Statistics Surveys
Vol. 12 (2018) 136–172
ISSN: 1935-7516
https://doi.org/10.1214/18-SS122

Pitfalls of significance testing and p-value variability: An econometrics perspective

Norbert Hirschauer

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), Germany
e-mail: 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, D-06120 Halle (Saale), Germany
e-mail: sven.gruener@landw.uni-halle.de

Oliver Mußhoff

Georg August University Goettingen
Department for Agricultural Economics and Rural Development, Farm Management
Platz der Göttinger Sieben 5, D-37073 Göttingen, Germany
e-mail: Oliver.Musshoff@agr.uni-goettingen.de

and

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), Germany
e-mail: claudia.becker@wiwi.uni-halle.de

Abstract: Data on how many scientific findings are reproducible are generally bleak and a wealth of papers have warned against misuses of the p-value and resulting false findings in recent years. This paper discusses the question of what we can(not) learn from the p-value, which is still widely considered as the gold standard of statistical validity. We aim to provide a non-technical and easily accessible resource for statistical practitioners who wish to spot and avoid misinterpretations and misuses of statistical significance tests. For this purpose, we first classify and describe the most widely discussed ("classical") pitfalls of significance testing, and review published work on these misuses with a focus on regression-based "confirmatory" study. This includes a description of the single-study bias and a simulation-based illustration of how proper meta-analysis compares to misleading significance counts ("vote counting"). Going beyond the classical pitfalls, we also use simulation to provide intuition that relying on the statistical estimate "p-value" as a measure of evidence without considering its sample-to-sample variability falls short of the mark even within an otherwise appropriate interpretation. We conclude with a discussion of the exigencies of informed approaches to statistical inference and corresponding institutional reforms.

JEL codes: B41, C18.
Keywords and phrases: Meta-analysis, multiple testing, p-hacking, publication bias, p-value misinterpretations, p-value sample-to-sample variability, statistical inference, statistical significance.

Received November 2017.

1. Introduction

Data on how many scientific findings are reproducible are generally bleak and a wealth of papers have warned against false findings in recent years. The reasons for false discoveries are manifold, but misuses and misinterpretations of statistical significance testing based on p-values are the most prominently decried ones. In the light of prevalent and persistent misunderstandings, even the American Statistical Association (ASA) felt compelled to issue a warning that the p-value can neither be used to determine whether a scientific hypothesis is true nor whether a finding is important (Wasserstein and Lazar 2016). In a *Nature* paper, Baker (2016: 151) comments: “This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics, says executive director Ron Wasserstein. The society’s members had become increasingly concerned that the P value was being misapplied in ways that cast doubt on statistics generally, he adds.” The ASA apparently considers inappropriate interpretations and uses of significance tests so serious a threat to statistics and science in general that, as a follow-up to its statement, it organized a symposium under the heading *Scientific Method for the 21st Century: A World beyond* $p < 0.05$ in fall 2017.

While a great variety of misuses and misinterpretations of the p-value are addressed in the literature concerned with the “reproducibility crisis,” we believe that they can be best systematized using a typology of four categories:

(1) One group of papers focus on **multiple testing that is left uncorrected for** in many studies. This leads to inflated claims of statistical significance. While the problem is well known in experimental research, it also arises in the regression-based analysis of observational data that is widely used in the economic and social sciences.[1] Even though multiple testing is *evident* in multiple regression analysis whenever researchers independently perform and interpret more than one test on one data set, many scientists ignore the problem

[1] In designed experiments, multiple hypothesis testing arises when we test several “contrasts of interest” (multiple treatments, multiple subgroups, multiple response variables, etc.). It also arises in non-experimental confirmatory analysis using multiple regression when, following widespread but often debatable practice, several hypotheses are subjected to “significance testing” one by one as if they were independent. Romano et al. (2010) note that it is quite common in empirical economic research to fit a multiple regression model and test several coefficients against the null. While there would be no multiple testing problem if one focused a priori on one hypothesis, the multiple testing problem (inflation of evidence) arises if one searches the list of p-values for significant results a posteriori. Romano et al. (2010) deplore that the latter case is much more common.

due to the "conventionality" of the model. Imagine a set of 10 hypotheses that are subjected *one by one* to significance testing. Assume the 10 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 statistically significant coefficient at the conventional threshold of $p^* = 0.05$ (Altman and Krzywinski 2017). Even more disastrous is *covert* multiple testing in conjunction with selective reporting. Testing alternative data, hypotheses, or analytical variants, and reporting only what has produced low p-values has been coined "p-hacking." Simmons et al. (2011: 1359) note that it is common "to explore various analytic alternatives, to search for a combination that yields 'statistical significance', and to then report only what 'worked'. The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%."

(2) A second class of papers tackle the **semantically induced misunderstandings** (cognitive biases) that the delusive language of frequentist statistics causes even among scientists. This group of papers stress the limitations of statistical significance testing that are present even if there are no multiple testing problems. They emphasize that the frequentist p-value concept can do much less to inform us about the reliability of scientific findings than what the persistently recurring colloquial associations with statistical terms such as significance, error probability, confidence interval, and inference suggest. This was also highlighted in the ASA-statement by Wasserstein and Lazar (2016) who stated that the p-value is neither the probability of a hypothesis nor a good measure of the evidence regarding a model or hypothesis. In the words of Greenland et al. (2016: 337), the departure point of these papers can be characterized as follows: "Misinterpretation and abuse of statistical tests, [...] have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so – and yet these misinterpretations dominate much of the scientific literature."

(3) A third class of papers is concerned with the **exaggerated focus on one-shot studies** (single-study bias) and the question of how to summarize the findings of individual studies to obtain an appropriate picture of the state of knowledge in a given field (meta-analysis). Borenstein et al. (2009: xxi) state that "rather than looking at any study in isolation, we need to look at the body of evidence." The essence of meta-analysis is best explained by comparing it to narrative reviews that simply count the studies that were declared statistically significant, or not, at the arbitrary threshold of $p^* = 0.05$. Contrasting the tallies ("vote counting"), together with the mistaken belief that studies with p-values on opposite sides of the conventional level of 0.05 are conflicting (Goodman 2008), leads to a wrong picture of the body of evidence. Borenstein et al. (2009: 14) claim that doing narrative reviews boils down to "*doing arithmetic*

with words" and that "when the words are based on p-values *the words are the wrong words*." They contend that this problem practically "gallops" through many research fields. In contrast, meta-analysis addresses the question of how to arithmetically synthetize meta summary statistics (usually the meta effect size and its p-value and confidence interval) based on the statistics provided in individual studies.

(4) Looking at the **publication bias** (Smith 1980) or file drawer problem, a fourth class of papers focus on researchers' incentives and the perverse effects of the scientific publishing system with its pressure to "publish [statistically significant results] or perish." Even if there were no inflation from multiple testing and no misinterpretation of statistical significance tests, the file drawer effect would distort the body of evidence towards results that can be declared "statistically significant." Starting with Sterling (1959) as an early precursor, researchers have increasingly realized this bias in recent years. Even back in the 1970s, Rosenthal (1979: 638) vividly described the harmful consequences that result from the preferences of researchers, reviewers and publishers for "significant" novelties: "The extreme view of this problem, the 'file drawer problem,' is that the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., $p > .05$) results." The file drawer effect does not only lead to the publication of unsubstantiated claims that are too rarely subjected to independent scrutiny. It may also lead to wrong medical treatment or policy recommendations with dire practical consequences.

Table 1 provides a summary of the pitfalls marked up above. While we will describe them one by one, it should be noted that they are intimately linked: first, misuses and misinterpretations often occur consecutively and may reinforce each other and accumulate. Second, mistakes and distortions in the research process may render subsequent procedures useless even if they are appropriate as such. For example, even the most elaborate meta-analysis aimed at consolidating the available evidence from prior studies will yield nonsense results in the presence of serious publication bias since it simply consolidates the distortion (Kline 2013: chapter 9).

We have attached the label "classical" to the pitfalls in Table 1 because they have already been widely discussed in the past. Despite the large body of literature that has accumulated over the last decades on these issues (Oakes 1986; Cohen 1994; Nickerson 2000; Ioannidis 2005; Armstrong 2007; Simmons et al. 2011; Motulsky 2014; Hirschauer et al. 2016; Wasserstein and Lazar 2016, and many others), misapplications of statistical significance testing continue to be an alarmingly "normal" practice for many scientists. With few exceptions (e.g., Krämer 2011; Ziliak and McCloskey 2008), the acknowledgement of the critical issues in statistical significance testing seems to be particularly low in economics. This can be partly ascribed to the fact that papers concerned with the pitfalls of significance declarations are not only scattered over disciplines often remote from economics but also focused on experimental approaches aimed at analyzing differences between treatment groups. Hence, an easily accessible informational resource for economists, who heavily rely on the regression-based

Table 1
Classical pitfalls of significance testing

	Flaws when performing significance tests	**Mistakes when interpreting significance tests**
Within study	**(1) Uncorrected multiple testing**(= inflated claims of statistical significance) • Unintentional disregard of evident multiple testing • *p-hacking*: covert testing of multiple analytical variants and selective reporting of those that yielded "statistical significance"	**(2) Semantically induced misinterpretations** • *Inverse probability error* (interpreting the p-value as the probability of the null) • *Sizeless stare* or even equation of significant effects with large or important effects • *False dichotomy* (interpreting not statistically significant results as confirmation of the null)
Across studies	**(3) Exaggerated focus on one-shot studies** (= disregard of prior knowledge) • Lacking meta-analysis • Improper meta-analysis (*vote counting*) • Lacking Bayesian analysis	**(4) Publication bias/file drawer effect** (= distortion towards positive results) • Selective preparation and submission • Selective reporting (*p-hacking*) • Selective publishing

study of observational data, is lacking. Furthermore, the p-value's sample-to-sample variability, even though it is a fundamental feature that severely limits its suitability to indicate the strength of evidence, has been underexposed in the economic literature so far (Berry 2016; Halsey et al. 2015).[2]

While some disciplines have adopted substantial reforms to abate inferential errors over the last decades, institutionalized reform efforts are at rather moderate levels in the social sciences. In clinical drug trials, for example, pre-registration and replication studies have by and large become a disciplinary standard (`http://www.who.int/ictrp/network/primary/en/`). Pre-registration precludes the confusion of "confirmatory"[3] and exploratory analysis because researchers are prevented from inflating statistical significance claims through

[2] A p-value is a summary statistic of the data. If the process that produced the data justifies using probability theory to interpret a p-value (e.g., in the case of random sampling or randomized assignment of treatments to units), then the p-value itself has a sampling distribution. We cannot tell from a single sample where our observed p-value falls in its distribution. It may come from the middle of the distribution or from one of the tails.

