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

Contributing authors

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, 06120 Halle (Saale), Germany

norbert.hirschauer@landw.uni-halle.de

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, 06120 Halle (Saale), Germany

sven.gruener@googlemail.com

Oliver Mußhoff

Georg August University Goettingen

Department for Agricultural Economics and Rural Development, Farm Management

Platz der Göttinger Sieben 5, 37073 Göttingen, Germany

Oliver.Musshoff@agr.uni-goettingen.de

Claudia Becker

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

Institute of Business Studies, Chair of Statistics

Große Steinstraße 73, 06099 Halle (Saale), Germany

claudia.becker@wiwi.uni-halle.de

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

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 multiple regression analysis. 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: statistical inference, *p*-value

1 Introduction

One might think, enough has been said about the misuses and misinterpretations of the *p*-value, both before and after the ASA-statement (WASSERSTEIN and LAZAR 2016). But the problems seem to be here to stay. Two related 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 have propagated cognitive biases that seduce even experienced researchers to make logically inconsistent and overconfident inferences, both when *p* is below and when it is above the conventional "significance" threshold. The following errors seem to be particularly widespread:[1]

1) use of *p*-values when there is neither a random sample nor a treatment after random assignment
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 testing and estimating and misinterpretation of "significant" results as evidence in favor of the coefficients/effect sizes estimated from a single sample
   - treatment of effects that are "statistically non-significant" as being zero (confirmation of the null)
4) inflation of evidence against the null caused by *p*-hacking or unconsidered multiple comparisons
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), GOODMAN (2008), 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), WASSERSTEIN and LAZAR (2016), ZILIAK (2016). While this list may seem impressive, it should be noted that it contains but a small selection of important references on *p*-value misconceptions from the last decades. Since we focus on the compilation of pragmatic and immediately actionable 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 *p*-value misconceptions (the don'ts) in the past.

The ASA-statement has highlighted that the *p*-value does not provide a good measure of evidence regarding a hypothesis. It is nonetheless a continuous measure of the strength of evidence against the null hypothesis, 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. 2017). Joining AMRHEIN et al. (2017), BERRY (2017), GELMAN and CARLIN (2017), GREENLAND (2017), MCSHANE et al. (2017), TRAFIMOW et al. (2017), 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. Since the *p*-value is so deeply anchored in the minds of most applied researchers, we believe that demanding drastic procedural changes, such as renouncing *p*-values or completely replacing frequentist approaches by Bayesian statistics, is not the most promising approach for mitigating the serious inferential errors that we see today. Dispensing with the dichotomy of significance testing but retaining the *p*-value and adopting small and manageable but efficient steps towards improvement seems to be more promising (AMRHEIN et al. 2017). Adequate steps will have to take account of the idiosyncrasies of each scientific discipline. For example, requirements in the biomedical sciences, which often focus on risk ratios or mean differences between experimental treatments, will at least partly differ from those in the social sciences including economics, which frequently resort to multiple regression analysis of observational data. In this paper, we compile a promising list of such steps with a focus on the use of *p*-values in multiple regression analysis.

2 Things to consider in general

Even if one is fully 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.[2] It is an even more difficult task to provide readers with an interpretative evaluation of the often large numbers of coefficient estimates in multiple regressions that avoids inferential errors. 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 represented in the large number of regression coefficient estimates and their associated *p*-values? We know that we should dispense with the extremely convenient

[2] Misunderstandings are exacerbated by the fact that few users of NHST seem to be familiar with the details of the differing perspectives adopted by FISHER (1925) and NEYMAN and PEARSON (1933) that have been amalgamated into what is today known as NHST (see, e.g., 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 argued that "significance testing" is 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). It seems that the assumptions of statistical decision theory rarely apply to empirical research based on multiple regression analysis whose regular objective is to investigate (causal) relationships (coefficients, effect sizes) 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.

but misleading dichotomous interpretation. This creates a big problem, however. Despite hundreds or even thousands of papers criticizing the misuses and misinterpretations of the *p*-value (the don'ts), we still lack appropriate wordings that describe the informational content of a *p*-value correctly *and* meaningfully. This is because the *p*-value does not provide a clear rationale or even calculus for statistical inference (GOODMAN 2008), or, as BERRY (2017: 896) formulates more drastically, because "as such it has no inferential content."

