Statistical Significance Testing in Economics

William Peden[a] and Jan Sprenger[b]

[a]Erasmus School of Philosophy, Bayle Building, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands;
[b]Center for Logic, Language and Cognition (LLC), Department of Philosophy and Education, Università degli Studi di Torino (University of Turin), Palazzo Nuovo, Via Sant'Ottavio 20, 10124 Torino, Italy.

ARTICLE HISTORY
Compiled October 26, 2021

KEYWORDS
econometrics; economic methodology; Tinbergen debates; significance testing; statistical inference

1. Introduction

The origins of testing scientific models with statistical techniques go back to 18th century mathematics. However, the modern theory of statistical testing was primarily developed through the work of Sir R.A. Fisher, Jerzy Neyman, and Egon Pearson in the inter-war period. Some of Fisher's papers on testing were published in economics journals (Fisher, 1923, 1935) and exerted a notable influence on the discipline. The development of econometrics and the rise of quantitative economic models in the mid-20th century made statistical significance testing a commonplace, albeit controversial tool within economics.

In the debate about significance testing, methodological controversies intertwine with epistemological issues and sociological developments. Our aim in this chapter is to expound these connections and to show how the use of, and the debate about, significance testing in economics differs from other social sciences, such as psychology.

Section 2 explains the basic principles of statistical significance testing with particular attention to its use in economics. In Section 3, we sketch how significance testing got entrenched in economics, and we highlight some early criticisms. Section 4 deals with Ziliak and McCloskey's criticism that significance tests, and the economists who apply them, confuse statistical and proper economic significance (i.e., effect size). Section 5 relates significance testing to the problem of publication bias in science and compares the debates about significance testing in economics and psychology. Section 6 wraps up and briefly discusses some suggestions for methodological improvement.

2. Statistical Significance Testing

There are two grand traditions of interpreting statistical testing procedures. Jerzy Neyman and Egon Pearson's behavioral, *decision-theoretic approach* contrasts two competing hypotheses, the "null" hypothesis H_0 and the "alternative" H_1, and conceptualizes statistical tests as an instrument for controlling the rate of erroneous decisions in rejecting or accepting one of them. Although such decision-theroetic formalisms have a natural affinity to rational choice theory and economic reasoning, econometric testing generally follows R.A. Fisher's *evidential approach* (Fisher, 3574, 1956) based on an interpretation of low p-values as evidence against the tested hypothesis.

Fisher's approach shall now be expounded. For Fisher, the purpose of statistical analysis consists in assessing the relation of a null hypothesis H_0 to a body of observed data. That hypothesis usually stands for there being no effect of interest, no causal relationship between two variables, or simply for a scientific default assumption. For example, suppose we run a simple linear regression model $y_i = \alpha + \beta\, x_i + \epsilon_i$ (with i.i.d.[1] error terms ϵ_i), where the data points x_i and y_i are paired realizations of the quantities of interest X and Y. A possible null hypothesis would be $H_0 : \beta = 0$, claiming that there is no systematic relationship between the variables X and Y. McCloskey and Ziliak (1996, 98) give the example of a regression analysis for purchasing power, comparing prices of goods abroad to those at home: $\beta = 1$ (and $\alpha = 0$) would then express perfect match between home prices (X) and abroad prices (Y), modulo random error.

Testing such null hypotheses for compatibility with the data is called a **null hypothesis significance test (NHST)**. The alternative hypothesis H_1 is typically unspecific and expresses the presence of an effect that differs from the hypothesized null effect (e.g., $H_1 : \beta \neq 0$ if the null is $H_0 : \beta = 0$). Now, the basic rationale of significance tests is that results that fall into the extreme tails of the probability distribution postulated by the null hypothesis H_0 compromise its tenability: "either an exceptionally rare chance has occurred, or the theory [=the null hypothesis] is not true." (Fisher, 1956, 39). The occurrence of such an exceptionally rare chance has both epistemological and practical consequences: first, the null hypothesis is rendered "objectively incredible" (Spielman, 1974, 214); second, the null should be treated as if it were false.

