Twenty steps towards an adequate inferential interpretation of p-values in econometrics

(Forthcoming: Journal of Economics and Statistics)

Contributing authors

Prof. Dr. Norbert Hirschauer (Corresponding Author)

Martin Luther University Halle-Wittenberg, Faculty of Natural Sciences III

Institute of Agricultural and Nutritional Sciences, Chair of Agribusiness Management

Karl-Freiherr-von-Fritsch-Str. 4, D-06120 Halle (Saale)

norbert.hirschauer@landw.uni-halle.de

Dr. Sven Grüner

Martin Luther University Halle-Wittenberg, Faculty of Natural Sciences III

Institute of Agricultural and Nutritional Sciences, Chair of Agribusiness Management

Karl-Freiherr-von-Fritsch-Str. 4, D-06120 Halle (Saale)

sven.gruener@landw.uni-halle.de

Prof. Dr. Oliver Mußhoff

Georg August University Göttingen

Department for Agricultural Economics and Rural Development, Farm Management

Platz der Göttinger Sieben 5, D-37073 Göttingen

Oliver.Musshoff@agr.uni-goettingen.de

Prof. Dr. Claudia Becker

Martin Luther University Halle-Wittenberg, Faculty of Law and Economics

Institute of Business Studies, Chair of Statistics

Große Steinstraße 73, D-06099 Halle (Saale)

claudia.becker@wiwi.uni-halle.de

Twenty steps towards an adequate inferential interpretation of *p*-values in econometrics

Abstract: We suggest twenty immediately actionable steps to reduce widespread inferential errors related to "statistical significance testing." Our propositions refer first to the theoretical preconditions for using *p*-values. They furthermore include wording guidelines as well as structural and operative advice on how to present results, especially in research based on multiple regression analysis, the working horse of empirical economists. Our propositions aim at fostering the logical consistency of inferential arguments by avoiding false categorical reasoning. They are not aimed at dispensing with *p*-values or completely replacing frequentist approaches by Bayesian statistics.

Keywords: *p*-value, statistical inference, inverse probability error, multiple testing

1 Introduction

Widely scattered over time and disciplines, a vast amount of criticism regarding the misuses and misinterpretations of the *p*-value (the don'ts) as well as a large number of suggestions for reform (the do's) have accumulated. One might think that enough has been said on this subject in the meanwhile, both before and after the ASA-statement (Wasserstein and Lazar 2016) that prominently red-flagged misuses and inferential errors. But the problems seem to be here to stay, particularly in disciplines such as economics that heavily rely on multiple regression analysis. Two features of the frequentist null hypothesis significance testing (NHST) framework are at the origin of most errors: first, the dichotomization of results depending on whether the *p*-value is below or above some arbitrary threshold (usually 0.05). Second, the associated terminology that speaks of "hypothesis testing" and "statistically significant" as opposed to "statistically non-significant" results. Dichotomization in conjunction with misleading terminology propagate cognitive biases that seduce researchers to make logically inconsistent and overconfident inferences, both when *p* is below and when it is above the "significance" threshold. The following errors seem to be particularly widespread:[1]

1) use of *p*-values when there is neither random sampling nor randomization
2) confusion of statistical and practical significance or complete neglect of effect size
3) unwarranted binary statements of there being an effect as opposed to no effect, coming along with
   - misinterpretations of *p*-values below 0.05 as posterior probabilities of the null hypothesis
   - mixing up of estimating and testing and misinterpretation of "significant" results as evidence confirming the coefficients/effect sizes estimated from a single sample
   - treatment of "statistically non-significant" effects as being zero (confirmation of the null)
4) inflation of evidence caused by unconsidered multiple comparisons and *p*-hacking
5) inflation of effect sizes caused by considering "significant" results only

[1] See, for example, McCloskey and Ziliak (1996), Sellke et al. (2001), Ioannidis (2005), Ziliak and McCloskey (2008), Krämer (2011), Ioannidis and Doucouliagos (2013), Kline (2013), Colquhoun (2014), Gelman and Loken (2014), Motulsky (2014), Vogt et al. (2014), Gigerenzer and Marewski (2015), Greenland et al. (2016), Hirschauer et al. (2016; 2018), Wasserstein and Lazar (2016), Ziliak (2016), Amrhein et al. (2017), and Trafimow et al. (2018). This list contains but a small selection of the literature on *p*-value misconceptions from the last 20 years. Since we focus on pragmatic steps for reducing inferential errors in the future (the do's), we do not engage in an extensive review of the vast body of literature that has accumulated on the don'ts in the past.

The ASA-statement highlights that the *p*-value does not provide a good measure of evidence regarding a hypothesis. In other words, it does not provide a clear rationale or even calculus for statistical inference (Goodman 2008). While Berry (2017: 896) might be pushing too hard when claiming that a *p*-value "as such has no inferential content," one must recognize that it is but a *graded* measure of the strength of evidence against the null, but only in the sense that small *p*-values will occur more often if there is an effect compared to no effect (Hirschauer et al. 2018).[2] Joining Berry (2017), Gelman and Carlin (2017), Greenland (2017), McShane and Gal (2017), McShane et al. (2017), Trafimow et al. (2018), and many others, we believe that degrading the *p*-value's continuous message into binary "significance" declarations ("bright line rules") is at the heart of the problem.[3] Since the *p*-value is deeply anchored in the minds of most scientists including economists, we believe that demanding drastic changes, such as renouncing *p*-values or replacing frequentist approaches by Bayesian methods, is not the most promising way to guard researchers from the inferential errors that we see today. Dispensing with the dichotomy of significance testing but retaining the *p*-value and adopting small and manageable steps towards improvement seems to be more promising (Amrhein et al. 2017). Such steps will have to account for the idiosyncrasies of each scientific discipline. For example, requirements in the medical sciences, which often focus on mean differences between randomized experimental treatments, will at least partly differ from those in economics, which frequently resorts to multiple regression analysis of observational data.