[3] The established term "confirmatory analysis" might itself be misleading because it is logically impossible to "confirm" a hypothesis and infer from the p-value whether the null hypothesis or an alternative hypothesis is true. Rather than introduce new methodological terminology, we nonetheless retain this term because it is widely used to describe the antipode of "exploratory analysis."

p-hacking. The low level of institutionalized reforms in the social sciences is not only due to a limited problem awareness but also the fact that the regression-based analysis of non-experimental data exhibits many characteristics (e.g., pre-existing data, specification search, predominance of bottom-up research and resulting model heterogeneity) that pose practical challenges for the effective adoption of approaches from other fields such as pre-registration, systematic replication, and meta-analysis.

When discussing the pitfalls of statistical significance testing, we should remember that the "p-value approach" was originally developed for *experimental research* with randomized assignment of treatments to units and the subsequent interpretation of differences between (small) treatment groups based on the *analysis of variance* (Fisher 1925; 1935). At present, however, it is broadly applied across many scientific fields and research contexts. It is important to conceptually distinguish the different research contexts: besides hypothesis testing in experiments, where the intended inference is about the causal effects of treatments, it is also used for testing hypotheses in the *regression-based study of observational data*, where the target of inference is about generalizing from a random sample to its population;[4] and besides *hypothesis testing* (*confirmatory study*), the p-value is also used as flagging device in the *exploratory search* for new hypotheses (*hypothesis generation*).

Against this background, we aim to provide a non-technical and easily accessible resource for statistical practitioners (with a focus on social scientists including economists) who wish to spot and avoid misinterpretations and misuses of statistical significance tests. Our paper focuses on misuses and misinterpretations of the p-value in **regression-based confirmatory studies** with a focus on statistical significance testing ("hypothesis testing") in the analysis of (non-experimental) **observational data**. We touch upon exploratory study ("hypothesis generation") only to the extent to which it is necessary to show how dangerous it is to blur the dividing line between exploratory and confirmatory study, and to point out what the latter must refrain from. Furthermore, despite the different types of inference, and despite our focus on the regression-based analysis of observational data, we address experimental approaches to the extent to which inferential errors (e.g., poor understanding of the summary statistic "p-value" itself, disregard of multiplicities, single-study bias, publication bias) arise in both research contexts.

A better understanding of inferential errors associated with the p-value will not only foster the logical consistency of inferential arguments in individual future studies but also help identify remedies in terms of organized approaches aimed at increasing the quality of published research. We therefore first provide a systematic overview of the "classical" pitfalls of statistical significance testing in section 2. This includes a simulation-based numerical illustration of

[4] The justification for using the p-value includes measurement error: in experiments, the error term and thus the standard error in the statistical model can reflect the fluctuations due to random assignment *or* measurement error. Similarly, in the random sampling case, the error term can be due to sampling error *or* measurement error. Conventional statistical significance testing does not distinguish between these two but addresses them jointly.

how meta-analysis could counteract the single-study bias. To complete the understanding that the p-value is a descriptive summary of a given data set but a poor tool for making direct inductive inferences in terms of obtaining an epistemic probability that a hypothesis is true (Goodman 2008), section 3 focuses on the p-value's variability over replications. It uses simulation to provide intuition that relying on the p-value as inferential tool without considering its sample-to-sample variability falls short of the mark even within an otherwise correct frequentist interpretation. Section 4 concludes with a discussion of what can be done on the institutional level to improve the quality of published research and statistical inference.

2. Classical pitfalls in significance testing

2.1. Uncorrected multiple testing

Unintentional disregard of evident multiple testing

Confirmatory studies based on experiments regularly analyze how various treatments (multiplicity of treatments) affect various outcomes (multiplicity of effects) across various subgroups (multiplicity of groups). Contrary to a *single* test, where the p-value denotes the probability of falsely rejecting the null when it is valid, conventional p-values do not reflect the probability under the null in the case of *multiple* testing. While adjustments for multiple tests on the same data set are common in many other areas of science, it is still, "with a few exceptions, [...] uncommon for the analyses of these data to account for the multiple hypothesis testing in economic experiments" (List et al. 2016: 1).

While even less accounted for, multiple testing is also present in confirmatory studies based on the *multiple* regression analysis of observational data whenever *multiple* tests are performed on one data set, i.e., whenever several hypotheses are tested one by one (see footnote 1). This inflates claims of statistical significance analogous to multiple testing in experimental study. Let's take a closer look why. Adopting the frequentist null hypothesis testing assumption of no association, significance testing of multiple regression coefficients represents a Bernoulli experiment: we independently repeat, for each regressor, a trial with two possible outcomes "significant" S and "not significant" $\overline{S}$, with probabilities $P(S) = 0.05 = p^*$, and $P(\overline{S}) = 0.95 = 1-p^*$. The chance of finding "significant" coefficients despite a fully random data structure is provided by the binomial distribution $B_{m,p^*}(k)$, with m indicating the number of regressors and thus null hypotheses, $p^* = 0.05$ the probability of falsely claiming significance, and $k \in \{0, \ldots, m\}$ the number of mistaken significance declarations (Altman and Krzywinski 2017).

Table 2 illustrates how multiple tests inflate claims of statistical significance. The implications are woeful if adjustment requirements are ignored. In this case, a researcher might be convinced of having found several "significant" results at the conventional level of $p^* = 0.05$ without realizing that (s)he faces an inflated

Table 2

Probability of k false significance declarations at the $p^* = 0.05$ *threshold in multiple regressions with independent hypotheses tests and completely random data structure*

	Number of multiple tests (null hypotheses)					
	$\boldsymbol{m = 1}$	$\boldsymbol{m = 2}$	$\boldsymbol{m = 3}$	$\boldsymbol{m = 5}$	$\boldsymbol{m = 10}$	$\boldsymbol{m = 20}$
$B_{m,0.05}\,(0) = \boldsymbol{P(k = 0)}$	0.95	0.903	0.857	0.774	0.599	0.358
$B_{m,0.05}\,(1) = \boldsymbol{P(k = 1)}$	0.05	0.095	0.135	0.204	0.315	0.377
$B_{m,0.05}\,(2) = \boldsymbol{P(k = 2)}$	–	0.003	0.007	0.021	0.075	0.189
$\boldsymbol{P\,(k \geq 1) = 1 - 0.95^m}$	**0.05**	**0.098**	**0.143**	**0.226**	**0.401**	**0.642**
Inflation factor: $P\,(k \geq 1)\,/0.05$	1.000	1.950	2.853	4.524	8.025	12.830

probability of 0.098 (0.143, 0.226, 0.401, 0.642) of finding at least one significant coefficient in a multiple regression with 2 (3, 5, 10, 20) regressors even though there is no association whatsoever in the data.

p-hacking: covert testing of multiple analytical variants and selective reporting

The arbitrary dichotomization of results into "significant" and "not significant," in conjunction with researchers' self-interested desire to obtain findings that can be declared "significant" is considered a major cause of the reproducibility crisis (cf., e.g., Amrhein et al. 2017; Berry 2017; Gelman and Carlin 2017; Greenland 2017; MacShane et al. 2017; Trafimow et al. 2017). This holds for both experimental and observational study. The term "*p*-hacking" has been coined to describe the behavior of researchers who try a multiplicity of analytical alternatives and then report only the one that produced the desired result (Simmons et al. 2011). In *p*-hacking there are several noteworthy features: first, the list of possible analytical alternatives is near endless in most research contexts. Second, problem awareness among researchers is frequently low, especially in the analysis of observational data, due to the ambiguity of how much specification search is appropriate.[5] Third, the multiple analytical variants that can be tried

[5] In designed experiments, we purposefully generate data to answer a given research question and (at least ideally) use a *pre-specified analytical model* that is not altered after seeing the data ("being blind to the data"). In contrast, regression-based analysis of observational data must get along with *pre-existing data* or survey data that are often not particularly well-suited to answer a given research question using a pre-specified model. It is therefore often considered appropriate or even imperative to fit the analytical model to the data (specification search). A simple example is when the violation of a distributional assumption is discovered and rendered "innocuous" by a log-transformation. A much more critical instance of altering a model after seeing the data is the removal of variables from an initial model because of multi-collinearity, i.e., high correlations between regressor variables that result in high standard errors and *p*-values. Omitting correlated variables reduces the *p*-values of the remaining coefficient estimates and thus invites overconfident inferential conclusions. Even worse, it is likely to bias our estimates (omitted variable bias). In other words, the very fact that in non-experimental confirmatory analysis it is not considered completely illegitimate to modify an analytical model after seeing the data creates analytical ambiguities (a gray area between "appropriate model fitting" and "*p*-hacking") that, as Simmons et al. (2011) note, facilitate the unconscious self-justification of choices that mesh with researchers' desire to obtain low *p*-values.

often seem to have little in common at first view. They share one noxious quality, however: selective reporting of covert multiple tests may disastrously inflate claims of statistical significance.