Notably, Fisher's ideas are close to Popper's falsificationism (Popper, 2002). They both agree the only purpose of an experiment is to "give the facts a chance of disproving the null hypothesis" (Fisher, 3574, 16). They also agree that failure to reject a hypothesis does not conclude positive evidence for the tested (null) hypothesis. However, while Popper gives a negative characterization of scientific knowledge, Fisher uses statistical tests for a positive account of experimental knowledge: the existence of causal effects can be *demonstrated* experimentally by means of (statistically) rejecting the hypothesis that the observed effect occurred due to chance.

The central concept of modern significance tests—the p-value—is now illustrated in a two-sided testing problem. Suppose we want to infer whether the mean θ of an unknown distribution is significantly different from the null value $H_0 : \theta = \theta_0$. We observe data $x := (x_1, \ldots, x_N)$, corresponding to N i.i.d. realizations of an experiment with (unknown) population mean θ and (known) variance σ^2. Then, one measures the discrepancy in the data x with respect to the postulated mean value θ_0 using the

[1] For convenience, "independent and identically distributed" will always be abbreviated as "i.i.d.".

standardized statistic

$$z(x) := \frac{\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0}{\sqrt{N \cdot \sigma^2}} \tag{1}$$

We may re-interpret equation (1) as

$$z = \frac{\text{observed effect} - \text{hypothesized effect}}{\text{standard error}} \tag{2}$$

The p-value then depends on the probability distribution of z given the null hypothesis:

$$p := p_{H_0}(|z(X)| \geq |z(x)|), \tag{3}$$

or in other words, the p-value describes the probability of observing a more extreme discrepancy under the null than the actually observed one. The lower the p-value, that is, the more the observed effect diverges from the effect postulated by the null hypothesis, the less does the null hypothesis explain the data.

Such p-values or "observed significance levels" play a large role in econometrics: they serve as an indicator of whether a finding is noteworthy, interesting, and ultimately publishable. Moreover, the conventional classification into various levels of significance is used for annotating correlation tables and for making them more readable: one asterisk behind an entry corresponds to $p < .05$ ("significant"), two asterisks to $p < .01$ ("highly significant"), and three asterisks to $p < .001$ ("very highly significant"). It is precisely this suggestive annotation practice that has attracted trenchant criticism in the last decades, since it obliterates the differences between statistical and genuine scientific significance. Before entering this debate, however, we first review the history of significance testing in econometrics, and some early debates. Specifically, we illustrate how the methodological controversies raised by significance testing are philosophically interesting, and that issues in the philosophy of statistics are consequential for important issues in economics.

3. Early Debates about Significance Testing in Econometrics

3.1. Econometrics and The Tinbergen Debates

We start with an important distinction between *economic statistics* and *econometrics*. Economic statistics consists of gathering and formulating descriptive statistical information about economic facts. By contrast, econometrics is the use of inferential statistics to formulate answers to theoretical questions, such as "Are interest rates pro-cyclical?" or "Is the fiscal multiplier greater than unity?"

The Tinbergen Debates occurred quite early in the history of econometrics. Jan Tinbergen was a Dutch economist who had already constructed the first econometric model of the business cycle in 1936. In a study for the League of Nations, Tinbergen sought both to popularise his approach to econometric research, to apply it to modelling actual economies, and to test a theory of the business cycle with these data. His approach to statistics was typical of the spirit of the day: econometrics cannot not prove economic theories right, but it could prove that a theory is incorrect (Tinbergen, 1939, 12). Tinbergen applied a large number of goodness-of-fit tests of statistical significance to the equations of his models (Morgan, 1990, 108–120). His approach was

sophisticated, and he might easily have expected a positive reaction from prominent contemporary business cycle theorists.

He would have been wrong. John Maynard Keynes (1939) scathingly criticised Tinbergen's research and stated that the results of Tinbergen's statistical work "probably have no value" (Keynes, 1939, 559). Keynes's principal objections were that Tinbergen's work failed to meet some requirements that Keynes considered to be vital: (1) full knowledge of the causally relevant factors; (2) these causal factors must be measurable and mutually independent; (3) the relationships must be linear; (4) one must know the relevant time lags and trends. Keynes thought that these conditions were more or less never met in economics, so he saw practically no role for significance testing by econometricians. This criticism led to a brief series of exchanges between Tinbergen (1940a,b)[2], and Keynes (1940), with other economists joining. Mary S. Morgan (1990, 123–124) argues that an underlying epistemological issue in the Tinbergen Debates was a disagreement between Tinbergen and Keynes on the potential function of statistical testing: for Keynes, its only possible role was to identify the strength of factors within a causal framework that had already been developed by theoretical analysis; for Tinbergen, statistical testing could also inform—though not determine—the theoretical analysis. This antagonism resurfaces in later critiques of significance testing by Ziliak and McCloskey.