Even if one is aware of the fundamental pitfalls of NHST, it is difficult to escape the categorical reasoning that is so entrancingly suggested by its dichotomous "significance" declarations.[4] Econometricians regularly face the challenge to provide an interpretative evaluation of numerous regression coefficients. Imagine a regression with several focal variables (predictors), a set of controls, and possibly even some secondary covariates (interaction terms, higher-order polynomials, etc.) introduced in the process of model specification. How should we evaluate and comment on the "evidence" as represent-

[2] The *p*-value is often per se not an informative measure of the strength of the evidence because the zero effect assumption is frequently an implausible "straw man" hypothesis from the very start. Leamer (1978: 89) pointedly noted that since "a large sample is presumably more informative than a small one, and since it is apparently the case that we will reject the null hypothesis in a sufficiently large sample, we might as well begin by rejecting the hypothesis and not sample at all.

[3] Gelman and Stern (2006: 328) pointed out that the problem with dichotomous interpretations is not merely that nothing substantial happens when moving from a *p*-value of let's say 0.049 to 0.051. Rather, "even large changes in significance levels can correspond to small, nonsignificant changes in the underlying quantities."

[4] Misunderstandings may be due to the fact that few users of NHST seem to be familiar with the differing perspectives of Fisher (1925) and Neyman and Pearson (1933) that have been amalgamated into what is today known as NHST (Ziliak and McCloskey 2008). In statistical decision theory inherited from Neyman and Pearson, the "world" (formally, the parameter space) is indeed divided into two mutually exclusive states – represented by null and the alternative hypothesis – between which a *decision* has to be made. The dichotomous setting results from assuming a 0/1 loss function and from constructing optimal decision rules that minimize the expected loss of a false decision (e.g., Lehmann and Romano 2010: 56ff.). While Fisher understood "significance testing" as a tool for inductive reasoning, statistical decision theory in the Neyman-Pearson tradition rejects this idea. Instead, it uses "hypothesis testing" to identify rules of *behavior* which "insure that in the long run of experience, we shall not be too often wrong" (Neyman and Pearson 1933: 291). As Berry (2017: 895) pointedly notes: "[…] believing or advertising something as true and acting as though it is true are very different kettles of fish." It therefore seems that the assumptions of statistical decision theory are ill-fitted to econometric research whose regular objective is to investigate relationships that link one or more "predictor" variables with a "response" variable. The "inductive belief" in the predictors being non-zero is more appropriately reflected by a gradual indicator than by a 0/1 decision setting.

ed in the large number of regression coefficients and their associated *p*-values? We know that we should dispense with the extremely convenient but misleading dichotomous interpretation, but we still lack harmonized wordings that describe the inferential content of a *p*-value appropriately.

Our difficulties are due to the fact that the *p*-value is not only a highly non-linear but also noisy summary statistic of the data at hand (Hirschauer et al. 2018). This implies that a difference between, let's say, a *p*-value of 0.20 and 0.19 does not indicate the same increase of the strength of evidence against the null as a difference between 0.04 and 0.03. It also requires realizing that all sample estimates, including *p*-values, may vary considerably over random replications. The *p*-value's inconclusive inferential content precludes per se making probability statements about hypotheses (Gelman 2016; Hirschauer et al. 2016). We must therefore avoid wordings that invite confusion with the Bayesian posterior probability, i.e., the epistemic probability that a scientific proposition about the world is true *given* the evidence in the data. Unfortunately, spontaneous interpretations of the *p*-value are often not correct as people, especially when confronted with "significance" language, seem to be prone to the "inverse probability error." That is, they often confuse the "conditional probability of data given a hypothesis" (*p*-value) with the "conditional probability of a hypothesis given the data" (posterior probability). Cohen (1994: 997) coined the term "inverse probability error" to highlight that the *p*-value "does not tell us what we want to know [i.e., the posterior probability], and [that] we so much want to know what we want to know that out of desperation, we nevertheless believe that it does."

On the one hand, we know that for a two-sided test "any *p*-value less than 1 implies that the test [null] hypothesis is not the hypothesis most compatible with the data, because any other hypothesis with a larger *p*-value would be even more compatible with the data" (Greenland et al. 2016: 341). Along the same lines but with a focus on experimental data, Goodman (2008: 136) notes that "the effect best supported by the data from a given experiment is always the observed effect, regardless of its significance." On the other hand, we commonly interpret the *p*-value as a "first defense line" against being fooled by the randomness of sampling (Benjamini 2016) when generalizing from our findings to the population. We should meet this defense-line interpretation with caution, however, because the *p*-value itself is but a statistic of a noisy random sample. In plausible constellations of noise and sample size, the *p*-value exhibits wide sample-to-sample variability (Halsey et al. 2015). This is paralleled by the variability of estimated coefficients over replications. We may easily find a large coefficient in one random sample (overestimation) and a small one in another (underestimation). Unbiased estimators estimate correctly on average (Hirschauer et al. 2018). We would thence need *all* estimates from frequent replications – *irrespective* of their *p*-value and their being large or small – to obtain a good idea of the population effect size. Based on a single sample, we have no way of identifying the *p*-value below (above) which the associated effect size estimate is too large (too small), but we are very likely to overestimate effect sizes when taking "significant" results at face value (Hirschauer et al. 2018). Even when finding a highly "significant" result (with, let's say, a *p*-value of 0.001), which ironically would be a highly appreciated case in conventional NHST, we cannot make a direct inference and assume the estimated effect to accurately reflect the population effect size (Bancroft 1944; Danilov and Magnus 2004). Quite on the contrary. "Under reasonable sample sizes and reasonable population

