziliak and McCloskey's first joint publication. And the central ideas that we criticize are all found in McCloskey's single-authored work. We were careful to refer to both Ziliak and McCloskey any time their *joint* work was in question. It would be poor scholarship to attribute to Ziliak ideas from McCloskey's single-authored works, simply because he later became a co-author. The depth of their collaboration on *The Cult of Significance* is irrelevant, as the book was neither published nor available to us at the time we wrote our paper or this rejoinder.

Ziliak and McCloskey also suggest something unsavory or nefarious in our pointing out that Deirdre was once Donald. McCloskey's personal story has no part in a paper on econometrics – nor did *we* give it one. The fact is that seven of the 15 works of McCloskey that we cite are published (and catalogued and indexed) as Donald, and eight as Deirdre. Scholarship demands that we note the identity, lest the reader wonder that Donald might be Deirdre's brother, father, uncle, or some other relation – or no relation at all. In other contexts, we might need to point out the identity of Cassius Clay with Muhammad Ali or Gaius Octavianus with Caesar Augustus. There is a certain self-absorption in McCloskey's thought that her personal story is so widely known that such pointers should be omitted.

3 An unanswered case

The second part of our argument was devoted to the evidence of their surveys in support of the charge that economists systematically confuse economic and statistical significance. We made four points.

First, their individual survey questions are badly formulated to answer the ostensible question, since some reflect good practices not logically connected to the distinction between economic and statistical significance, some reflect practices that are good in some contexts and not others, and some reflect bad practices. Ziliak and McCloskey do not controvert our observation, nor do they defend any of the questions that we challenge.

Second, scoring economists by simple summation of scores on individual questions, especially when, given the uncontroverted first point that the questions are not all germane to the object of the survey and given that some questions are not independent, results in an arbitrarily weighted – and, indeed, meaningless – score. Ziliak and McCloskey (p. 45) answer as if we had raised a simple index number problem and proposed using one question as the alternative to their sum: 'Does that mean that it's better to measure auto theft alone when looking into crime?' Indeed not; but neither should one form a crime index as auto thefts + auto thefts before noon + murders + traffic stops + contributions to charity + number of times guns are fired, which (see the first point) gives a pretty fair analogy to the kind of index that they construct.

Third, good practice in subjective scoring requires training and calibration to maintain consistency among observations and between the different surveys. Appropriate techniques are well known in non-economic social sciences. Yet, Ziliak and McCloskey offer no evidence in their papers (or in reply to direct, personal inquiries) that appropriate procedures have been followed. Their only reply is off point: they misrepresent our point to be an objection to subjectivity, when it was clearly an objection to their failure to apply procedures that insure consistency and comparability among subjective judgments.

Fourth, and related to the third point, they provide no protocols that make it publicly comprehensible what rules govern the mapping from the surveyed papers to the scores. They do nothing in their reply to assure us that any replicable procedures were followed. Instead, they personalize the issue. We reject their interpretation of our email correspondence as partial and misleading. But the facts of that correspondence are logically irrelevant to our point: they give the reader no articulated basis for understanding what the scores on individual questions really mean relative to individual papers.[4]

The only insight they provide is found in their comments on individual papers. We have already examined a selection of these in detail; there is no need to repeat it. Any reader who doubts our interpretations should simply look at the original papers and compare what we say to what Ziliak and McCloskey say. Ziliak and McCloskey's reply mainly repeats their previous interpretation without really addressing ours.