There were other methodological debates that arose out of Tinbergen's work in the late 1930s (e.g., Haavelmo, 1940). For instance, in a 1940 letter, Oskar Lange argued that Tinbergen's results lacked "statistical significance" because he had failed to take serial correlation (i.e., autocorrelation, a positive or negative association between a variable and its future/past values) into account and serial correlation is exactly one of the reasons why we want a business cycle model in the first place. Hence, Lange thought that Tinbergen's methods were either flawed or redundant (Louçã, 2007; Orcutt and Irwin, 1948). This criticism illustrates how some economists in the 1940s used "statistical significance" to refer to an *epistemic achievement*, rather than its contemporary sense of controlling error rates or p-values below a certain level (typically 0.05).

Another example was the reaction of a young Milton Friedman:

> Tinbergen's results cannot be judged by ordinary tests of statistical significance. The reason is that the variables with which he winds up, the particular series measuring these variables, the leads and lags, and various other aspects of the equations besides the particular values of the parameters (which alone can be tested by the usual statistical technique) have been selected after an extensive process of trial and error *because* they yield high coefficients of correlation [with that sample data]. (Friedman, 1940, 659)

Here, Friedman is contending that statistical significance tests of an economic model are only appropriate if they are out-of-sample tests[3] Unlike Keynes, Friedman thought that significance testing had a potential function within theory choice. However, he regarded work like Tinbergen's as only useful for identifying whether models are worth testing further, with more data (Friedman, 1940, p.660). In Hans Reichenbach's terminology, Friedman saw in-sample testing as limited to the "context of discovery", where we *develop* economic theories, and not the "context of justification" where we *evaluate* theories on the basis on their theoretical and empirical merits (Reichenbach, 1938). The methodological value of out-of-sample testing continues to be debated in

[2]Anticipating some later comments about statistical significance and economic significance, Tinbergen (1940b, 143–145) distinguishes between "statistical independence"—the absence of a correlation between variables—and "economic independence"—a type of *causal* independence.

[3]That is, the models are not tested with the same data used to estimate their parameters.

economics (e.g., Gelfond and Murphy, 2016).

3.2. The "Con" in Econometrics

Economists were always somewhat sceptical of post-war econometrics. However, the end of post-war economic stability in the 1970s and poor performance of econometric models during that decade encouraged intense methodological reflections (Hayek, 1989; Hendry, 1980; Lucas, 1976; Sims, 1980). One of the most important was Edward Leamer's "Let's Take the Con Out of Econometrics" (Leamer, 1983). Leamer covers many methodological issues, but his general critique is that significance testing in econometrics normally involves a large number of often unrealistic assumptions, e.g. that the model's error terms are uncorrelated. Due to the influential instrumentalist manifesto of Milton Friedman (1953), economists are generally comfortable with unrealistic assumptions. However, if the data are relevant to the model *only in conjunction with unrealistic assumptions*, then these assumptions cannot be regarded any more as useful idealizations, and be used as such in statistical testing. Leamer grants that the assumptions may be *approximately true*, but the results of the test then depend on their exact details. Therefore, Leamer recommends (and helped develop) sensitivity analysis: given that economic theory or background knowledge cannot uniquely specify the statistical assumptions to make in our tests, we should report the consequences of adopting various assumptions. This will help identify test results that are very sensitive to particular assumptions, and increase the robustness of our statistical practices. Although Leamer's paper has been cited thousands of times, he is unimpressed by subsequent efforts towards more robust significance testing in econometrics (Leamer, 2010).

Leamer also advocates a Subjective Bayesian methodology over the (alleged) objectivity of classical significance testing. One of his reasons is that even sophisticated econometric models can easily be tested using modern computers. This makes some types of statistical malpractice much easier (Leamer, 1983, 36–37). Before the 1970s, testing complex econometric models was extraordinarily difficult or even impossible. This changed in the early 1970s, transforming econometric practice. While this transformation was useful in many ways, it also created methodological hazards, as Leamer noted: computational limitations had been a partial shield against p-*hacking*, that is, the use of questionable research practices (e.g., selective reporting of studies, removing outliers, adding further covariates, etc.) in order to obtain a statistically significant result (i.e., $p < .05$). In a context where a test could take weeks or months, even with the aid of computers, statistically significant results were genuinely surprising and systematic p-hacking was not feasible. Modern computers can carry out analogous tests in seconds, so statistically significant results are less surprising for economists. There is also some empirical evidence of p-hacking in economics (Brodeur et al., 2016). Strategies to counteract this danger are discussed in Section 5.