effect sizes, it is the abnormally large sample effect sizes that result in *p*-values that meet the 0.05 level, or the 0.005 level, or any other alpha level, as is obvious from the standpoint of statistical regression" (Trafimow et al. 2018). Hence, even seemingly neutral representations such as "the retail prices of product A exceed the retail prices of product B by 20% on average ($p < 0.001$)" may be misleading because they insinuate that the evidence against the null can be translated into evidence in favor of the concrete effect that we happened to find in a single sample (Amrhein et al. 2017).

The problem of correctly interpreting *p*-values is exacerbated by the fact that the disregard of multiple comparisons, which are pervasive in econometric analyses, inflates the strength of evidence against the null and makes the *p*-value essentially uninterpretable. A *p*-value is a summary statistic that tells us how incompatible the data are with the specified statistical model including the null hypothesis. In a pre-specified *single* regression, the *p*-value represents the conditional probability of finding the observed effect (or even a larger one) in random replications *if* the null hypothesis were true. While there is no multiple testing problem if one focuses a priori on one hypothesis and model, a multiple testing problem arises whenever researchers independently perform and interpret more than one test on one data set. Disregarding the multiple testing problem is quite common in multiple regression analysis. For one thing, this is due to the fact that it is widespread practice to subject several hypotheses, which are implicitly presumed to be independent, to significance testing one by one, and then to search the list of *p*-values for "significant" results. Imagine for illustration sake a set of ten hypotheses and assume the ten corresponding regressors to be independent and completely non-predictive. Despite the completely random probabilistic structure, there is a 40% chance ($1-0.95^{10}$) of finding at least one coefficient with $p \leq 0.05$ if we perform ten independent tests (Altman and Krzywinski 2017). Furthermore, while obfuscating the dividing line between confirmatory and exploratory research, we must not forget that econometricians commonly fit regression models to pre-existing data and retain one model as the final ("best") model after multiple models have been tried out and evaluated by using some measure of model fit (e.g., likelihood ratio or Akaike Information Criterion). We arrive at overconfident conclusions if we assess the strength of evidence in only the "best" model even though multiple analytical alternatives had been tried out before (Danilov and Magnus 2004; Forstmeier et al. 2016). The same applies, to an even greater extent, when researchers self-interestedly explore multiple analytical variants and selectively choose one variant that "works" in terms of producing statistical significance (*p*-hacking; Simmons et al. 2011).

2 Suggestions for the use and interpretation of *p*-values

While some disciplines and especially those that traditionally use experimental designs, such as the medical sciences, have adopted substantial reforms to abate inferential errors related to the *p*-value, economists at large do not seem to play a very active part in the debate and the reform efforts. We believe that this is not so much due to economists not recognizing the problems associated with significance testing, but rather due to their not knowing what to do instead as long as no disciplinary consensus has been reached. With a view to the *p*-value's deep entrenchment in the current research practice and the apparent need for both guidance and (some degree of) consensus, the objective of this

paper is to contribute to the debate by systematically compiling suggestions – none of them new and none of them our own – that jointly seem to represent the most promising set of concrete and immediately actionable steps to reduce inferential errors.[5] Being economists, we focus on suggestions that are relevant for correctly interpreting the results of multiple regression analysis, which is the working horse in econometric research. Being pragmatic, we focus on suggestions that are concerned with the analysis of single-sample data, even though we are aware of the advantages of multiple-study designs, meta-analysis, and Bayesian approaches for making valid inferences.

Contenting ourselves for the time being with compiling small incremental steps must not be understood as opposition towards more substantial change in the future. Quite on the contrary. We hope that our suggestions will help prepare the field for better study designs and inferential tools, and especially more pre-registration, replication, and meta-analytical thinking in the long run that take systematic account of the fact that all estimates (standard errors, *p*-values, effect sizes) vary over random replications (Gelman 2016) because our best unbiased estimator estimate correctly *on average* (Hirschauer et al. 2018). More immediately, however, we hope that they will serve as a discussion base or even tool kit that is directly helpful, for example, to editors of economics journals who aim at revising their editorial policies and guidelines to increase the quality of published research. In brief, we address the question of how a typical econometric study, which for the time being refrains from *Bayesian statistics* and continues to use *frequentist statistics*, should proceed to avoid the inferential errors that are so pervasive at present. It is important to note that some suggestions, such as displaying standard errors, could be criticized as asking for redundant information. Readers of a research paper could in principle compute standard errors when effect sizes and *p*-values are provided. Mathematical redundancy is not a good argument, however. Instead, the question is how we should present information to avoid cognitive biases and foster the logical consistency of inferential arguments through good intuition.

Our suggestions are best preceded by a quote by Vogt et al. (2014: 242; 244) who note that the classical tools for statistical inference (including *p*-values) are inherently based on a random process of data generation: "in research not employing random assignment or random sampling, the classical approach to inferential statistics is inappropriate. […] In the case of a random sample, the *p*-value addresses the following question: 'If the null hypothesis were true of the population, how likely would we have been to obtain a sample statistic this large or larger in a sample of this size?' […] In the case of random assignment, the *p*-value targets the following question: 'If the null hypothesis were true about the difference between treated and untreated groups, how likely is it that we would have obtained a difference between them this big (or bigger) when studying treatment and comparison groups of this size?' […] If the experimental and control groups have not been assigned using probability techniques, or if the cases have not been sampled from a population using probability methods, inferential statistics are not applicable. They are routinely applied in inapplicable situations, but an error is no less erroneous for being widespread."