The literature concerned with the pitfalls of significance tests and declarations has extensively discussed the various forms of *p*-hacking (see Hirschauer et al. 2016 for an overview). We do not intend to summarize this literature once more. However, to facilitate the understanding in which ways *p*-hacking represents a covert and thence especially harmful type of multiple testing, we briefly describe the four main forms of *p*-hacking that can be distinguished.

a) **Testing multiple data sets:** Researchers might be tempted to explore whether *p*-values can be reduced when the number of units in a sample is manipulated. There are several possibilities: the reduction of sample size, for example through a tentative elimination of outliers, is one possibility. Another one is the increase of sample size if an original sample yielded "disappointing" *p*-values. A general feeling that larger samples are better may impede the awareness that this constitutes a test of multiple sample sizes. Finally, researchers might try which *p*-values they can obtain when analyzing multiple data subsets (subgroups, cf., List et al. 2016). Separately testing a treatment in 20 arbitrary subgroups and displaying only significant results, for example, leads to "false positive probabilities" as shown in the last column of Table 2 even when the effect is nil in all subgroups (Kerr 1998; Motulsky 2014).

b) **Testing multiple data transformations:** Researchers might also be tempted to check which (combination) of many conceivable data transformations produces lower *p*-values than the original data. Possibilities are plentiful: downgrading of measurement scales (e.g., age classes instead of age in years), log-transformation, squaring, and the synthetization of various variables including ratios and interaction terms. Some of these manipulations may be claimed to be statistically appropriate in the light of the theory, research question, and data – for example when researchers realize, after seeing the data, that distributional assumptions are violated (see footnote 5). They inflate claims of statistical significance, however, if they are driven by a researcher's significance-pursuing behavior.

c) **Testing multiple variable sets:** Besides the number of units in a sample, researchers might also be tempted to try out different predictor and response variables. In experimental study, this implies pronouncing statistical significance for selected results after a multiplicity of treatments were tested for a multiplicity of outcomes. Regression-based confirmatory analysis may be threatened by similar practices. The choice of variables to be included in a regression is often ambiguous: Which theoretical (latent) constructs are to be included as predictor variables? Which manifest variables (e.g., survey items) should be used to measure these constructs? Which control variables should be considered? Mining for a combination of variables that yields low *p*-values inflates statistical significance claims. A researcher who studies, for example, how people's attitudes towards organic farming affect their willingness to pay for organic products produces a distortion if (s)he keeps on searching for a survey item for “attitude” until (s)he finds one that produces a “significant” result.

d) **Testing multiple estimation models:** Selecting statistical tests and models also offers ample scope for decisions that “improve” p-values. For example, when facing an ambiguous choice of whether to use a simple OLS estimator or a panel data model, it would be good scientific practice to transparently compare the results of both models. However, the rules of good scientific practice are occasionally broken and data analyses are not performed as planned in a prior study design but ad hoc adjusted according to the criterion of which analytical model yields low p-values. Scientific transparency is lost when the results of competing models are neither explicitly reported nor comparatively discussed.

Disregarding multiplicities and notably p-hacking are problems that arise when confirmatory and exploratory study are conceptually mixed up. Exploring potentially interesting associations among a set of variables as well as explorative model fitting and variable selection can be an adequate firststep of the research process. In exploratory study, p-values may help identify what might be worth investigating with new data in the future (Head et al. 2015; Motulsky 2014).[6] This exploratory exercise, however, must be clearly distinguished from confirmatory analysis; i.e., the exploratory search for new hypotheses must not be presented as hypothesis testing after results are known (HARKing; cf., Kerr 1998). Attaching the label “significance *test*” to p-values in exploratory studies is misleading per se since no *testable* hypotheses exist. As reminder for researchers and readers alike, Berry (2016: 2) thence suggests to include a “black-box warning” into all exploratory studies: “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.” In contrast, not accounting for multiple testing in confirmatory analysis inflates significance claims. It is nonetheless common in economics (Romano et al. 2010) – even though ever-increasing numbers of models are tested and variables included in regression models due to increased computing power and better data availability (Ioannidis and Doucouliagos 2013).

Multiple testing adjustment requirements

We engage in multiple comparisons every time we perform and interpret significance tests for more than one statistical hypothesis based on *one* data set. Covert multiple testing (p-hacking) is hard to detect and overcome because it withholds the information that is needed to correct for multiplicities. By contrast, adjustment tools are available for those who are ready to account for the multiple tests they have made in accordance with their research interests. If, in

[6] This seems to be what Fisher had in mind when noting that low p-values *signify* “worth a second look” if we have little to no prior knowledge (cf., Gigerenzer et al. 2004; Lecoutre and Poitevineau 2014; Nuzzo 2014).

Table 3
Multiple testing adjustments for $m = 3$ consistent with a single-test threshold of 0.05

Conventionally computed ("raw") p-values Rank r of raw$p$-values	0.01 1	0.03 2	0.04 3
Family wise error rate (FWER) corrections			
Significance threshold after Bonferroni adjustment: $0.05/m$	**0.0167**	0.0167	0.0167
Significance threshold after Bonferroni-Holm adjustment: $0.05/(m + 1 - r)$	**0.0167**	0.0250	0.0500
False discovery rate (FDR) correction			
Significance threshold after Benjamini-Hochberg adjustment: $0.05 \cdot r/m$	**0.0167**	**0.0333**	**0.0500**

a multiple testing setting, we are interested to make claims of statistical significance that are comparable to the one we would make for a single significance test (i.e., at a threshold comparable to the conventional $p^* = 0.05$), we have to adjust the significance levels.[7]

Both the Bonferroni and the Bonferroni-Holm adjustment are used to control the so-called *family wise error rate* (FWER or multiple type I error rate). The logic behind the FWER correction is to restrict the probability of rejecting even one null hypothesis when it is true, irrespective of how many of the other null hypotheses are valid (Hochberg and Tamhane 1987; Pigeot 2000). The computationally easiest way to arrive at a family wise error rate is the Bonferroni correction (see Fisher 1935). Assuming the conventional significance threshold of 0.05, the Bonferroni correction implies that the adjusted threshold of $0.05/m$ is used for each of a total number of m tests. For large m, this leads to extremely small significance levels. This, in turn, usually results in only few null hypotheses being rejected. In our example for $m = 3$ tests (see Table 3), only the effect associated with the "raw" p-value of 0.01 would be declared significant since the adjusted threshold is 0.0167. Hence, alternative and less conservative corrections have been suggested. The Bonferroni-Holm adjustment (Holm 1979) is an example. It calculates different thresholds for each one of the m tests. The respective threshold is computed according to the formula $0.05/(m + 1 - r)$ and increases with the rank r of the raw p-values. The rationale is to identify the smallest raw p-value that is above the adjusted threshold and then declare all smaller raw p-values "significant." Following this rule in our example, we would still only declare the raw p-value of 0.01 "significant" even though the Bonferroni-Holm thresholds are generally less rigorous than the Bonferroni thresholds.[8]

[7] Although adjusting the significance level is necessary in most cases of multiple testing, there are special constellations where an inherent adjustment takes place within the very system of tested hypotheses. This is the case for so-called *closed families of hypotheses* in the special case of a coherent test procedure. The least-significant-differences test of Fisher (1935) in case of one-way ANOVA falls into this class when three groups are compared, but not for more than three groups. For details, see Pigeot (2000) or Didelez et al. (2006).

[8] Another concept, the *global level of simultaneous tests*, can be directly linked to the binomial distributions as shown in Table 2. The global level of simultaneous tests is aimed at restricting the probability of rejecting one or more of the tested null hypotheses when *all of them* are valid.

A different perspective is adopted when using the concept of the *false discovery rate* (Benjamini and Hochberg 1995) according to which different significance levels for each of the m tests are computed based on the formula $0.05 \cdot r/m$. Here, the rationale is to identify the largest raw p-value that is below or equal to the adjusted threshold and then declare this and all smaller raw p-values "significant." The FDR correction produces less rigorous thresholds than the FWER correction. With the FDR rule, we would declare all three raw p-values "significant." To provide intuition, one could say that the FDR is aimed at restricting the rate of valid null hypotheses being rejected relative to the total number of rejected hypotheses. One should keep in mind that, despite its delusive naming, the FDR (along with the other non-Bayesian multiple testing adjustments) shares the crucial limitation of the frequentist p-value approach: it does not estimate the post-study probabilities of hypotheses given the data. Analogous to the p-value, the frequentist FDR cannot work backwards and inform us about the probability of real-world phenomena (see section 2.2) – even though we as researchers would want to have that kind of (inherently Bayesian) information to assess the trustworthiness of our inductive conclusions.[9]

Contrary to multiple variables of interest that are subjected one by one to significance testing, control variables, which often populate regression models in large numbers, need not to be considered in multiple testing adjustment. Instead, they have to be clearly identified as control variables and separated from the m variables of interest. Disregarding control variables is only adequate, however, if we explicitly distinguish between confirmatory and exploratory analysis. This may require to clearly divide studies in two parts: a confirmatory part where we perform a pre-defined number of multiple tests m, and an exploratory part where we look at potentially interesting correlations in the control variables. As has been said before, conventional p-values are an adequate focusing aid to identify what might be worth investigating with new data. If this exploratory search is not presented as confirmatory, which itself would be p-hacking (illegitimate hidden testing) and HARKing (cf., Kerr 1998; Gigerenzer and Marewski 2015; Motulsky 2014), no multiple testing corrections are needed.