4. The Cult of Statistical Significance? Effect size vs. p-values

Another forceful criticism of significance tests concerns their relation to effect size, and the confusion between p-values and proper effect size measures. The economists Deirdre McCloskey and Stephen Ziliak (henceforth, ZMC) have made this point in a series of papers and books (McCloskey and Ziliak, 1996; Ziliak and McCloskey, 2004, 2008). We illustrate the difference with one of their favorite examples (Ziliak

and McCloskey, 2008, ch. 1). Assume that we have to choose between two diet cures, based on pill A and pill B. Pill A makes us lose 10 pounds on average, with an average variation of 5 pounds.[4] Pill B makes us lose 3 pounds on average, with an average variation of 1 pound. Which one leads to more significant loss? Naturally, we opt for pill A because the effect of the cure is so much larger.

However, if we translate the example back into significance testing, the order is reversed. Assume the standard deviations are known for either pill. Compared to the null hypothesis of no effect at all, observing a three pounds weight loss after taking pill B is a more significant result evidence for the efficacy of that cure than observing a ten pounds weight loss after taking pill A:

$$z_A(10) = \frac{10-0}{5} = 2 \qquad\qquad z_B(3) = \frac{3-0}{1} = 3$$

Thus, there is a notable discrepancy between our intuitive judgment and the one suggested by the p-values. This occurs because statistical significance is supposed to be "a measure of the strength of the signal relative to background noise" (Hoover and Siegler, 2008, 58). On this score, pill B indeed performs better than pill A. According to Ziliak and McCloskey, however, economists and policy-makers are primarily interested in the effect size, not the signal/noise ratio: they do not want to ascertain the presence of *some* effect, but to demonstrate a *substantial* effect. Due to the importance of the latter for practical decisions and economic policy—Ziliak and McCloskey call this "policy oomph"— we should actually focus on *practically meaningful effect sizes* rather than significance levels. The former can be measured by standardized regression coefficients, or specific statistics such as Cohen's d or Pearson's r^2 (for the strength of correlations), but not by p-values. An effect need not be statistically significant to be big and remarkable (like pill A), and a statistically significant effect can be quite small and uninteresting (like pill B).

This fundamental difference is, however, frequently neglected: in practice, the level of significance acts often as a cue to scientific importance (compare Cohen, 1994, for a similar diagnosis in psychology). By scrutinizing the statistical practice in the top journal *American Economic Review*, as well as by surveying the opinion of economists on the meaning of statistical significance, McCloskey and Ziliak (1996) derive the conclusion that statistical concepts and tools are frequently abused. For example, researchers tend to confound economic with statistical significance (70% of the papers do not make an explicit distinction), relate effect sizes to the scientific context (72%), or engage in "sign econometrics", where the sign, but not the size of a coefficient is commented on (53%). Worse still, there is no visible improvement over time. According to Ziliak and McCloskey, the discussion in statistics and methodology journals has done nothing to little to alleviate the problems since the proportion of articles with questionable use of significance tests has not decreased over time (Ziliak and McCloskey, 2004, 2008).

In total, ZMC's critique of significance testing can be summarized as: (1) significance tests encourage reasoning fallacies (e.g., statistical significance = null hypothesis refuted or effect economically meaningful); (2) they are not a suitable tool for economic analysis (because they do not aim at effect size or "policy oomph"); (3) they give rise to a culture of mindless use of statistical inference without proper considerations of economic and policy implications (see also Gigerenzer, 2004). Few statistics or econometric textbooks highlight this difference, and thus prevent the proper appreciation

[4]The concept of "average variation" is intuitively explicated as the statistical concept of standard deviation, which is, for a random variable X, defined as $\sqrt{E[(X - E(X))^2]}$.

of the limits of significance tests in upcoming generations.