[5] While drawing on the critical literature on NHST in general, this compilation was especially influenced by the work of Ziliak and McCloskey (e.g., 2008). It also owes much to the recent discussions on how to remove *p*-values from their "dichotomous pedestal" and correctly interpret them as graded evidence against the null (e.g., Amrhein et al. 2017, Gelman and Carlin 2017, McShane et al. 2017, Trafimow et al. 2018).

(a) Fundamental prerequisites for using the p-value

Suggestion 1: Do not use neither *p*-values nor other inferential tools such as standard errors or confidence intervals if you already have data for the whole population of interest. In this case, no generalization (inference) from the sample to the population is necessary and you can directly describe the population properties. Do not use *p*-values either if you have a non-random sample that you chose for convenience reasons instead of using probability methods. Being inherently based on probability theory and a random process of data generation, *p*-values are not interpretable for non-random samples.[6]

Suggestion 2: Be clear that the function of the *p*-value is different depending on whether the data generating process is random sampling or random assignment. In the random sampling case, you are concerned with generalizing from the sample to the population (external validity). In the random assignment case, you are concerned with the internal validity of an experiment in which you randomly assign treatments to subjects. In an experiment, the *p*-value is a continuous measure of the strength of evidence against the null hypothesis of there being no treatment effect. It is *no help whatsoever* for generalizing to the population from which the entirety of experimental subjects have been recruited. They may, or may not, be a random sample of the population.

Suggestion 3: When using *p*-values as a tool that is to help generalize from a sample to a population, provide convincing arguments that your sample represents at least *approximately* a random sample. To avoid misunderstandings, transparently state how and from which population the random sample was drawn and, consequently, to which population you want to generalize.

(b) Wording guidelines for avoiding misunderstandings

Suggestion 4: Use wordings that ensure that the *p*-value is understood as a *graded* measure of the strength of evidence against the null. Make sure that readers realize that no particular information is associated with a *p*-value being below or above some threshold such as 0.05 (see also suggestion 19).[7]

Suggestion 5: Avoid wordings that insinuate that the *p*-value denotes an epistemic (posterior) probability that you can attach to a scientific hypothesis (the null) *given* the evidence you found in your data. Stating that you found an effect with an "error probability" of *p* is misleading, for example. It

[6] Denton (1988: 166f.) points out that "where there is a sample there must be a population" and that conceiving of the population can be difficult. The easiest case is a sample drawn from a finite population such as a country's citizens. A less intuitive sample-population-relationship arises when we generate a sample by conducting an "experiment" such as flipping a coin *n*-times. Here, the population is an imaginary set of infinitely repeated coin flips. When we are not able to repeatedly generate a random sample because we already have the data for the whole population of interest (e.g., the macro-data of a country), maintaining the *p*-value's probabilistic foundation poses serious conceptual challenges. Here, the frequentist statistician would have to imagine an infinite "unseen parent population" (or super population) and a "noisy generating process" from which we observed one realization ("great urn of nature"). Using an example, Denton critically notes in this context: "The idea of a probability process underlying the balance of trade in the fourth quarter of last year […] does not evoke wild enthusiasm from everybody. […] However, some notion of an underlying process – as distinct from merely a record of empirical observations – has to be accepted for the testing of hypotheses in econometrics to make any sense." We would add that researchers should at least explicitly state if their use of *p*-values is based on the notion that the data are a random realization from an unseen parent population.

[7] This is different from statistical decision theory where, based on restrictive assumptions, a dichotomous "$p < 0.05$-decision" would be not only conventional but optimal. We deliberately focus on an alternative perspective here, which is driven from the problems experienced in the practice of multiple regression analysis.

suggests the false interpretation that the *p*-value is the probability of the null – and therefore the probability of being "in error" when rejecting it. Consequently, avoid the term "error probability."

Suggestion 6: Avoid wordings that insinuate that a low *p*-value indicates a large or even practically or economically relevant size of the estimate, and vice versa. Use wordings such as "large" or "relevant" but refrain from using "significant" when discussing the effect size – at least as long as threshold thinking and dichotomous interpretations of *p*-values associated with the term "statistical significance" linger on in the scientific community (see also suggestion 19).

Suggestion 7: Do not suggest that high *p*-values can be interpreted as an indication of no effect ("evidence of absence") even though in the NHST-approach "non-significance" leads to non-rejection of the null hypothesis of no effect. Do not even suggest that high *p*-values can be interpreted as "absence of evidence." Doing so would negate the evident effects that you observed in the data.

Suggestion 8: Avoid formulations and representations that could suggest that *p*-values below 0.05 can be interpreted as evidence in favor of the just-estimated coefficient. Formulations claiming that you found a "statistically significant effect of z" should be avoided, for example, because they mix up estimating and testing procedures. The strength of evidence against the null cannot be translated into evidence in favor of the concrete estimate that one happened to find in a sample.

Suggestion 9: Avoid using the terms "hypothesis *testing*" and "*confirmatory* analysis" or at least put them into proper perspective and communicate that it is logically impossible to infer from the *p*-value whether the null hypothesis or an alternative hypothesis is true. We cannot even derive probabilities for hypotheses based on what has delusively become known as "hypothesis *testing*." In the usual sense of the word, a *p*-value cannot "test" or "confirm" a hypothesis, but only describe data frequencies under a certain statistical model including the null hypothesis.[8]

Suggestion 10: Restrict the use of the word "evidence" to the concrete findings in your data and clearly distinguish this evidence from your inferential conclusions, i.e., the generalizations you make based on your study and all other available evidence (see also suggestion 14).