Besides multiple hypotheses tested on the same data set in a given multiple regression, an additional multiplicity problem may arise due to specification search. In other words, it was easy to determine the number of tests m that must be corrected for in our multiple regression example where we assumed that a defined number m of pre-defined hypotheses were tested separately. However, in many research settings, the determination of the number of multiple tests and even the need for the consideration of multiple testing may be less obvious. While obfuscating the dividing line between confirmatory and exploratory research, many observational studies engage to some extent in model fitting and retain one model as the final ("best") model after dif-

[9] To avoid confusion, the *frequentist* false discovery rate must be clearly distinguished from the term's Bayesian interpretation as used, for example, by Hirschauer et al. (2016) and Motulsky (2014). The *Bayesian* false discovery rate (or: Bayesian error rate) is the post-study probability of "no effect" and therefore the probability of a faulty scientific claim when rejecting the null.

ferent models have been evaluated using some measure of model fit (see footnote 5). This gives rise to many questions: how should we deal with variables including controls, interaction terms, higher-order polynomials etc. that are included in the regression besides the original variables in the course of the analysis? How should we consider tests of multiple model specifications in the first place? The answer is straightforward in principle: whenever we perform multiple tests on *one* data set, they need to be considered to prevent the inflation of statistical significance claims. While we know that we arrive at overconfident conclusions if we assess the strength of evidence in only the "best" model even though multiple alternatives had been tested (Forstmeier et al. 2016; Simmons et al. 2011), we often lack the knowledge of *how many* analytical variants have been tried out after seeing the data. This problem would be mitigated if confirmatory research were completely based on pre-registration, where the research hypotheses as well as the statistical model to test these hypotheses or at least the analysis plan would be transparently specified *before* seeing the data (Nosek et al. 2018).

We may summarize that the appropriate multiple testing adjustment depends on the research setting. It is an often ambiguous choice on which we cannot further elaborate in this paper. For a general overview and description of the great variety of multiple comparison adjustments and their eligibility criteria the reader is referred to Bretz et al. (2010) or Westfall et al. (2011).

2.2. Semantically induced misinterpretations of the p-value

Even if significance claims are not inflated, serious misunderstandings lurk due to widespread cognitive biases in the interpretation of conditional probabilities and the delusive technical terms of frequentist statistics that contradict everyday language. The fact that a vast body of literature has decried these misinterpretations over the last four decades (see Hirschauer et al. 2016 for an overview), has apparently been of little avail. The ubiquity and persistence of faulty interpretations of both experimental and observational data are, not least, caused by the fact that they have been perpetuated over decades through academic teaching and even through best-selling statistics textbooks (Haller and Krauss 2002; Gigerenzer et al. 2004; Lecoutre and Poitevineau 2014; Krämer 2011; Nickerson 2000).

In regression-based econometric applications, the most common misinterpretations of the p-value are best understood when realizing that the majority of analyses are performed as if using the following misleading guidelines: (1) Run a multiple regression. (2) Compute coefficients and p-values. (3) Declare coefficients with p-values below a threshold (usually 0.05) "statistically significant." (4) Do not reflect on the arbitrary dichotomization of results into "statistically significant" and "not statistically significant." (5) Implicitly attach a high trustworthiness or probability (possibly even $1 - p$) to statistically significant coefficients. (6a) Do not discuss the effect size or (6b) even suggest that a "significant effect" is also large or important or (6c) simply claim the just-estimated effect to be real. (7) Attach the label "statistically nonsignificant" to coefficients with p-values above 0.05 instead of noting that they are "not statistically significant" at the arbitrary threshold. (8a) Interpret your "statistically nonsignificant" results as minor or not noteworthy or (8b) even suggest that they can be interpreted as a proof of no effect.

As a consequence of such misleading practices, researchers and readers alike will not only overrate the inferential content of findings that have been declared "statistically significant" but also of findings that have been declared "statistically nonsignificant." These dichotomous declarations will make them draw similarly dichotomous existence/relevance conclusions that – to borrow a quote from H.L. Mencken in the New York Evening Mail from November 16, 1917 – are "neat, plausible, and wrong." The first neat but wrong conclusion is that effects with the label "statistically significant" can be considered to be real (and may be large) with a high probability. The second neat but wrong conclusion is that the label "statistically nonsignificant" is an indication or even proof of no or little effect. Interpretations along this erroneous dichotomy is entrenched practice for many researchers, a practice that is based on deeply engrained beliefs that result both from wishful thinking (desire for a "neat" interpretation) and the inevitably wrong connotation of the word "significance" in everyday language.

Unfortunately, the p-value, if correctly interpreted, has much less inferential content than what colloquial associations with the terms "statistically significant" and "statistically nonsignificant" suggest (Berry 2016). Following the misleading guidelines from above mirrors a serious misunderstanding of what the p-value, which is merely a summary statistic of a given data set, can tell about reality. It means falling prey to three fallacies that have been given distinct names: Cohen (1994) used the term *inverse probability error* to describe the belief that the p-value is the conditional probability of (falsely rejecting) the null hypothesis given the data under study. Instead, the p-value is the conditional probability of finding the observed effect (or even a larger one) in random replications *if*, as a thought experiment, we assumed the null hypothesis to be true. Per definition, it cannot work inversely and inform us on the underlying reality. But looking for answers to their scientific questions about reality, statistical practitioners often confuse frequentist and Bayesian probabilities and adopt a Bayesian interpretation of frequentist measures such as the p-value.[10] McCloskey and Ziliak (1996) coined the expression *sizeless stare* for the disregard of effect size or the implicit equation of statistical significance with relevance. In the light of the increasing availability of large samples, the naïve equation of significance with relevance becomes more and more misleading because any effect, even if very small and irrelevant, eventually becomes statistically signifi-

[10] Cohen (1994: 997) succinctly described the harmful mixture of wishful thinking and semantic confusion that causes the inverse probability error: "[the p-value] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is 'given these data, what is the probability that H0 is true?' But [. . .], what it tells us is 'given that H0 is true, what is the probability of these (or more extreme) data?' "

cant in large samples.[11] Hirschauer et al. (2016) used the term *false dichotomy* to describe the logical fallacy that misleads people to first adopt an ill-founded either-or perspective and then use the ensnaring label "statistically nonsignificant" that finally makes them interpret p-values above 0.05 (absence of statistical significance) as an indication or even confirmation of the null (evidence of absence).[12]

Colloquial associations that rashly equate "statistically significant" with "scientifically trustworthy" may furthermore prevent researchers from realizing that a meaningful interpretation of the p-value requires acknowledging its probabilistic nature. Vogt et al. (2014: 242; 244) note that the classical tools for statistical inference (including p-values) are inherently based on probability theory. They conclude that "in research not employing random assignment or random sampling, the classical approach to inferential statistics is inappropriate. [...] 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."

In some instances, it is not clear whether we have a random sample or not. Denton (1988: 166f.) notes that "where there is a sample there must be a population." He points out that conceiving of the population can be difficult. The easiest case is a sample drawn from a finite population such as the entirety of a country's citizens. Slightly less intuitive is an experiment such as flipping a coin. Here, the population is an imaginary set of coin flip experiments that are infinitely repeated under constant conditions. More conceptual challenges arise in the case of observational data that are not a proportion of a larger population but rather seem to represent the whole population (e.g., the macro-data of a country). Here, the frequentist statistician must introduce an infinite "unseen parent population" and a "generating process" from which we observe one random realization under noise. Emphasizing that there are many people who will not subscribe to the idea of a probability process underlying such observational data, Denton notes with regard to such circumstances: "However, some notion of an underlying process [...] has to be accepted for the testing of hypotheses in econometrics to make any sense" (Denton 1988: 167).

Using p-values in non-experimental study requires that the data represent (at least approximately) a *random sample* of a defined parent population, or that

[11] It is unlikely that any two real-world variables exhibit zero correlation. This is why Lecoutre and Poitevineau (2014: 50) call the null hypothesis a "straw man" that significance testing tries to knock down. Similarly, Leamer (1978: 89) notes 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."