Our wording difficulties are due to the obscure as well as noisy and inconclusive informational content of the *p*-value that as such even precludes making probability statements about hypotheses (GELMAN 2016). 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, while realizing that there is evidence in the data, 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 noisy statistic of data obtained from one-time random sampling. 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). We must not forget that unbiased estimators estimate correctly on average (HIRSCHAUER et al. 2017). 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. 2017). 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 the conventional NHST-approach, we cannot make a direct inference and assume the estimated effect to accurately reflect the population effect size (BANCROFT 1944). 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 .05 (or the .005) criterion" (TRAFIMOW et al. 2017: 10).[3] Hence, even seemingly neutral, non-dichotomous 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).

[3] See, for example, DANILOV and MAGNUS 2004 for considerations on correcting overestimation.

The problem of interpreting *p*-values is further exacerbated by the fact that multiple comparisons, which are inherent to multiple regression, inflate the strength of evidence against the null as indicated by the *p*-value. Since the extent of multiple comparisons varies between studies, *p*-values cannot be compared across different studies. 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 *single* regression, a *p*-value (for example 0.05) represents the conditional probability of finding the observed effect (or even a larger one) in random replications *if* the null hypothesis were true. In contrast, in a *multiple* regression with, let's say, ten focal predictor variables, we make ten comparisons in that we assess the strength of evidence against the null as many times as there are variables of interest. Even if all ten null hypotheses were true, we would have a 40.1% ($1-0.95^{10}$) probability of finding at least one coefficient with $p \leq 0.05$. Finding a low *p*-value for a coefficient in a multiple regression represents much weaker evidence against the null than finding the same *p*-value in a single regression. While obfuscating the dividing line between confirmatory and exploratory research (see suggestion 9, 11, and 15), we must furthermore not forget that applied researchers commonly retain one model as the final ("best") model after different models have been tried out and evaluated by using some measure of model fit (e.g., likelihood ratio or Akaike Information Criterion). We necessarily produce inflated effects and arrive at overconfident conclusions if we assess the strength of evidence in only the "best" model even though multiple alternatives had been tried out before (DANILOV and MAGNUS 2004; FORSTMEIER et al. 2016).

Remembering that the *p*-value is but a summary statistic of the data at hand is important because we must avoid all wordings that invite confusion with the posterior (or Bayesian) probability, i.e., the epistemic probability that a hypothesis or scientific proposition about the world is true *given* the evidence from the data. There is a big tension between the correctness and the intuitive meaningfulness of the *p*-value interpretation 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). "Inverse probability error" is a term coined by COHEN (1994: 997) to emphasize 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."

3 Suggestions for the use and interpretation of *p*-values in multiple regressions

Widely scattered over time and disciplines as well as journals, a huge amount of criticism regarding the use of *p*-values (the don'ts) as well as a large number of suggestions for reform (the do's) have accumulated. However, neither abundant criticisms of misuses nor grand visions of how to replace the *p*-value through other tools of statistical inference such as Bayesian statistics have been of much avail. Not even the ASA-statement seems to have produced much change so far (MATTHEWS 2017). We believe that this is not so much due to researchers not recognizing the problems associated with conventional significance testing, but rather due to their not knowing what to do instead. With a view to the apparent need for both

guidance and (at least some degree of) consensus among scientists, this commentary discusses and suggests immediate reforms that seem to be realistic in the light of the *p*-value's deep entrenchment in current research practice.

We systematically compile suggestions (do's) – 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.[4] 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. In brief, the criteria for the selection of suggestions were as follows: (i) their suitability for furthering a *correct* and *meaningful* interpretation of *p*-values associated with regression coefficient estimates in single studies; (ii) their capacity to provide small and efficient changes that are *manageable* for all those who are so much accustomed to using *p*-values that they are probably not ready (yet) to meet the huge challenges of fully Bayesian analyses within a multiple regression framework.