(c) Things to do and discuss explicitly

Suggestion 11: Do explicitly state whether your study is *exploratory* and thus aimed at generating new research questions/hypotheses ("ex post hypotheses"), which might be termed "hypothesizing after results are known," or whether you aim at producing new evidence with regard to *pre-specified* research questions/hypotheses ("ex ante hypotheses"). If your paper contains both types of study, explicitly communicate *where* you change from the study of pre-specified issues to exploratory search.

Suggestion 12: In *exploratory* search for potentially interesting associations, do never use the term "*hypotheses* testing" because you have no testable ex ante hypotheses. But large effect sizes in con-

[8] In specification search, researchers regularly resort to "hypothesis testing" routines and conventional *p*-value thresholds. These routines are *not* concerned with inductive inference in terms of judging one assumption being "better" than another. Instead, they are about making *decisions* between competing models. A classical question is whether the null hypothesis, which is usually used to represent the normal distribution assumption upon which the more convenient model is based, should be rejected in favor of a more complex model. These decision-rules are based on arbitrary weights that are assigned to type I and – implicitly – type II errors (cf. footnote 4).

junction with low *p*-values may be useful as a flagging device to identify ex post hypotheses that might be worth investigating with new data in the future. To prevent overhasty generalizations from such an unconstrained "search for discoveries" in a sample, it might be worthwhile considering Berry's (2017: 897) recommendation to use the following warning: "Our study is exploratory and we make no claims for generalizability. Statistical calculations such as *p*-values and confidence intervals are descriptive only and have no inferential content."

Suggestion 13: If your study is (what would be traditionally called) "confirmatory" (see suggestion 9), i.e., aimed at producing evidence regarding *pre-specified* research questions/hypotheses, exactly report in your paper the list of questions/hypotheses that you drafted as well as the model you specified regarding structure and variable set *before* seeing the data. In the results section, clearly relate findings to these ex ante questions or hypotheses. While the study of ex ante specified hypotheses is conventionally termed "confirmatory analysis" and "hypotheses testing," these terms should be avoided or at least put into proper perspective. They might mislead people to expect categorical yes/no answers that we cannot give (see also suggestion 9).

Suggestion 14: When studying pre-specified questions or hypotheses, clearly distinguish two parts in the analysis: (i) the description of the empirical *evidence* (estimated effect sizes) that you happened to find in your single study (What is the evidence in the data?); (ii) the *inferential reasoning* that you base on this evidence under consideration of the study design, *p*-values, confidence intervals, and external evidence (What should one reasonably believe after seeing the data?). If applicable, a third part should outline the recommendations or *decisions* that you would make all things considered including the weights attributed to type I and type II errors (What should one do after seeing the data?).

Suggestion 15: If you fit your model to the data even though you are concerned with pre-specified hypotheses, explicitly demonstrate that your data-contingent model specification does *not* constitute "hypothesizing after the results are known." When using *p*-values as an inferential tool that is to help make inferences, *explicitly* consider and comment on multiple comparisons. Doing so, distinguish between (i) multiple comparisons that you make when you perform more than one test on one data set in your final multiple regression model, and (ii) the multiple comparisons that you make if you tried multiple models before retaining one model as the "best" model. If appropriate, use robustness checks to show how substantially stable your findings are over a reasonable range of analytical variants.

Suggestion 16: Explicitly distinguish between statistical and scientific inference. In the random sampling case, for example, *statistical inference* is concerned with the fact that even a random sample does not exactly reflect the properties of the population (sampling error). Generalizing from a random sample to its population is only the first step of *scientific inference*, which is the totality of reasoned judgments (inductive generalizations) that we make in the light of our own study and the available body of external evidence. We might want to know, for example, what we can learn from a random sample of a country's agricultural students for its student population, or even people in general. Be clear that a *p*-value can do *nothing* to assess the generalizability of results beyond the parent population (here: the country's *agricultural* students) from which the random sample has been drawn.

(d) Operative rules

Suggestion 17: Provide information regarding the size of your estimate (point estimate). In many regression models, a meaningful representation of magnitudes will require going beyond coefficient estimates and displaying marginal effects or other measures of effect size.

Suggestion 18: Do not use asterisks (or the like) to denote different levels of "statistical significance." Doing so could instigate erroneous categorical reasoning.

Suggestion 19: Provide *p*-values if you use the graded strength of evidence against the null as an argument (amongst others) to make inferences. However, do not classify results as being "statistically significant" or not. That said, avoid using the terms "statistically significant" and "statistically non-significant" altogether. Dispensing with these two categorical labels would enable you for the first time to use "relevant" and "significant" as interchangeable terms. However, to avoid confusion, it might be better to steer clear of the term "significant" altogether.

Suggestion 20: Provide standard errors for all effect size estimates. Additionally, provide confidence intervals for the focal variables of interest associated with your pre-specified research questions/hypotheses.

We believe that our suggestions represent practically significant steps towards improvement, but we do not expect that all empirical economists will endorse all of them at once. Some suggestions, such as providing effect size measures and displaying standard errors, are likely to cause little controversy. Others, such as renouncing dichotomous significance declarations and giving up the term "statistical significance" altogether, will possibly be questioned. Opposition against giving up the conventional and "neat" yes/no declarations is likely to be fueled by the fact that no consensus has yet been reached as to which formulations are appropriate and foolproof to avoid cognitive biases and communicate the correct meaning of frequentist concepts such as *p*-values and confidence intervals.