[12] The inverse probability error and the false dichotomy fallacy, even though jointly found in many instances, are inconsistent in themselves. Committing the inverse probability error, one would believe that $p = 0.01$ indicates a 1%-probability of the null being true and thus a 1%-probability of making a false claim when rejecting it. Similarly, one would have to believe that a "nonsignificant" result with $p = 0.15$, for example, indicates a 15%-probability of the null. After this error, one cannot logically interpret a "nonsignificant" result as a confirmation of the null.

one accepts the notion that they are a *random realization* from an unseen parent population. In many contexts, the fact that even a truly random sample does not exactly reflect the properties of the population (random sampling error) is the least of worries. Data are frequently not obtained through random sampling but affected by various types of selection bias. These problems are often more serious than sampling error. Not having a random sample in observational study has fundamental implications for the utilization of the p-value that, arithmetically, can be computed for any data: (1) The p-values computed from non-random samples have no meaningful interpretation. (2) The same holds when the sample itself represents a finite population of interest. In this case, random sampling error is completely eliminated because we already have the population properties. (3) The p-value in observational study can only be meaningfully interpreted if researchers explicitly define *from which* population the random sample has been drawn and thence *to which* population a statistical inference is to be made. (4) *Statistical* inference such as the generalization from a random sample to its population is only the first step of *scientific* inference. Scientific inference is the totality of reasoned judgments (inductive generalizations) that we (can) make in the light of our own observations and the available body of evidence found elsewhere. There is no easy-to-apply recipe or even calculus for making such judgements. We instead have to critically rethink each situation. When trying to answer the question, for example, of what we can learn from a sample of agricultural students for a country's student population or even its citizens, or even human beings in general, we must keep in mind that a p-value can do *nothing* to assess the generalizability of a result beyond the parent population from which the random sample has been drawn.

2.3. Exaggerated focus on one-shot studies

Economic analyses are frequently based on multiple regression in which a "dependent" or response variable (usually denoted as y) is modeled as a function of several "independent" variables (usually denoted as x_j). The independent variables are often divided into focal variables of interest (*focal predictors*) and variables that are used to control for confounding influences (*control variables*). The direct outcomes of multiple regression analysis are the estimated coefficients $\hat{\beta}_j$ (regression slopes) that relate the predictors x_j to response y. If several studies have tackled the same x–y relation (e.g., between education x_1 and wage earnings y), we are naturally interested in summarizing these findings.

In many areas of science, it is common to use methods known as meta-analysis to obtain a concrete picture of the existing knowledge regarding a specific research question. Contrary to narrative reviews, meta-analyses are aimed at computationally synthesizing the results from prior studies. Unfortunately, the heterogeneity of multiple regression models in economics often make quantitative meta-analysis a difficult if not impossible task. In what follows, we first outline the meta-analytical endeavors and its challenges in economics. We then use a simulation-based example to illustrate the functional principle of meta-analysis.

Arguably going back to Stanley and Jarrell (1989), some economic meta-analyses have been carried out on selected issues over the last decades. Zelmer (2003), Cooper and Dutcher (2011), Engel (2011), and Lange (2016) addressed economic experiments. Non-experimental meta-analyses were conducted by Card and Krueger 1995, Crouch (1995), Loomis and White (1996), Fitzpatrick et al. (2017), and Van Houtven et al. (2017). More than ten years ago, even a whole issue of the *Journal of Economic Surveys* (cf., Roberts 2005) was dedicated to the meta-analysis of regression coefficients. In the meanwhile, there is also a *Meta-Analysis of Economics Research Network* that provides a platform for economic meta-analyses. Nonetheless, the practice of meta-analysis is less common in economics than other fields.[13] What is more, even the statistical literature has mainly dealt with the synthetization of measures such as (standardized) mean differences or (risk or response) ratios that are primarily used in non-economic experimental fields (e.g., medical sciences). In contrast, summarizing coefficients from multiple regressions, which are the working horse of economists, has attracted less attention (Becker and Wu 2007). The limited use of meta-analysis in economics can be attributed to the fact that economic research is mainly a non-programmed bottom-up research exercise. As such, it produces an enormous quantity of empirical results on topical issues, but is also plagued by an enormous heterogeneity of regression model specifications (Bruns 2017). Attempts to summarize findings that deal with the same x–y relation (e.g., education-wage) are thus hampered by a deficient or even lacking comparability of the regression coefficients across prior studies.

The comparability of regression slopes across studies is severely constrained by some basic features found in most economic research fields: first, the metrics (scales and units of measurement) of independent and dependent variables usually differ across studies. Second, even if all variables are identically measured, using structurally different models implies that the estimated coefficients are usually beyond comparison. Third, even if identical metrics and model structures are used, comparability is jeopardized when models with different sets of independent (control) variables are estimated. Fourth, even if metrics and estimation models are similar, different studies may have drawn their samples from different populations. Consequently, there may simply be no data base to do a meta-analysis because each single study covers a different parent population. Regression coefficients for the education-wage relation in France, Ghana, and the US, for example, cannot be meaningfully synthesized into one summary coefficient.

We use the univariate *weighted least squares* approach to demonstrate what meta-analysis is about in principle. We assume that we are to summarize 20 individual studies based on different sample sizes ($n = 20, 30, 40, 50$). For the

[13] Even a cursory look at economic publications will show that the critique by Stanley and Jarrell (1989: 162) still applies: "The reviewer often impressionistically chooses which studies to include in his review, what weights to attach to the results of these studies, how to interpret the results, [...]. Traditionally, economists have not formally adopted any systematic or objective policy for dealing with the critical issues which surround literature surveys. As a result, reviews are rarely persuasive to those who do not already number among the converted."

sake of easy intuition, we avoid all complications, such as different metrics between studies, by simulating 20 samples (random realizations) from a "reality" characterized by the linear relationship $y = \beta_0 + \beta x + e$, with $\beta_0 = 1$, $\beta = 0.2$, $x \in \{0.5, 1.0, 1.5, \ldots, n/2\}$, and $e \sim N(\mu; \sigma)$, with $\mu = 0$ and $\sigma = 5$. In each single study, an OLS regression is used to estimate $\hat{\beta}_0$ and the focal coefficient $\hat{\beta}$.

Using the weighted least squares method, the summary coefficient $\hat{\beta}^{sum}$ that synthetizes the coefficients of the single studies is computed as follows (cf., Becker and Wu 2007: 7):

$$\hat{\beta}^{sum} = \sum_{i=1}^{I} \hat{\beta}_i \cdot w_i \Big/ \sum_{i=1}^{I} w_i, \qquad \text{with } w_i = 1/\widehat{SE}_i^2 \tag{1}$$

where I is the number of single studies and $\hat{\beta}_i$ is the coefficient estimated in the ith study. The weight w_i that is attributed to the coefficient from each study i is the reciprocal of its squared standard error estimate $\widehat{SE}_i^2$; and the ratio $w_i / \sum_{i=1}^{I} w_i$ denotes the percentage weight of each study. The standard error of the summary coefficient $\hat{\beta}^{sum}$ is:

$$\widehat{SE}^{sum} = \left(1 \Big/ \sum_{i=1}^{I} w_i\right)^{0.5}. \tag{2}$$

Table 4 describes the results of the meta-analysis. The p-values are computed based on the assumption of standard normally distributed test scores and one-sided tests. This reflects the assumption that the researchers who presumably had carried out the 20 previous studies had qualitative prior knowledge indicating a non-negative relation between x and y.

Several noteworthy findings and conclusions can be derived from Table 4:

1. A large majority of studies (14 in 20) have not found a statistically significant result. This might mislead narrative reviewers to contrast tallies and conclude that the results of these 20 studies represent contradictory evidence or even overall a confirmation of no effect.
2. Meta-analysis is capable of leaving behind the arbitrary either-or interpretation within each study. Instead, it synthesizes – with adequate weights – the informational content of all studies given the fact that "the effect best supported by the data from a given experiment [or random sample] is always the observed effect, regardless of its significance" Goodman (2008: 136).
3. The meta effect size $\hat{\beta}^{sum} = 0.195$ approximates the true $\beta = 0.2$ quite well. The meta p-value is 0.000000003. This shows that even a great majority of studies that are not statistically significant can together represent a "highly statistically significant" effect. This is due to the fact that meta-analysis is capable of including the informational content of studies even if they are too small to produce statistical significance.
4. Given the meta effect size and its very low p-value, we would be confident that the real-world level of β lies above 0 – and we would commonly refer

Table 4

Meta-analysis for 20 single studies, each based on a simulated random sample from a reality characterized by the x–y relation: $y = 1 + 0.2x + e$*, with* $e \sim N(0; 5)$

Study No. i	Observations n per study	Estimated coefficient $\hat{\beta}$	Standard error $\widehat{SE}$	p-value[a)]	Weight (%)
1		0.008	0.267	0.487	1.56
2		−0.050	0.366	0.554	0.83
3	20	0.138	0.424	0.373	0.62
4		−0.083	0.294	0.611	1.29
5		0.166	0.471	0.362	0.50
6		0.121	0.201	0.274	2.76
7		0.240	0.249	0.167	1.80
8	30	0.560	0.206	0.003*	2.63
9		0.441	0.205	0.016*	2.65
10		0.330	0.178	0.032*	3.51
11		0.244	0.156	0.059	4.58
12		0.141	0.123	0.125	7.38
13	40	0.253	0.127	0.023*	6.92
14		0.213	0.143	0.069	5.44
15		0.177	0.146	0.113	5.21
16		0.284	0.095	0.001*	12.44
17		0.165	0.120	0.085	7.72
18	50	0.076	0.085	0.186	15.50
19		0.268	0.120	0.013*	7.70
20		0.097	0.112	0.192	8.95
Meta-analysis of all 20 studies (total no. of observations: 700)		**0.195**	**0.033**	**0.000***	**100.00**
Single large regression (over all 700 observations)		**0.212**	**0.027**	**0.000***	–

[a)] p-values below the conventional threshold of $p^* = 0.05$ are indicated by *.

to its estimated value of approximately 0.2 even though we realize that the p-value does not provide a clear rationale or even calculus for statistical inference (Goodman 2008).