Contenting ourselves for the time being with compiling small incremental steps for single-study designs 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. 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 (economic) journals who reflect on best practices and try to revise their editorial policies and guidelines in order 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 *p*-values, should proceed and which wordings it should use to avoid the many inferential errors that are so pervasive at present. It is important to note that some suggestions, such as displaying random 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 emphasize that the classical tools for statistical inference (including *p*-values) are inherently based on probability theory: "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

[4] While drawing on the critical literature on NHST in general, the selection of these suggestions 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. 2017).

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."

(a) Fundamental prerequisites for using the p-value

Suggestion 1: Do not use neither *p*-values nor other inferential tools such as random errors or confidence intervals if you have (a 100% sample of) the population of interest. In this case, no generalization from the sample to the population (statistical inference) is necessary and you can directly describe the population properties. Do not use *p*-values either if you simply provide descriptive statistics or if you have a non-random sample that you have chosen for convenience reasons instead of using probability methods. Being inherently based on probability theory and repeated random sampling, displaying *p*-values for a non-random sample is meaningless and *no help whatsoever* for making statistical inferences.

Suggestion 2: When using *p*-values as an inferential aid in studies based on random sampling or random assignment, be clear that the function of the *p*-value is different in the two cases. In the random sample case, you are concerned with generalizing from the sample to the population. In the random assignment case, you are concerned with the internal validity of an experiment in which you randomly assign experimental subjects to groups that you subject to different treatments. For random assignments, the *p*-value is a continuous measure of the strength of evidence against the null hypothesis of there being no treatment effect in the experiment. It is *no help whatsoever* to assess the generalizability of results towards the population from which the experimental subjects themselves have been recruited. They may, or may not, be a random sample of a certain population.

Suggestion 3: Be aware that random samples from a population are often costly to come by and therefore frequently not available. 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 to which population you want to generalize.[5]

[5] Emphasizing the *p*-value's probabilistic foundation, DENTON (1988: 166f.) points out that "where there is a sample there must be a population." He notes 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 studying observational macro-data, maintaining the *p*-value's probabilistic foundation poses serious conceptual challenges. One would have to imagine an "unseen parent population" and a noisy generating process from which we observe a random realization.

(b) Wording guidelines for avoiding misunderstandings

Suggestion 4: Use wordings that ensure that the *p*-value is understood as a *continuous* measure of the strength of evidence against the null. Make sure that the reader realizes that no particular information is associated with a *p*-value being either below or above some particular threshold such as 0.05 (see also suggestion 19).[6]

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 suggests the false interpretation that the *p*-value is the probability of the null – and therefore the probability of being wrong ("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 directly be translated into evidence in favor of the concrete estimate that one happened to find in a sample.

Suggestion 9: Do not use neither the term "hypothesis *testing*" nor the term "*confirmatory* analysis." 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*." *p*-values cannot "test" or "confirm" any hypothesis at all, but only describe data frequencies under a certain statistical model including the null hypothesis.[7]

---

[6] 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.

[7] In specification search, researchers try to identify a model that reasonably fits the data, i.e., *decisions* are to be made between competing models. Doing so, researchers often resort to statistical tests based on "hypothesis testing" routines resting upon conventional *p*-value thresholds. Despite this label, the *p*-value in statistical tests is not an epistemic probability of one model being "better" than another. Instead, these routines reflect conventional decision-rules of when the null hypothesis, which is usually used to represent the more convenient simple model such as one based

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, or whether it is aimed at producing new evidence with regard to *pre-specified* research questions/hypotheses. While the latter is conventionally termed "confirmatory analysis," this term should be avoided. It might mislead people to expect categorical yes/no answers that we cannot give (see suggestion 9). Your paper may also contain both types of analysis. If so, explicitly communicate *where* you change from the study of pre-specified issues to exploratory search.

Suggestion 12: In the *exploratory* search for potentially interesting associations, large effect sizes in conjunction with low *p*-values may be useful as a flagging device to identify what might be worth investigating with new data in the future. To prevent overhasty generalizations from such an unconstrained "search for discoveries" in a single 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 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 (econometric) model you specified regarding structure and variable set *before* seeing the data. In the results section, clearly relate findings to the initial questions or hypotheses.

Suggestion 14: When studying pre-specified questions or hypotheses, clearly distinguish two parts in your analysis: (i) the description of the *evidence* (estimates) that you actually happened to find in your single study (What is the evidence in this data?); (ii) the *inferential reasoning* that you base on this evidence under consideration of *p*-values, confidence intervals, the study design, and all relevant external evidence (What should one reasonably believe after seeing this 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 this data?).