First of all, we need intuitive and correct formulations that ensure that the *p*-value is understood as a *graded* measure of the strength of evidence *against the null*. Which wording is appropriate to convey the information of a *p*-value of, let's say, 0.37 as opposed to 0.12 or 0.06 or 0.005 – for large and small effects, respectively? Our troubles do not come as a surprise since the difficulties of translating the *p*-value concept adequately into natural language are at the heart of the problem. Berry (2017: 896) puts it in a nutshell: "I forgive nonstatisticians who cannot provide a correct interpretation of $p < 0.05$. *p*-Values are fundamentally un-understandable. I cannot forgive statisticians who give understandable—and therefore wrong—definitions of *p*-values to their nonstatistician colleagues. But I have some sympathy for their tack. If they provide a correct definition, then they will end up having to disagree with an unending sequence of 'in other words'. And the colleague will come away confused […]." While this statement may seem overly pessimistic, we agree with the problem description. The only way out is to *find* and *agree* on formulations that convey the limited but existing informational content of the *p*-value in both a *correct* and *meaningful* way, lest we better abandon its use altogether.

We also need formulations that provide an intuitive and correct interpretations of confidence intervals (CI). Imagine observing a mean difference of 10 g in daily weight gains between two groups of ani-

mals that were randomly assigned to different dietary treatments. Finding a 95% CI of [8, 12], it would not be correct to say that the difference is between 8 g and 12 g with 95% probability. Not much change for the better is obtained by replacing "probable" by "plausible" or "confident." Stating, for example, that we can be 95% confident that the difference is between 8 g and 12 g" is extremely deceptive. While such statements sound like uncertainty statements, they promise too much certainty. In the words of Gelman (2016), they are to be qualified as "uncertainty laundering" because they neglect the inherent uncertainty of the CI itself. A correct interpretation requires realizing that CI (analogous to *p*-values) are noisy and vary from one random sample to the other. A 95% CI only means that 95% of CI computed for repeatedly drawn random samples will capture the "true" value (Greenland et al. 2016). Most formulations in empirical papers and even textbooks, however, seem to communicate in one way or the other that a CI provides the probability that the "true" effect size (population parameter) is within the stated interval. They thus insinuate that we could make epistemic probability statements regarding population effect sizes based on the results of a single study. Such statements must be reserved to Bayesian analysis (here: the Bayesian posterior probability interval), however.

3 Reforms under way and outlook

Both the accumulation of knowledge and technological developments in computing continuously shift what constitutes best methodological practice in statistical analysis. Given these dynamics, sticking traditions as well as tight but rarely scrutinized journal guidelines may slow down or even prevent necessary change. While overly rigid formal rules in the process of publication are detrimental with respect to dynamic adjustments, guidelines can also be a pertinent means to disseminate new best practice procedures and induce overdue change in inert disciplinary traditions.

Trying to get an impression of the reforms in publishing economic research, we asked the editors of 100 leading economics journals (Scimago Journal & Country Rank) about policy changes with respect to the use of *p*-values.[9] Overall, journals seem still a long way from translating a significant portion of recent reform suggestions into concrete journal policies. Despite the prominence of the current *p*-value debate, a substantial share of editors believe that their reviewing systems are sufficiently effective to prevent inferential errors. Consequently, they do not see a need to bring about formal change. Some journal editors, however, seem to be seriously worried about the misuses and misinterpretations of the *p*-value. What is more, some editorial boards are deliberating concrete steps or have already started reforming their guidelines to prevent misleading practices and inferential errors. It was interesting to

[9] We asked the editors informally via e-mail whether they had specific journal policies with respect to the use of *p*-values in journal papers. The editors of 33 journals responded, 6 of them stated that they could only provide little information (e.g., because editorial boards were being reshuffled), 19 claimed that there was no need for reforms, and 8 said that they had already revised their journal guidelines or deliberated on doing so. Perhaps more noteworthy than the exact numbers of this "flashlight survey" is an information provided by Penny Goldberg, the former editor of the *American Economic Review*, regarding the practices of six of the most prestigious economics journals (communication from October 2017): "Indeed many Economics journals (the AER, the four AEJs, and Econometrica) have adopted a new policy according to which authors should report coefficients and standard errors, or coefficients and confidence intervals, but should not use asterisks or other symbols or fonts to denote statistical significance. […] The concern that led to this policy was that readers (incl. referees) often get fixated on asterisks and dismiss results that are "barely significant", forgetting that the threshold one uses for asterisks is based on a convention. Reporting standard errors seems more objective."

learn that these reforms represent a subset of our suggestions. For example, leading journals, such as the *American Economic Review*, *Econometrica*, and the four AEJs (*Applied Economics*, *Economic Policy*, *Macroeconomics*, *Microeconomics*), now request authors not to use asterisks or other symbols to denote "statistical significance." While this seems a small change, it breaks with a convention that many economists considered to be set in stone. The basic idea behind banning asterisks is to prevent overconfident yes/no conclusions. The *American Economic Review* and *Econometrica* furthermore request authors to explicitly report effect sizes and display standard errors (or even confidence intervals) *instead of* *p*-values in results tables, but they do not explicitly ban *p*-values. This is consistent with a suggestion put forward by many critical voices in the recent debate – to demote *p*-values from their pedestal and consider them as a tool amongst many that help make appropriate inferences (cf., e.g., Amrhein et al. 2017; McShane et al. 2017; Trafimow et al. 2018). The fact that some leading economics journals, which are widely considered as beacons for best practice, have initiated modest but sensible changes with regard to the use of the *p*-value is a promising signal. Goodman (2017: 559) notes that "norms are established within communities partly through methodological mimicry." If a field's flagship journals, opinion leaders, and professional associations take up the lead, they may be able to set a trend. "Once the process starts, it could be self-enforcing. Scientists will follow practices they see in publications; peer reviewers will demand what other reviewers demand of them."