5. Low p-values do not indicate results that are "more trustworthy" than others. Considering only significant studies, for example, would introduce a distortion (see section 3) and we would find a summary coefficient of 0.3109. That is, the results of *all* studies jointly represent the body of evidence and are valuable and *necessary*, irrespective of their p-values, to provide an approximately correct picture of the real-world regularity.
6. The 20 single studies in our illustrative example were *not* distorted but based on 20 random realizations (simulations). If the single studies were distorted due to publication bias (see section 2.4), the basic *weighted least squares* method of meta-analysis, which is unable to control for such biases, would simply summarize the distortion.[14]

[14] While the *weighted least squares* approach is not able to control for publication bias, meta-regression (cf., Stanley and Doucouliagos 2012; Stanley and Jarrell 1989) has been suggested to control for the idiosyncrasies of model specifications in previous studies and notably for publication bias.

7. Being in the comfortable position to know all raw data, we also carried out a single large regression over all $n = 700$ observations that serves as benchmark for the meta-analytical calculus. The estimated $\hat{\beta}^{700} = 0.212$ from the large regression is slightly above the true effect size $\beta = 0.2$ and the computed meta effect size of $\hat{\hat{\beta}}^{sum} = 0.195$.

Finally, one should bear in mind that conventional meta-analysis stays within the confines of the frequentist approach: it does not provide probabilities of scientific propositions given the data, or, as Kline (2013: 307) notes with regard to experimental data, "a standard meta-analysis cannot answer the question, What is the probability that the treatment has an effect?" Only Bayesian methods can provide the post-study probabilities of scientific propositions that researchers and users of scientific results are ultimately interested in (see footnote 10). While it is possible to combine Bayesian methods with meta-analytical approaches (Howard et al. 2000; Kline 2013: 307), as the number of studies increases, Bayesian methods become less important. This is due to the fact that we consider an increasing number of observations and thus increasing evidence by including more and more studies. Correspondingly, the number of studies that remain unconsidered is declining. In the extreme, we include *all* prior studies and consequently have an uninformative (flat) prior beyond these studies. If so, the meta p-value approximates the post-study Bayesian error rate (Zyphur and Oswald 2015).

The complications in multiple regressions that restrict the feasibility of meta-analyses, as well as the eligibility of the meta-analytical approaches that are available to deal with these problems, are beyond this paper's scope. For further study of meta-analysis, the reader is referred to Becker and Wu (2007), Card (2012), Kline (2013: chapter 9), and Schmidt and Hunter (2014). For an introduction to Bayesian methods, the reader is referred to Hartung et al. (2008: chapter 12), Pitchforth and Mengersen (2013), and Zyphur and Oswald (2015).

2.4. Publication bias

The ill-directed incentives of the present publication system produce a bias towards statistical significance (Fanelli 2011). This bias is arguably more covert and higher in economics than other fields (Fanelli 2010), not least because replication is not a popular exercise among non-experimental economists (Evanschitzky and Armstrong 2010). While some researchers may honestly but erroneously believe that "starless" results are not interesting enough to warrant publication (Sterne et al. 2008), the major problem is the one highlighted by the "publish or perish" witticism: in our competitive research system, most researchers are under pressure to produce journal papers with novel findings. If papers with statistically significant findings ("positive" results) are more likely to be published, researchers are likely to adopt one or several selection strategies: *selective preparation* means not to conduct (replication) studies that are likely to be a "waste of time" because they do not promise to produce statistically significant novelties. *Selective submission* implies one does not submit papers

Table 5
Selected methods for detecting publication bias

Method	**Authors**
Comparison of published and unpublished studies	Song et al. (2000); Sterne and Egger (2005)
Caliper test	Berning and Weiß (2016); Gerber and Malhotra (2008)
p-curve analysis	Head et al. (2015); Simonsohn et al. (2014)
Funnel plots	Egger et al. (1997); Light and Pillemer (1984)
Capture-recapture method	Bennett et al. (2004); Poorolajal et al. (2010)
Fail-safe N	Rosenberg (2005); Rosenthal (1979)
Selection models	Kicinski et al. (2015); Silliman (1997)

with results that are not statistically significant because they are unlikely to be published. *Selective reporting* means p-hacking and the selective presentation of an analytical variant that "worked best" in terms of producing significance. Beyond the single researcher's sphere of influence, there is finally the problem of *selective publishing*. Even if researchers are conscientious and scrupulous enough to refrain from self-interested selection practices, reviewers and editors have ample discretion to promote papers for publication that contain seeming novelties. As result of all these selection processes, "significant" findings are overrepresented and studies with negative results tend to stay in researchers' "file drawers" (Rosenthal 1979) without ever being presented to the public. In other words, they are missing but not missing at random and thence cause a bias for significance. This leads to the definition by Kline (2013: 274) according to which publication bias implies that published studies have more "statistically significant" findings and larger effect sizes than unpublished studies (including the ones that have not been made in the first place).

Over the last decades, a variety of meta-analytical methods have been developed to gauge publication bias (see Table 5). For a description of these methods and their respective potential to identify selection procedures, the reader is referred to Cooper et al. (2009), Rothstein et al. (2005), Song et al. (2000), or Weiß and Wagner (2011). In this paper, we will have to limit ourselves to briefly describing a few selected approaches.

A first attempt to gauge publication bias is to compare gray literature (e.g., discussion and conference papers) with published studies (Song et al. 2000). Larger effect sizes (and smaller p-values) in published compared to non-published papers are an indication of publication bias caused by selective submission and/or selective editorial policies. Besides the problem that unpublished studies are less disseminated, neither selective preparation nor p-hacking can be identified through this comparison. This is why the assessment of publication bias is often based on "anomalies" in the structure of test statistics. A widely applicable method is the *caliper test* (Gerber and Malhotra 2008). Its idea is to compare the frequency of reported test scores within a small band above and below the usual significance thresholds. Gerber and Malhotra (2008: 6) claim that "there is no reason to expect that, in the narrow region just above and below the critical value, there will be substantially more cases above than below the critical value unless the.05 level was somehow affecting what is being published." Instead of

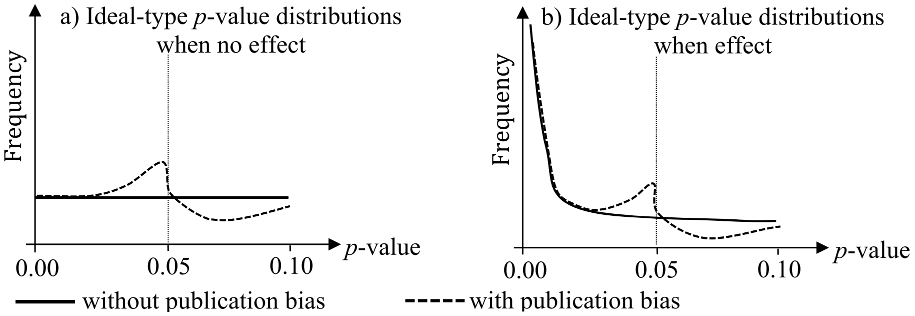


Fig 1. *Illustration of the basic idea of p-curve analysis*

test scores, *p-curve analysis* looks directly at the distribution of p-values (Head et al. 2015; Simonsohn et al. 2014). Figure 1 illustrates the idea: while we do not know whether there is an effect or not, we know that the p-value distribution is uniform if there is no effect. We also know that it has an exponential form with a right skew (Head et al. 2015) if there is an effect. In both cases, anomalies around the critical threshold should not occur and, if found, are taken as an indication of publication bias. Whereas an overrepresentation of p-values below 0.05 can be attributed to p-hacking, all selection procedures jointly contribute to the underrepresentation of values above the threshold.

In many research fields, publication bias itself has become an important object of study. Joober et al. (2012: 149), for example, report that in some medical areas almost no negative studies exist. They also find that publication bias has increased in many fields over the last years. The issue has also been taken up in the political and sociological sciences (cf., e.g., Auspurg and Hinz 2011; Gerber et al. 2010). In a recent study using the caliper test, Berning and Weiß (2016) find strong evidence for publication bias in papers published from 2001 to 2010 in three flagship journals of the German social sciences (Kölner Zeitschrift für Soziologie und Sozialpsychologie, Zeitschrift für Soziologie, and Politische Vierteljahresschrift). With a few early exceptions (e.g., Denton 1985; Lovell 1983), the awareness and study of publication bias has been less pronounced in economics in the past. But recently, a large-scale study by Brodeur et al. (2016) analyzed the distribution of about 50,000 test statistics that were published from 2005 to 2011 in three of the most prestigious economic journals (American Economic Review, Journal of Political Economy, Quarterly Journal of Economics). They find a considerable overrepresentation of marginally significant test statistics as well as a sizeable underrepresentation of marginally "nonsignificant" statistics. Their "interpretation is that researchers inflate the value of just-rejected tests by choosing 'significant' specifications" (Brodeur et al. 2016: 1).

Many suggestions have been made to mitigate publication bias (see Munafò et al. 2017 for an overview). Song et al. (2013) and Weiß and Wagner (2011) propose to strengthen alternative publication outlets. They also call for a general change of editorial policies towards giving equal publishing chances to all scientific results, including replications and negative findings. To reduce the risk of selective reporting and increase the chance of being published independent of whether positive or negative results are eventually found, Munafò et al. (2017) suggest to go beyond the after-study provision of raw data, and peer-review and register complete study designs *before* they are carried out. Across a variety of disciplines, various initiatives try to institutionalize efforts counteracting distorting selection procedures. With a view to the dire consequences of publication bias in medical research, a global initiative *All Trials Registered/All Results Reported* was launched in 2013. Along the same lines, the *Journal of Negative Results in BioMedicine*, the *PLOS ONE Journal*, and the *All Results Journals* explicitly encourage replication studies and pursue policies of publishing positive and negative results.