The echo of Ziliak and McCloskey's work in the economic community and beyond was mainly positive, but not exclusively so. For eminent statisticians like Arnold Zellner or methodologists like Nathan Berg (2004), Edward Leamer (2004) and Bruce Thompson (2004), the critique of (the mindless use of) NHSTs fits into a broader project of changing and improving scientific method. Even if they don't all agree about the right direction for the future, they agree that NHSTs are flawed and need to be replaced. A second group agrees with ZMC on their central points, but nuance their criticism. For example, Joel Horowitz (2004) observes that significance testing with all its problems may well be inevitable when we want to test whether an economic model is well-specified, or when we are interested in the *existence* of an effect rather than its magnitude. Finally, a third group (e.g., Elliott and Granger, 2004; Hoover and Siegler, 2008) defends NHSTs and argues (1) that testing theories is inevitable; (2) there is a need for non-subjective method of theory testing and NHSTs provide it; (3) ZMC's focus on effect size is not without problems (e.g., it does not adequately quantify the uncertainty of the estiamte); (4) misuse of procedures is not unique to NHSTs, but occurs in all parts of statistical reasoning. That latter point is also shared by Gigerenzer (2004) who warns that shifting to another framework (e.g., Bayesian reasoning) may just lead to a reiteration of mindless statistical inferences with different tools.[5]

5. New Challenges: Publication Bias, the Replication Crisis, and Comparison with Psychology

The last years have seen an increasing distrust in scientific findings due to concerns about publication bias and lack of replicability. Publication bias means bias in what is published with respect to what is researched, and it is related to significance tests since they often act as a gatekeeper for the whether a finding is "interesting" or "publishable". The standard approach to significance tests makes it hard to interpret non-significant results, and to draw any substantive inference from them. By contrast, as shown by Ziliak and McCloskey, it is easy to lure oneself into identifying a statistically significant result with an important scientific finding. While this mechanism of suppressing non-significant results, called the *file drawer effect*, and its impact on the published literature, has been identified long ago in theoretical models (e.g., Ioannidis, 2005; Rosenthal, 1979; Rozeboom, 1960; Sterling, 1959), it has been ignored by economists for a long time, especially by those not involved in methodological debates. Recently, the severity of the problem has been demonstrated by systematic replication projects for experimental findings in psychology, medicine and economics, revealing a disappointingly low replication rate in all featured disciplines (e.g., Collaboration 2015 for psychology and Camerer et al. 2016 for experimental economics).

The problem is not only that significance tests lead to a depreciation of non-significant findings (even if they are methodologically sound), but that researchers often use questionable research methods, such as selective reporting of results, adding covariates, eliminating outliers, etc., in order to obtain a significant (and therefore publishable) finding. Such *p-hacking* is easy to achieve with modern computation tools. Meta-analytic techniques such as funnel plots and *p*-curves have provided evidence

[5]However, surveys of statistical research practice reveal that fallacious interpretations of inference procedures are particularly frequent for NHSTs as compared to confidence intervals or Bayesian inference (e.g., Cumming, 2012; Fidler, 2005).

of publication bias in various research areas (Simonsohn et al., 2014; Weiß and Wagner, 2011); the file drawer effect and p-hacking are plausible causes of these findings. Specifically in economics, Brodeur et al. (2016) have found a two-humped distribution of p-values, suggesting not only an excess of just significant results, but also that researchers try to "work away" ambiguous significant p-values either toward significance or toward clear non-significance. Crucially, one need not assume mischievous or badly trained researchers for explaining such findings—complex data analysis problems require many judgment calls and researchers may be unconsciously influenced by their own biases and incentives when making such decisions.

Olken (2015) discusses compulsory preregistration of the data analysis plan (i.e., before collecting or analyzing the data) as an antidote to p-hacking and *HARKing*—that is, hypothesizing after the results are known, and relabeling exploratory as confirmatory research. Researchers complying with such a plan would decide in advance on primary and secondary outcome variables, measurement scales, statistical tests, covariates, and so on, as to minimize the potential for p-hacking. However, Olken's judgment is mixed: while such plans may be both efficient and feasible for controlled trials in psychology or medicine, econometric data analysis typically deals with complex datasets, a high number of secondary outcome variables, and aims at unraveling hidden theoretical mechanisms. The tradeoff between unbiasedness, efficiency and nuance implied by compulsory preregistration may be non-trivial in economics. Finally, the Registered Reports model, where the publication of an article is decided on the basis of the research question (e.g., Chambers, 2013), the experimental design and the data analysis plan, may helpful to counter the file drawer effect and publication bias for disciplines where simply structured controlled trials are the norm (e.g., medicine, psychology). However, it is is not easily transferable to the type of (observational) data analysis that economists typically undertake.