Suggestion 15: If you specify/adapt your initial model contingent on the data even though you aim to carry out a study concerned with pre-specified hypotheses, explicitly justify that your data-contingent model specification does *not* constitute "hypothesizing after the results are known." When making inferences, *explicitly* consider and comment on multiple comparisons that inflate the strength of the evidence against the null as indicated by the *p*-value. Doing so, distinguish between (i) the multiple comparisons

---

on a normal distribution assumption, 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.

that you make because you study multiple variables in your final 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 ("robust") your findings are over a reasonable range of analytical variants.

Suggestion 16: In inferential reasoning, explicitly distinguish between statistical and scientific inference. *Statistical inference* and the *p*-value are concerned with the random sampling error, i.e., the fact that even a random sample will not exactly reflect the properties of the population. 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, its citizens, or even human beings in general. Be clear in your inferential reasoning that a *p*-value, being a probabilistic concept, 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 for coefficient estimates or marginal effects if you feel that graded evidence against the null is useful for making inferences despite the unknown but often wide sample-to-sample variability of *p*-values. 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 enables you for the first time to use "relevant" and "significant" as interchangeable terms without causing confusion.[8]

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

We hope that we give an impulse to applied researchers to comply with these single-study suggestions in all cases in which coordinated multiple-study designs are not feasible. We furthermore believe that the quality of published research could be improved by incorporating these suggestions as best practice rules

[8] There might be cases where the restrictive assumptions of statistical decision theory, which divide the "world" into two mutually exclusive states between which a decision has to be made, are useful. In the natural sciences, for example, one may have to decide upon the size of a certain parameter within a series of consecutive experiments where based on previous experimental runs the parameter values are successively refined for the subsequent run.

into journal guidelines. With a view to the inflation of the strength of evidence through covert multiple comparisons (*p*-hacking), SIMMONS et al. (2012) propose a formalization outside of the paper that seems worthwhile considering. Similar to the standard "no-competing-interests" statements, they suggest to oblige authors to make a *formal* "no-*p*-hacking" declaration. The problem is that it is difficult to unambiguously define the practices that are outlawed as *p*-hacking. A substantiated selection of an analytical approach is not *p*-hacking. But results will be biased if researchers covertly engage in multiple comparisons and selectively publish only the analytical variant that "worked" in that it produced lower *p*-values than other variants (HIRSCHAUER et al. 2016). For a formal declaration to make sense, journals must clearly specify outlawed practices. Some people may think that, given the perverse publish-or-perish conditions that many researchers face today, a formal no-*p*-hacking declaration is just an empty phrase. However, we believe that it could produce a practically significant reinvigoration of science ethics' call for transparency and integrity that leads to published research better reflecting reality than what we have seen in the past. In this sense, we agree with SIMMONS et al. (2012: 6) that "changes need not to be judged in terms of their perfection, but merely in terms of their improvement."

While we believe that our suggestions represent practically significant steps towards improvement, we do not expect that all researchers will endorse all of them at once. With a view to their acceptance and immediate viability, there seem to be three categories: some suggestions, such as the eschewal of asterisks and the requirement to display random 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. And two suggestions, both concerned with leaving behind categorical yes/no declarations, require more than just debate before they can be implemented. They require solving problems that linger on because the existing statistical literature does not yet meet applied researchers' needs for guidance on how to do things in a post $p < 0.05$ era.

First, we do not yet have formulations at our disposition that ensure that the *p*-value is understood as a *continuous* 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. This is what BERRY (2017): demands: "We

created a monster [the *p*-value]. And we keep feeding it, hoping that it will stop doing bad things. It is a forlorn hope. No cage can confine this monster. The only reasonable route forward is to kill it." Contrary to Berry, we believe that we should only dispense with the dichotomy of significance testing and categorical reasoning but retain the *p*-value itself because of its familiarity and potential usefulness. However, if retaining the *p*-value is to make sense, we need practical guidance for applied researchers who, rightly leaving behind dichotomous significance declarations, are in need of correct and understandable formulations of what a *p*-value means. Debate and consensus on such guidance could possibly be brought about through organized approaches and concerted actions of statistical associations and the professional associations and journals in the various disciplines.