Besides avoiding mistakes within a single study, making appropriate inferences requires considering the body of evidence including all prior knowledge in the field under research. Several reforms beyond the realms of the single study have been suggested. Meta-analyses that systematically consolidate the body of evidence and Bayesian methods that formally consider prior knowledge are prominent examples. Furthermore, two reforms on the institutional level are practically important in some disciplines but only nascent in others. First, many leading journals now oblige authors to provide their raw data and analytical protocols in the appendix to facilitate *replication* studies scrutinizing a study's findings. While compulsory sharing of raw data and analytical protocols seems to slowly trickle down to more and more journals, institutionalized efforts to actually promote replication studies are weak in economics compared to other fields. According to Duvendack et al. (2015, 2017), most of the 333 economic Web-of-Science journals still give low priority to replication. The same holds for initiatives targeted at counteracting publication bias. While a global initiative *All Trials Registered/All Results Reported* was launched in 2013 in the medical sciences, similar efforts are rare in economics. Among the few exceptions are *The Replication Network,* and *Replication in* Economics. Both platforms are aimed at fostering the scrutiny of scientific claims and at counteracting publication bias by providing not only databases for replications but also equal opportunities for publishing positive and negative results. Two recent activities in economics indicate that problem awareness in economics is growing: the *American Economic Review* published eight short papers about *Replication* in its 5th issue of 2017; and Economics published a special issue *The practice of replication* in 2018.

Another important reform on the institutional level is *pre-registration*. Pre-registration goes beyond the mere appeal to honestly report *pre-specified* hypotheses and analysis plans. Instead, it obliges researchers to disclose their hypotheses, data, and analytical approach *before* running the analysis. Anal-

yses that deviate from the pre-analysis plan must be justified in the final paper. Pre-registration is aimed at preventing covert multiple comparisons (*p*-hacking) and at providing equal chances of being published independent of which results are eventually found. In other words, it is to prevent not only selective reporting but also selective publishing and thus the bias towards "statistically significant" findings (cf., Rosenthal 1979), which seems to be widespread even in economic flagship journals (Brodeur et al. 2016). Contrary to fields such as clinical drug trials for which pre-registration is standard (*http://www.who.int/ictrp/network/primary/en/*), it is as yet rare in the social sciences. However, two innovative initiatives have been launched recently. The *American Economic Association* started pre-registering randomized controlled trials on its AEA RCT platform in 2013. Since 2017 study designs and analysis plans have to meet some formal criteria to be published before the data is collected. An even further reaching pilot program was started by the *Journal of Development Economics* in 2018 to test how pre-results peer reviews ("blind reviews") can contribute to better science in economics. The pilot gives researchers who intend to use data that are yet to be collected "the opportunity to have their prospective empirical projects reviewed and approved for publication *before* the results are known."

The poor development of replication, meta-analysis, and pre-registration in economics stems from the discipline's culture and its focus on observational study, data-driven model specification, and multiple regression analysis. In other words, there are questions to be answered before approaches from other fields can be transplanted to (non-experimental) economic research. It is not clear, for example, how pre-registration or replication should work within a culture in which it is not only common but highly appreciated to specify regression models *after* seeing the data ("model fitting"). If one accepts the model fitting exercise, one would have to pre-register full decision trees. Furthermore, pre-registration of studies that are based on pre-existing data that are already available to the research community before registering do not seem to make much sense. In addition, the question arises of how to carry out quantitative meta-analysis and consolidate the body of evidence when even within a narrow field of research there are often as many data-dependent model specifications as studies. The fact that economic research is mainly a bottom-up research exercise is responsible for the lacking comparability across studies. Non-programmed bottom-up research produces a large quantity of empirical results on topical issues, but is plagued by an enormous heterogeneity of empirical measures and model specifications. Besides differing measures for the focal variables of interest, databased models are regularly populated by differing interaction terms, transformed variables, lagged variables, higher-order polynomials, and control variables. Given the heterogeneity of econometric models, applied economists need consensus regarding the legitimacy and meaning of specification search as well as regarding best practices for replication and meta-analysis.

Acknowledgment

We owe a special debt to Andrew Gelman (Columbia University) who gave us helpful comments and criticism on our suggestions. Any remaining errors are our own. We would like to thank the German Research Foundation for financial support.

References

Altman, N., Krzywinski, M. (2017): Points of significance: P values and the search for significance. Nature Methods 14(1): 3-4.

Amrhein, V., Korner-Nievergelt, F., Roth, T. (2017): The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research (https://peerj.com/preprints/2921.pdf).

Bancroft, T.A. (1944): On biases in estimation due to the use of preliminary tests of significance. Annals of Mathematical Statistics 15(2): 190-204.

Benjamini, Y. (2016): It's not the *p*-values' fault. The American Statistician 70(2): Supplemental Material to the ASA Statement on P-Values and Statistical Significance (http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5354.pdf).

Berry, D. (2017): A p-Value to Die For. Journal of the American Statistical Association 112(519): 895-897.