Compared to economists, psychologists worry more about flawed incentives in their discipline and the shortcomings of significance tests (e.g., Bakker et al., 2012; Cumming, 2012; Schmidt, 1996). They have a tradition of critical reflection on NHST and relating it to philosophical questions that is almost absent in economics (e.g., Cohen, 1994; Meehl, 1967). However, also in psychology, the scathing criticisms of significance tests have *not* led to an abolition of that practice, even if significance levels are now routinely accompanied by standard errors and effect size estimates. Where does this inertia come from? Why does the persuasive force of pro-reform arguments not result in real change? Martin Altman (2004) and Bruce Thompson (2004) identify various sources, which may also play a role in economics. The problem is—like in politics—not a lack of awareness or knowledge distribution, but a lack of willingness to implement them at anything faster than an excruciatingly low pace, if at all. Barriers to change may be the time delay caused by the bureaucratic constraints on procedures within large professional associations such as the AEA and the APA, fear of the cognitive dissonance that would result from giving up a practice that one has followed for ages, and lack of commitment on behalf of key figures such as senior practitioners and editors-in-chief of major journals who could "force" authors to abide by more sophisticated procedures. We are curious whether the sense of urgency created by the replication crisis and the evidence of p-hacking and publication crisis leads to genuine methodological reform and to more sustainable statistical practice.

6. Conclusion

Significance testing in economics has been hotly contested from its very beginnings. However, the focal points of the criticism have shifted. Early criticisms, like the ones found in the Tinbergen-Keynes-Friedman debate, sustained that the assumptions of significance tests are seldom, if ever, met for economic data. While unrealistic assumptions are common in economic modeling, they invalidate significance tests, according to critics, when they are the crucial nexus between the tested theoretical model and the data.

These criticisms did not prevent significance testing from becoming an highly influential and widespread tool in economics, and econometrics in particular. And widespread use came with frequent misuse, especially when p-values became standard measures of statistical evidence. Critics like Ziliak and McCloskey object that statistical significance levels are no good indicators of actual relevance for economic policy decisions (because the p-value is by itself not an indicator of effect size). According to ZMC, this confusion has created a lot of damage to economic science and society as a whole. Recently, the debate about significance testing has got a new twist due to increasing evidence of p-hacking and questionable research practices, and doubts about the replicability and trustworthiness of research findings. These phenomena were first identified in other sciences such as psychology and medicine, but they are a problem for economics, too.

Responses to these challenges are grouped into three categories (e.g., Romero, 2019): (1) *statistical reform*, such as use of confidence intervals and Bayesian models, (2) *methodological reform*, such as pre-registration of experiments and data analysis plans, and (3) *social reform*, such as setting up crediting researchers also with confirmatory research and non-significant findings. Which combination of these three approaches is the best reply to save the reliability of significance testing in economics is an exciting question for future research.

References

Altman, M. (2004). Statistical significance, path dependency, and the culture of journal publication. *The Journal of Socio-Economics*, 33(5):651 – 663.

Bakker, M., Wicherts, J., and van Dijk, A. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, 7:543–554.

Berg, N. (2004). No-decision classification: an alternative to testing for statistical significance. *The Journal of Socio-Economics*, 33(5):631 – 650.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49:609–610.

Cohen, J. (1994). The Earth is Round ($p < .05$). *Psychological Review*, 49:997–1003.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349. Retrieved from http://science.sciencemag.org/content/349/6251/aac4716.full.pdf.

Cumming, G. (2012). *Understanding the New Statistics*. Routledge, New York.

Elliott, G. and Granger, C. W. (2004). Evaluating significance: comments on "size matters".

The Journal of Socio-Economics, 33(5):547 – 550.

Fidler, F. (2005). *From Statistical Significance to Effect Estimation: Statistical Reform in Psychology, Medicine and Ecology.*

Fisher, R. (1923). Statistical tests of agreement between observation and hypothesis. *Economica*, (8):139–147.