A second and analogous problem arises because it is equally hard to provide a correct wording and good intuition regarding the meaning of the confidence interval (CI). Imagine observing a mean difference of 10 g in daily weight gains between two randomly assigned animal groups that were subjected 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" with words such as "likely/plausible" or "confident." Stating, for example, that the CI indicates the precision of the estimation and that "in other words, we can be 95% confident that the difference is between 8 g and 12 g" is extremely deceptive. This holds even though researchers could argue that they use the word "confidence" as a technical term that by convention is attached to the said interval. Even though 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). As in the case of *p*-values, we must realize that providing a correct technical definition of a difficult-to-understand concept is not enough. It is likely to provoke an unending sequence of false "in-other-words statements." In fact, we rarely see a proper and understandable interpretation of CI in empirical papers and even textbooks. Most formulations seem to communicate in one or the other way that a CI describes the probability that the specified range contains the "true" value with the stated probability (e.g., 95%). They thus insinuate that we could make an epistemic probability statement regarding the population effect size based on the results of a single study. Such a statement must be reserved to Bayesian analysis (here: the Bayesian posterior probability interval), however. The lack of appropriate wordings is especially serious since CI have been recommended as being a part of the solution for the *p*-value problem (e.g., CUMMING 2014). But to be a part of the solution, we first need guidance and consensus regarding the wordings that are able to communicate the informational content of a CI not only *correctly* but also *meaningfully*. Again, the problem that is not restricted to a particular applied science discipline, and the process of finding a solution is likely to benefit from concerted actions of expert statisticians and the professional associations in the various disciplines.

4 Reforms under way and reforms still to be undertaken in publishing economic research

Measures addressed to author(s) of a single study

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 rigid formalizations such as tight but rarely scrutinized journal guidelines may slow down or even prevent necessary change. A particular challenge arises for interdisciplinary journals which need sufficiently flexible rules to cater for the fact that they receive manuscripts from collaborative research teams and authors with different traditions in statistical analysis. While overly rigid and sticking rules are detrimental with respect to interdisciplinary research and dynamic adjustments, guidelines can also be a pertinent means to communicate new best practice procedures and induce overdue change in disciplinary traditions and researchers' inert habits.

Trying to get an impression of "what is going on" in the practice of publishing research, we asked the editors of 100 leading economic journals (cf., http://www.scimagojr.com/journalrank.php?category=2002) about policy changes with respect to the use of *p*-values. 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 significant share of journal editors believe that their reviewing system is sufficiently effective to prevent inferential errors. Consequently, they do not see a need to bring about formal change. The editors of some journals, however, seem to be seriously worried about the misuses and misinterpretations of the *p*-value. What is more, some editorial boards are deliberating concrete future steps or have already started using their guidelines to prevent misleading practices and false inferential conclusions. 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 represents a distinct breach of a convention that many researchers considered to be set in stone for a long time. The basic idea behind banning asterisks is to prevent overconfident categorical conclusions induced by arbitrary thresholds. Other noteworthy editorial policy changes in leading economic journals (e.g., American Economic Review, Econometrica) include the request to explicitly report effect sizes (marginal effects) and display standard errors (or even confidence intervals) in results tables. With respect to the *p*-value, they call upon authors to display standard errors *instead of p*-values in results tables but do not ban the use of *p*-values when interpreting results in the text. This is consistent with a suggestion put forward by many critical voices in the recent debate, namely to demote *p*-values from their pedestal and consider them as a tool amongst many that may help researchers make appropriate inferences (cf., e.g., AMRHEIN et al. 2017; MACSHANE et al. 2017; TRAFIMOW et al. 2017).

The fact that some of the leading journals 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 their general function as beacons for best practice, the guidelines of flagship journals may become more direct agents of change due to prevalent submission practices. Researchers often submit their papers to leading journals first. If declined, they regularly try alternative publication outlets and submit their papers, written according to the guidelines of the flagship journal, more or less unchanged to less prominent journals. This might cause a trickle-down effect that generates new best practice standards for the less prominent journal.