Institutionalized efforts to strengthen the practice of replication and pre-registration seem be weak in economics compared to other fields such as the medical sciences that spearheaded the development. In a study of all 333 economic Web-of-Science journals, Duvendack et al. (2015) find that most of them still give very low priority to replication. Pre-registration of studies also seems to lag behind other fields. No economic journal, for example, is among the approximately 40 journals that, according to the The *Center for Open Science*, have adopted a policy of peer reviewing and registering study designs before results are known (Duvendack et al. 2017). But things are changing. A topical initiative is the call of the *economics-ejournal* in which researchers are asked to select a published study as a candidate for replication and to discuss how they would carry out the replication. There are also some noteworthy replication platforms for economists such as *The Replication Network* and *Replication in Economics* that provide data bases of replications and the opportunity to publish replication studies. The issue has also attracted the attention of professional economic societies such as The American Economic Association that concerned itself with replication on its 2017 *annual meeting* and has launched a *pre-registration scheme for randomized controlled trials*. In this scheme, a study's design is peer-reviewed based on its methodological quality and registered if accepted. Peer-review of a study's design and formal registration by a prestigious institution are meant to prevent p-hacking and contribute to equal chances of being published independent of which results are eventually found. Reaching even further, the *Journal of Development Economics* recently started a pilot to test whether pre-registration in conjunction with "blind reviews" can improve the quality of empirical research in economics. The pilot provides researchers with the opportunity to have their prospective studies reviewed and approved for *publication before* the data are collected.

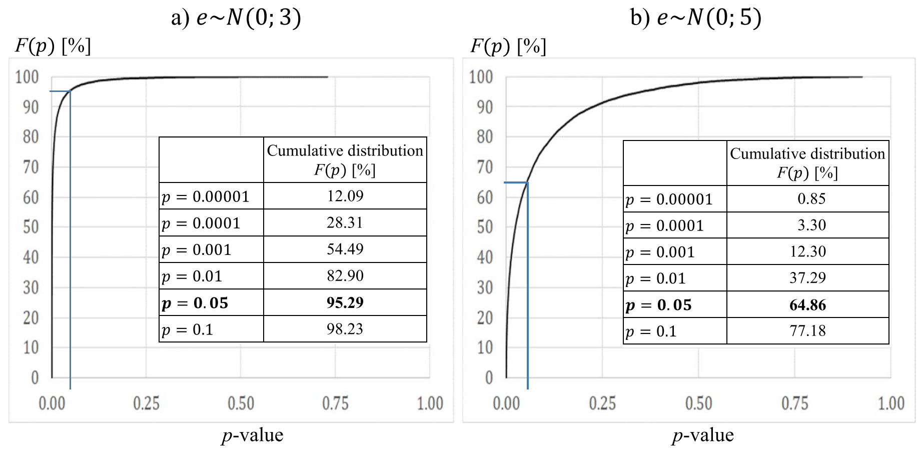
3. Disregard of p-value sample-to-sample variability

Practical empirical research often ignores the implications that result from the fact that a p-value is but a statistical estimate and inherently based on probability theory. Consequently, results with small p-values are often declared significant and then claimed to represent substantial evidence for the existence of a real phenomenon or even simply assumed to be real (Halsey et al. 2015). But even if researchers do not commit the inverse probability error of interpreting the p-value as the post-study probability of making a false existence claim or even erroneously consider a low p-value as strong evidence in favor of a specific alternative hypothesis, many believe that, after having found a low p-value, repetition of an experiment or repeated random sampling will produce a similar statistical verdict (Goodman 1992). With regard to the random sampling case, which we will use for our argument and numerical illustrations hereafter, Halsey et al. (2015: 180) note that "P-values are only as reliable as the sample from which they have been calculated. A small sample taken from a population [with big noise] is unlikely to reliably reflect the features of that population." They therefore provide a very unreliable signal of what is going to happen in replications unless they are very low (Cumming 2008).

While most researchers realize that small samples and big noise increase the *level* of the p-value that is to be expected, they are less likely to be fully aware of the fact that they also considerably increase the sample-to-sample *variability* of the p-value. The erroneous belief that the p-value indicates the likelihood that significant results can be replicated has been called "replication fallacy" (Gigerenzer et al. 2004). There is no way to assess the chance of replicating "significant" results unless one has at least an approximate estimate of power. The replication fallacy can be partly attributed to the fact that, contrary to other statistical estimates, no measure of the p-value's variability is usually reported even though it can be of considerable magnitude (Boos and Stefanski 2011). What is more, conventional notation abstains from advertising that the p-value is but an estimate (for example by using the notation $\hat{p}$). The disregard of the p-value's variability, which is particularly widespread in multiple regression analysis, is reflected in studies that do not discuss, let alone quantitatively assess, statistical power.

Ignoring or underestimating the variability of the p-value over replications goes hand in hand with underestimating the variability of effect size estimates. Together, this generates the risk of overrating the inferential value of results that are found in a single study. We use simulation to illustrate this risk with a focus on the regression-based confirmatory analysis of observations obtained from random sampling. The simulation provides an intuitive numerical illustration of how the random sampling error impacts on the variability of both p-values and coefficient estimates over replications (repeated random sampling). This illustration is aimed at counteracting the widespread misconception that an effect size estimate accompanied by a small p-value is by itself a reliable indication of the true effect size.

We carry out two simulations to illustrate why relying on the p-value without considering its sample-to-sample variability falls short of the mark even within an otherwise correct interpretation. In each simulation, we generate 10,000 random samples based on a presumed "reality" characterized by the linear relationship $y = 1 + \beta x + e$, with $\beta = 0.2$. The two simulations differ in their normally distributed error terms, which are $e{\sim}N(0;3)$ and $e{\sim}N(0;5)$, respectively. The sample size is $n = 50$, with x varying from 0.5 to 25 in equal steps of 0.5. For



	Cumulative distribution $F(p)$ [%]
p = 0.00001	12.09
p = 0.0001	28.31
p = 0.001	54.49
p = 0.01	82.90
p = 0.05	**95.29**
p = 0.1	98.23

	Cumulative distribution $F(p)$ [%]
p = 0.00001	0.85
p = 0.0001	3.30
p = 0.001	12.30
p = 0.01	37.29
p = 0.05	**64.86**
p = 0.1	77.18

Fig 2. *p-value distribution over 10,000 replications ($y = 1 + 0.2x + e$; sample size $n = 50$) (One-sided test; normal test statistic)*

each reality (i.e., for the $\sigma = 3$ and the $\sigma = 5$ case), we run OLS-regressions for each of the 10,000 samples. Since we use an estimator that perfectly fits the association in the data, we can interpret the heterogeneous results in these 10,000 samples as direct effect of the random sampling error.

Figure 2 visualizes the variability of the p-value over the 10,000 OLS-regressions in each of the two simulations. The left-hand side shows the cumulative distribution of the p-value for the normally distributed error term $e{\sim}N(0;3)$. The right-hand side shows it for the error term $e{\sim}N(0;5)$. Let's take an exemplary look at the cumulative distributions for a p-value of 0.01 and 0.10. In the $\sigma = 3$ case, $p \leq 0.01$ is obtained in 8,290 of the 10,000 samples (i.e., in 82.90% of all samples), and we find p-values above 0.10 in only 1.77% ($= 1 - 0.9823$) of all replications. In the $\sigma = 5$ case, the probability of obtaining $p \leq 0.01$ in a random sample drops to 37.29% while the probability of obtaining p-values above 0.10 rises to 22.82% ($= 1 - 0.7718$).

The cumulative distributions that are tabulated for selected p-values illustrate that the size of the error term has a considerable impact on the p-value's sample-to-sample variability. This is why many statisticians call for complementing the p-value approach with statistical power considerations. Statistical power is defined as the conditional probability of a significant result over many replications if an effect is true. Having generated the data, we know the true effect $\beta = 0.2$. We thence also know the true power to be the cumulative distribution F of the p-value at the 0.05 level. For $e{\sim}N(0;3)$, we find a true power of 95.29% (i.e., 9,529 out of 10,000 samples yield $p \leq 0.05$). In contrast, the true power is only 64.86% in the case of $e{\sim}N(0;5)$.

Comparing the two p-value distributions in Figure 2 helps show that, besides a single study's p-value, its variability – and in dichotomous significance testing

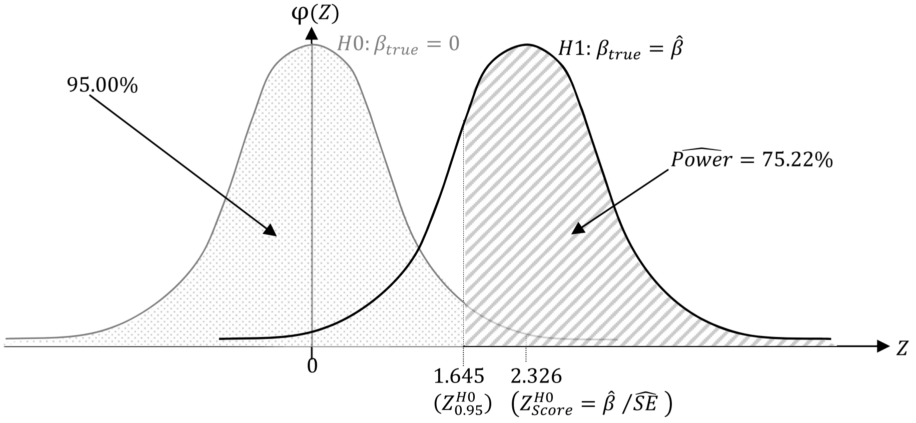


FIG 3. *Estimated power for* $p = 0.01$ *based on the assumption* $\hat{\beta} = \beta_{true}$ *and* $\widehat{SE} = SE_{true}$

the statistical power[15] – determines the informational value of a finding. High-powered studies are more reliable in that they indicate not only lower average p-values but also a lower variability of p-values over random replications. In many cases, power is below the approximately 65% found in our $e{\sim}N(0;5)$ simulation. We should therefore cautiously interpret a p-value and realize that p-values found in one study, albeit very low, should be put into perspective by providing a measure of their variability.