Fisher, R. (1935). The mathematical distributions used in the common tests of significance. *Econometrica*, 3(4):353–365.

Fisher, R. A. (1935/74). *The Design of Experiments.* Hafner Press, New York. Reprint of the ninth edition from 1971. Originally published in 1935 (Edinburgh: Oliver & Boyd).

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Hafner, New York.

Friedman, M. (1940). Review of "Business Cycles in the United States of America, 1919-1932" by J. Tinbergen. *American Economic Review*, 30(3):657–660.

Friedman, M. (1953). *Essays in positive economics.* University of Chicago Press., Chicago.

Gelfond, R. and Murphy, R. H. (2016). A call for out-of-sample testing in macroeconomics. *Libertas: Segunda Epoca*, 1(1):1–1.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5):587 – 606.

Haavelmo, T. (1940). The inadequacy of testing dynamic theory by comparing theoretical solutions and observed cycle. *Econometrica*, 8(4):312–321.

Hayek, F. v. (1989). The pretence of knowledge (nobel lecture). *American Economic Review*, 79(6):3–7.

Hendry, D. F. (1980). Econometrics—alchemy or science? *Economica*, 47(188):387–406.

Hoover, K. D. and Siegler, M. V. (2008). Sound and fury: Mccloskey and significance testing in economics. *Journal of Economic Methodology*, 15(1):1–37.

Horowitz, J. L. (2004). Comments on "size matters". *The Journal of Socio-Economics*, 33(5):551 – 554.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2.

Keynes, J. M. (1939). Professor Tinbergen's method. *The Economic Journal*, 49(195):558–577.

Keynes, J. M. (1940). On a method of statistical business-cycle research. A comment. *The Economic Journal*, 50(197):154–156.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1):31–43.

Leamer, E. E. (2004). Are the roads red? comments on "size matters". *The Journal of Socio-Economics*, 33(5):555 – 557.

Leamer, E. E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives*, 24(2):31–46.

Louçã, F. (2007). *The years of high econometrics: A short history of the generation that reinvented economics.* Routledge, London and New York.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46.

McCloskey, D. and Ziliak, S. (1996). The standard error of regressions. *Journal of Economic Literature*, 34(1):97–114.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34:103–115.

Morgan, M. (1990). *The history of econometric ideas.* Cambridge University Press, Cambridge.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80.

Orcutt, G. and Irwin, J. (1948). A study of the autoregressive nature of the time series used for tinbergen's model of the economic system of the united states, 1919–1932. *Journal of the Royal Statistical Society. Series B*, 10(1):1–53.

Popper, K. R. (1959/2002). *The Logic of Scientific Discovery.* Routledge, London. Reprint of the revised English 1959 edition. Originally published in German in 1934 as "Logik der Forschung".

Reichenbach, H. (1938). *Experience and prediction. An analysis of the foundations and the*

structure of knowledge. The University of Chicago Press, Chicago.

Romero, F. (2019). Philosophy of Science and the Replicability Crisis. *Philosophy Compass*, 14:e12633.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin*, 86(3):638–641.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological bulletin*, 57:416–442.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1:115–129.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). *P*-curve: A Key to the File Drawer. *Journal of Experimental Psychology: General*, 143:534–547.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48.

Spielman, S. (1974). The Logic of Tests of Significance. *Philosophy of Science*, 41(3):211–226.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, 54:30–34.

Thompson, B. (2004). The "significance" crisis in psychology and education. *The Journal of Socio-Economics*, 33(5):607 – 613.

Tinbergen, J. (1939). Statistical testing of business–cycle theories. Technical report, League of Nations Economic Intelligence Service., Geneva.

Tinbergen, J. (1940a). Econometric business cycle research. *Review of Economic Studies*, 7(2):73–90.

Tinbergen, J. (1940b). On a method of statistical business-cycle research. a reply. *The Economic Journal*, 50(197):141–154.

Weiß, B. and Wagner, M. (2011). The Identification and Prevention of Publication Bias in the Social Sciences and Economics. *Jahrbücher für Nationalokonomie und Statistik*, 231:661–684.

Ziliak, S. T. and McCloskey, D. N. (2004). Size matters: the standard error of regressions in the american economic review. *The Journal of Socio-Economics*, 33(5):527 – 546.

Ziliak, S. T. and McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press, Ann Arbor, Mich.