Measures beyond the single study

Avoiding mistakes within the single study is a necessary but not sufficient condition for making correct inferences. Instead, we need to embed each study in its wider context and consider the body of evidence including all external ("prior") knowledge in the field under research. Several propositions towards improvement beyond the realms of the single study have been made. Meta-analysis that systematically consolidates the body of evidence and the formal consideration of prior knowledge through Bayesian analysis are prominent examples. Furthermore, two reforms on the level of scientific institutions (journals, scientific associations) are practically important in some disciplines but only nascent or non-existing in others.

First, many leading journals now oblige authors to provide their raw data and analytical protocols in the appendix of their paper to facilitate *replication* studies aimed at 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 strengthen replication are weak in economics compared to other fields. According to DUVENDACK et al. (2015), 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, for example, 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.

A second important reform on the institutional level is *pre-registration*. Pre-registration goes not only beyond the post-study provision of raw data and analytical protocols but also beyond the mere appeal to honestly report *pre-specified* hypotheses and analysis plans in the paper. Instead, it obliges researchers to disclose their hypotheses, data, and analytical approach *before* running the analysis. Analyses that deviate from the pre-analysis plan must be justified and explained in the final paper. Pre-registration is aimed at preventing *p*-hacking and 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 clinical drug trials for

which pre-registration is standard (http://www.who.int/ictrp/network/primary/en/), it is still rare in the social sciences. There are, however, two new initiatives. Within the "$1 Million Preregistration Challenge," the Center for Open Science (https://cos.io/prereg/) provides $1,000 to 1,000 researchers who pre-register their research projects. Because existing registries were not considered a good fit for the needs of the social sciences, the American Economic Association launched an initiative in 2017 to register randomized controlled trials on its AEA RCT platform (https://www.socialscienceregistry.org/). After peer-approval of the study design and analysis plan, research projects are accepted and published before they are implemented.

The poor development of replication, pre-registration, and meta-analysis in economics stems from the discipline's culture and its preoccupation with 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 replication or preregistration should work within a disciplinary culture in which it is not only common but also highly appreciated practice to specify regression models *after* seeing the data ("model fitting"). Related to that, 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 guidance and consensus regarding best practices for specification search, replication, and meta-analysis. We hope that professional associations in the field of economics take up the cause and organize a debate on the urging question of how to systematically build up knowledge and scrutinize scientific claims derived from nearly limitless variants of databased models.

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.

References

Amrhein, V., Korner-Nievergelt, F., Roth, T. (2017): The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ, DOI 10.7717/peerj.3544.

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 (DOI: 10.1080/01621459.2017.1316279).

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

Cumming, G. (2014): The New Statistics: Why and How. Psychological Science 25(1): 7-29.

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.

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

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

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

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 (doi:10.1111/brv.12315).

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.

Gelman, A., Carlin, J. (2017): Some natural solutions to the p-value communication problem-and why they won't work. Blogsite: Statistical Modeling, Causal Inference, and Social Science.

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

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 (https://www.nature.com/magazine-assets/d41586-017-07522-z/d41586-017-07522-z.pdf).

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. (2017): Pitfalls of significance testing and p-value variability – implications for statistical inference. The Centre for Statistics working paper series 07/2017, Georg August University Goettingen (https://www.uni-goettingen.de/de//511092.html).

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

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

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.

Matthews, R. (2017): The ASA's *p*-value statement, one year on. The Royal Statistical Society. Significance 14(2): 38-40.

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

McShane, B., Gal, D. (2017): Statistical Significance and the Dichotomization of Evidence. Journal of the American Statistical Association 112(519): 885-895 (DOI: 10.1080/01621459.2017.1289846).

McShane, B., Gal, D., Gelman, A., Robert, C., Tackett, J.L. (2017): Abandon Statistical Significance. http://www.stat.columbia.edu/~gelman/research/unpublished/abandon.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. (2012): A 21 word solution. Dialogue. The Official Newsletter of the Society for Personality and Social Psychology 26(2):4-7.

Trafimow, D. et al. (2017): Manipulating the alpha level cannot cure significance testing – comments on "Redefine statistical significance". PeerJ Preprints 5:e3411v1 (https://doi.org/10.7287/peerj.preprints.3411v1).

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.