Yet, they do add a twist: they claim that they did not assert that economists ignore the magnitude of their coefficients but 'that the typical economist doesn't care about magnitudes *when "formally," "statistically" testing and deciding upon the importance of a variable*. For that job the economist strongly tends to substitute fit for oomph …' (p. 40). But this reformulation is just a shell game, obfuscating their own claim with the notion of 'formal,' 'statistical' tests that involve coefficient magnitudes, when they have never suggested what such tests might look like. (Compare this new formulation to the unobfuscated version in McCloskey and Ziliak 1996, p. 97, cited in our original paper: 'a difference can be permanent … without being "significant" in other senses … [a]nd … significant for science or policy and yet be insignificant statistically …'). Even here they advise that one should act on big coefficients ('oomph') despite large uncertainty – indeed one should act on coefficients so noisily measured that the magnitudes and even the signs of the coefficients are subject to great uncertainty. We disagree, and they call this begging the question. In fact, we have not assumed the thing to be proved. We gave a detailed analysis of why this is a bad idea (the 'licorice-jelly-bean mistake'), to which they simply do not reply. That we assume this result when discussing individual papers is no more begging the question than assuming (without re-proving) trigonometry when constructing maps.

4 Fish story

Ziliak and McCloskey are after big fish: 'We economists will need to redo almost all the empirical and theoretical econometrics since Hotelling and Lawrence Klein and Trygve Haavelmo first spoke out loud and bold' (p. 41). In asserting this claim, Ziliak and McCloskey employ the old – and wildly popular – rhetorical device, the 'fish story.'

More than 20 years ago, McCloskey (1985, p. 139) wrote: 'It would be arrogant to suppose that one knew better than thousands of intelligent and honest economic scholars.' Yet, Ziliak and McCloskey's attempts to reform econometrics are clearly based in the supposition that they know better than their 'thousands upon thousands of significance-testing econometric colleagues,' who are characterized as 'scientific mice' for not engaging in debate with them (p. 39). Yet, we are characterized as acting in 'sweaty desperation,' having set ourselves the task 'of denying the obvious' (p. 44). If it is

all so obvious, then our econometric colleagues are worse than pusillanimous mice; they are either unintelligent or dishonest. We think that they are neither.

We also disagree with McCloskey's earlier formulation. There is nothing arrogant about pointing out mistakes, even when they are widespread. But one must make the case. We have argued that Ziliak and McCloskey have not made their case. Stripped of its irrelevancies (such as the spleen about 'the barons and baronesses at Cornell and Princeton'; p. 44) and the misrepresentations (for example, we nowhere advocate 'mechanical' application of tests of statistical significance nor do we assert that sampling error is the only – or even the most important – error to which empirical modeling and observation are subject), Ziliak and McCloskey's reply fails to offer non-fallacious counterargument to any of our points.

It is not that all is for the best in the best of all possible worlds. We are not Panglosses. There is plenty of work to be done by econometricians and economic methodologists. But Ziliak and McCloskey have not landed their fish; they haven't even set the hook.

Acknowledgements

We thank Thomas Mayer and Roger Backhouse for comments on an earlier draft.

Notes

1. And, we note, neither did Fisher: 'I am casting no contempt on [Neyman-Pearson] acceptance procedures, 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' (Fisher 1955, pp. 69–70).
2. The calculations are specious for several reasons. We offered the count of JSTOR hits merely to demonstrate that economists do, in fact, sometimes use confidence intervals, not as a measure of their importance. We offered other indicators of economists' use of confidence intervals, which Ziliak and McCloskey allow to pass without comment. It would be hard to use a JSTOR search to get at an accurate measure of the use of confidence intervals: All searches are subject to both type I and type II error, though some searches are relatively easy to formulate in a way that minimizes error: think of the difference between Googling 'Ebenezer Scrooge' and 'Jim Smith,' when one has particular individuals in mind. Identifying actual use of confidence intervals from keywords is one of the hard searches, as is identifying the relevant base, which is not all articles but all articles in which confidence intervals might be appropriate. Even for the inappropriate base that they choose, there is no need for Ziliak and McCloskey to make up the numbers: a glance at Table 2 of our original paper gives the base that they guess to be 25,000 actually to be 18,200. Using this base, confidence intervals show up in 10% of cases, rather than their 7% (and this is lower bound, since the possibility of type I error is high in this case). And if we take the base to be the number of articles falling in the statistical family plus those in the confidence family (9598+1788), then the lower bound is 16%. (Of course, there may be some double-counting here, correcting for which would lower the base and raise the lower bound.) Ziliak and McCloskey still regard this as low, since they say that '100% *should* be reporting confidence intervals' (p. 51). But that is not right for their base of *all* articles, roughly half of which are not empirical. But it is also not right if the logical foundations of tests of statistical significance are bankrupt, since, as we have shown (and Ziliak and McCloskey concede), confidence intervals rest on the same foundation. Either it does not *fundamentally* matter which we report or both confidence intervals and tests of statistical significance should be reported in *0%* of the articles.