Brodeur, A., Lé, M., Sangnier, M., Zylberberg, Y. (2016): Star Wars: The Empirics Strike Back. American Economic Journal: Applied Economics 8(1): 1-32.

Cohen, J. (1994): The earth is round (p < 0.05). American Psychologist 49(12): 997-1003.

Colquhoun, D. (2014): An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science 1: 140216; http://dx.doi.org/10.1098/rsos.140216: 1-16.

Danilov, D., Magnus, J.R. (2004): On the harm that ignoring pretesting can cause. Journal of Econometrics 122(1): 27-46.

Denton, F.T. (1988): The significance of significance: Rhetorical aspects of statistical hypothesis testing in economics. In: Klamer, A., McCloskey, D.N., Solow, R.M. (eds.): The consequences of economic rhetoric. Cambridge: Cambridge University Press: 163-193.

Duvendack, M., Palmer-Jones, R., Reed, W.R. (2015): Replications in Economics: A Progress Report. Econ Journal Watch 12(2): 164-191.

Duvendack, M., Palmer-Jones, R., Reed, W.R. (2017). What Is Meant by "Replication" and Why Does It Encounter Resistance in Economics? American Economic Review, 107(5): 46-51.

Fisher, R.A. (1925): Statistical methods for research workers. Edinburgh: Oliver & Boyd.

Forstmeier, W., Wagenmakers, E.-J., Parker, T.H. (2016): Detecting and avoiding likely false-positive findings – A practical guide. Biological Reviews of the Cambridge Philosophical Society 92(4): 1941-1968.

Gelman, A., Carlin, J. (2017): Some natural solutions to the *p*-value communication problem–and why they won't work. Journal of the American Statistical Association 112(519): 899-901.

Gelman, A., Stern, H. (2006): The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician 60(4): 328-331.

Gelman, A., Loken, E. (2014): The Statistical Crisis in Science. American Scientist 102: 460-465.

Gelman. A. (2016): The problems with p-values are not just with p-values. American Statistician, supplemental material to the ASA statement on p-values and statistical significance, 10, 2016.

Gigerenzer, G., Marewski J.N. (2015): Surrogate Science: The Idol of a Universal Method for Statistical Inference. Journal of Management 41(2): 421-440.

Goodman, S. (2008): A dirty dozen: Twelve p-value Misconceptions. Seminars in Hematology 45: 135-140.

Goodman, S.N. (2017): Change norms from within. Nature 551: 559.

Greenland, S. (2017): Invited Commentary: the Need for Cognitive Science in Methodology. American Journal of Epidemiology 186(6): 639-645.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G. (2016): Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31(4): 337-350.

Halsey, L.G., Curran-Everett, D., Vowler, S.L., Drummond, B. (2015): The fickle P value generates irreproducible results. Nature Methods 12(3): 179-185.

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. (2018): Statistics Surveys 12: 136-172.

Hirschauer, N., Mußhoff, O., Grüner, S., Frey, U., Theesfeld, I., Wagner, P. (2016): Die Interpretation des p-Wertes – Grundsätzliche Missverständnisse. Journal of Economics and Statistics 236(5): 557-575.

Ioannidis, J., Doucouliagos, C. (2013): What's to know about the credibility of empirical economics? Journal of Economic Surveys 27(5): 997-1004.

Ioannidis, J.P.A. (2005): Why Most Published Research Findings are False. PLoS Medicine 2(8): e124: 0696-0701.

Kline, R.B. (2013): Beyond Significance Testing: Statistics Reform in the Behavioral Sciences (2nd ed.). American Psychological Association, Washington.

Krämer, W. (2011): The Cult of Statistical Significance – What Economists Should and Should Not Do to Make their Data Talk. Schmollers Jahrbuch 131(3): 455-468.

Lehmann, E.L., Romano, J.P. (2010): Testing statistical hypotheses. 3rd ed. New York: Springer.

McCloskey, D.N., Ziliak, S.T. (1996): The Standard Error of Regressions. Journal of Economic Literature 34(1): 97-114.

McShane, B., Gal, D. (2007): Statistical Significance and the Dichotomization of Evidence. Journal of the American Statistical Association 112(519): 885-908.

McShane, B., Gal, D., Gelman, A., Robert, C., Tackett, J.L. (2017): Abandon Statistical Significance (https://arxiv.org/pdf/1709.07588.pdf).

Motulsky, J.J. (2014): Common Misconceptions about Data Analysis and Statistics. The Journal of Pharmacology and Experimental Theurapeutics 351(8): 200-205.

Neyman, J., Pearson, E.S. (1933): On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London A 231: 289-337.

Rosenthal, R. (1979): The file drawer problem and tolerance for null results. Psychological Bulletin 86(3): 638-641.

Sellke, T., Bayarri, M.J., Berger, J.O. (2001): Calibration of *p*-Values for Testing Precise Null Hypotheses. The American Statistician 55(1): 61-71.

Simmons, J.P., Nelson, L.D., Simonsohn U. (2011): False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22(11): 1359-1366.

Trafimow, D. et al. (2018): Manipulating the alpha level cannot cure significance testing. Frontiers in Psychology 9 (https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00699/full).

Vogt, W.P., Vogt, E.R., Gardner, D.C., Haeffele, L.M. (2014): Selecting the right analyses for your data: quantitative, qualitative, and mixed methods. New York: The Guilford Publishing.

Wasserstein, R.L., Lazar N.A. (2016): The ASA's statement on p-values: context, process, and purpose, The American Statistician 70(2): 129-133.

Ziliak, S.T. (2016): Statistical significance and scientific misconduct: improving the style of the published research paper. Review of Social Economy 74(1): 83-97.

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