3. It is also worth noting that Edgeworth 1907 presents observational data on the wasps and bees of exactly the sort that Ziliak and McCloskey deprecate as sample size N=1.
4. Anyone who *is* interested in the details of the email correspondence should contact Hoover (kd.hoover@duke.edu), who would be happy to provide a copy of the complete correspondence.

References

Edgeworth, F.Y. (1885), "Methods of Statistics," *Jubilee Volume of the Statistical Society,* Royal Statistical Society of Britain, pp. 181–217.

Edgeworth, F.Y. (1907), "Statistical Observations on Wasps and Bees," *Biometrika*, 5(4), 365–386.

Elliott, G., and Granger, C.W.J. (2004), "Evaluating Significance: Comments on 'Size Matters,'" *Journal of Socio-Economics*, 33(5), 547–550.

Fisher, R. (1955), "Statistical Methods and Scientific Induction," *Journal of the Royal Statistical Society,* Ser. B (Methodological), 17(1), 69–78.

Haavelmo, T. (1944), "The Probability Approach in Econometrics," *Econometrica*, 12(Supplement), iii–vi, 1–115.

Hoover, K.D., and Perez, S.J. (1999), "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search," *Econometrics Journal*, 2(2), 1–25.

Hoover, K.D., and Perez, S.J. (2004), "Truth and Robustness in Cross Country Growth Regressions," *Oxford Bulletin of Economics and Statistics*, 66(5), 765–798.

Horowitz, J.L. (2004), "Comments on 'Size Matters,'" *Journal of Socio-Economics*, 3(5), 551–554.

Johnston, J. (1972), *Econometric Methods* (2nd ed.), New York: McGraw-Hill.

Kruskal, W. (1980), "The Significance of Fisher: A Review of *R.A. Fisher: The Life of a Scientist*" [by Joan Fisher Box]," *Journal of the American Statistical Association*, 5(372), 1019–1030.

McCloskey, D.N. (1985), *The Rhetoric of Economics* (1st ed.), Madison, WI: University of Wisconsin Press.

McCloskey, D.N. (1998), *The Rhetoric of Economics* (2nd ed.), Madison, WI: University of Wisconsin Press.

McCloskey, D.N. (1999), "Other Things Equal: Cassandra's Open Letter to Her Economist Colleagues," *Eastern Economic Journal*, 25(3), 357–363.

McCloskey, D.N. (2002), *The Secret Sins of Economics*, Chicago, IL: Prickly Paradigm Press. www.prickly-paradigm.com/paradigm4.pdf.

McCloskey, D.N., and Ziliak, S.T. (1996), "The Standard Error of Regressions," *Journal of Economic Literature*, 34(1), 97–114.

Romer, C.D. (1986), "Is Stablization of the Postwar Economy a Figment of the Data?," *American Economic Review*, 76(3), 314–334.

Ziliak, S.T., and McCloskey, D.N. (2004), "Size Matters: The Standard Error of Regressions in the *American Economic Review*," *Journal of Socio-Economics*, 33(5), 527–546 (also published in *Econ Journal Watch*, 1(2), 331–358. www.econjournalwatch.org/main/index.php?issues_id=3).