Unfortunately, researchers usually face but one random realization and ignore the true effect, the p-value's variability, and the true power (Halsey et al. 2015). In each of the 10,000 random realizations, a researcher would estimate a different coefficient $\hat{\beta}$, a different standard error $\widehat{SE}$, and a different p-value. If we naïvely assumed that the coefficient $\hat{\beta}$ and the standard error $\widehat{SE}$ that we happened to estimate from one data set were true ($\beta_{true} = \hat{\beta}$; $SE_{true} = \widehat{SE}$), we could estimate the power. Figure 3 shows that under this assumption the power estimate can be directly derived from a given p-value. We illustrate the case for $p = 0.01$, which corresponds to a test score $Z_{Score}^{H0} = 2.326$. Assuming a one-sided significance test at the p^* =0.05 level, we know the critical test value $Z^* = Z_{0.95}^{H0} = 1.645$. Presuming $\hat{\beta} = \beta_{true}$ as alternative hypothesis $H1$, we can estimate $\widehat{Power} = 1 - \Phi^{H1}\,(1.645) = 75.22\%$.

Table 6 further illustrates the effect of random sampling error. It displays p-values (and their cumulative distribution F), coefficient estimates $\hat{\beta}$ and standard error estimates $\widehat{SE}$ (and their resulting scores under the null Z_{Score}^{H0}), and power estimates $\widehat{Power}$ for six selected samples out of the 10,000 replications

[15] Power is a zeroth order (lower) partial moment of the p-value distribution over replications. As such, it contains only a part of the distributional information. This part, however, is sufficient if one confines oneself to the dichotomous "statistically significant" vs. "not statistically significant" distinction; i.e., "statistical power quantifies the repeatability of the P value, but only in terms of the either-or interpretation" (Halsey et al. 2015: 180).

TABLE 6

p-values and coefficient and power estimates for six out of 10,000 simulated samples (with size $n = 50$ each)

	$e{\sim}N(0;3)$			$e{\sim}N(0;5)$			Z^{H0}_{Score}	$\widehat{Power}$
	$F(p)$ [%]	$\hat{\beta}^{\sigma=3}$ [a]	$\widehat{SE}^{\sigma=3}$	$F(p)$ [%]	$\hat{\beta}^{\sigma=5}$	$\widehat{SE}^{\sigma=5}$		[%]
$p = 0.00001$	12.09	0.224	0.053	0.85	0.491	0.115	4.265	99.56
$p = 0.0001$	28.31	0.248	0.067	3.30	0.317	0.085	3.719	98.10
$p = 0.001$	54.49	0.174	0.056	12.30	0.304	0.098	3.090	92.58
$p = 0.01$	82.90	0.162	0.070	37.29	0.249	0.107	2.326	75.22
$p = 0.05$	95.29	0.112	0.068	64.86	0.170	0.103	1.645	50.00
$p = 0.1$	98.23	0.083	0.065	77.18	0.127	0.099	1.282	35.82

[a] In general, lower *p*-values go hand in hand with larger coefficient estimates. The first three *p*-value rows show, however, that there are simulation runs in which the coefficient estimates do not follow the inverse order of the *p*-value. This is because both the mean and the variance vary from sample to sample.

(with size $n = 50$ each) that we simulated for the $e{\sim}N(0;3)$ and the $e{\sim}N(0;5)$ "realities," respectively. In each simulation, we first ordered the 10,000 random samples from low to high according to the obtained *p*-values. We then identified the samples that yielded a *p*-value closest to the ones displayed in the first column. For example, in the $\sigma = 3$ case, the 1,209th sample yielded $p = 0.00001$ (the cumulative distribution of this *p*-value is therefore 12.09%), a coefficient estimate $\hat{\beta}^{\sigma=3} = 0.224$, and a standard error $\widehat{SE}^{\sigma=3} = 0.053$. In the $\sigma = 5$ case, only 85 out of the 10,000 samples produced a *p*-value lesser than or equal 0.00001. The 85th sample yielded $p = 0.00001$, a coefficient estimate $\hat{\beta}^{\sigma=5} = 0.491$, and a standard error $\widehat{SE}^{\sigma=3} = 0.115$. The identical *p*-value in both cases resulted from the identical $Z^{H0}_{Score} = 4.265 = \hat{\beta}^{\sigma=3}/\widehat{SE}^{\sigma=3} = \hat{\beta}^{\sigma=5}/\widehat{SE}^{\sigma=5}$. Using a one-sided significance test at the $p^* = 0.05$ level (corresponding to a critical test value $Z^* = Z^{H0}_{0.95} = 1.645$) and using alternative hypotheses $H1^{\sigma=3}$ and $H1^{\sigma=5}$ based on the assumption that the respective coefficient and standard error estimates are true, we obtain an estimated $\widehat{Power} = 99.56\% = 1 - \Phi^{H1^{\sigma=3}}(1.645) = 1 - \Phi^{H1^{\sigma=5}}(1.645)$ in both cases.

While low *p*-values are often associated with a high reliability of estimated effects, Table 6 reflects an important fact: an unbiased estimator (in our case the OLS-estimator) estimates correctly *on average*. While we would find in both realities an *average* coefficient across all 10,000 simulation runs that is *very* close to 0.2, we overestimate the effect size in the case of highly significant results. In our simulation, for example, $p = 0.00001$ was accompanied by a coefficient estimate of 0.491 in the $\sigma = 5$ case. In other words, under reasonable sample sizes and population effect sizes, it is the *abnormally* large sample effect sizes that produce "highly significant" *p*-values (Trafimow et al. 2017). If, out of the 10,000 simulated samples, we only considered the 6,486 samples that yielded significant results at the 0.05 level (i.e., satisfied the $Z^{H0}_{Score} \geq 1.645$ condition), we would find an average coefficient estimate of 0.257. This is not surprising: by averaging over "significant" results only, we right-truncate the distribution of the p-value which, in turn, involves a left-truncation of the distribution of the coefficient.

The p-value's variability over replications undermines its already weak informative value. Even if there were *no* uncorrected multiple testing, *no* misinterpretation of the p-value, and *no* distorting selection procedures, researchers might still be overconfident and believe that at least a very low p-value of, let's say, 0.005 or 0.001 provides a trustworthy indication of the true effect. Unfortunately, we cannot deduce even that much by just looking at the p-value. The consequences for conventional significance testing are quite sobering: if we are oblivious to the p-value's unknown sample-to-sample variability, we will grossly overestimate its limited informational content in a single study; and if we account for the p-value's sample-to-sample variability, we must concede that its suitability to indicate the strength of evidence is *very* limited. Let us briefly summarize the main reasons behind this sobering insights:

1. The variability of the p-value over random replications may be high or low. Being reduced to having to analyze data from one random realization, we do not know the degree of variation (see footnote 2). We would need some prior assumption regarding the true effect size.
2. In plausible constellations of noise and sample size, we can very easily find a significant result in one random sample and not find a significant result in another.
3. The variability of the p-value is paralleled by the variability of the estimated coefficient. We may thence find a large coefficient in one random sample and a small one in another.
4. Unbiased estimators estimate correctly *on average*, but we have no way of identifying the p-value below which (above which) we overestimate (underestimate) the effect size. In our example, a p-value of 0.001 was associated with a coefficient estimate of 0.174 in the $\sigma = 3$ case (= underestimation of the effect size). In the $\sigma = 5$ case, it was linked to a coefficient estimate of 0.304 (= overestimation). That is, even in the case of a highly significant result, we cannot make a direct inference regarding the effect.
5. If we rashly claimed a just-estimated coefficient to be true, we would not have to be worried if it cannot be replicated. For example, if an effect size and standard error estimate associated with a p-value of 0.05 were real, we would *necessarily* have a mere 50% probability of finding a statistically significant effect in replications (one-sided test). Things improve with lower p-values. But even at the 0.01 level, we have only a 75% probability of re-finding significance (see Table 6).

Our simulations provided an intuitive numerical illustration of the vagaries of sampling and the often neglected fact that it would be an inferential error to interpret the p-value as the probability of a hypothesis. Instead, it is but a measure that indicates how (in)compatible the particular data at hand (sample) is with a specified statistical model including the null hypothesis (Wasserstein and Lazar 2016). While the p-value is only a statement about a data set *conditional* on the null hypothesis of no effect, small p-values nonetheless give a hint that there might rather be an effect than none, in the sense that small p-values will occur more often if there is an effect compared to no effect (cf., the ideal-type p-value distributions without publication bias in Figure 1). However, due to the p-value's data dependence and the imponderables of random sampling (particularly in the case of small samples and big noise), large p-values may occur in a considerable fraction of random sampling replications even if there is an effect, and vice versa (cf., Figure 2). Hence, we must be cautious when interpreting p-values and realize that the p-value does not provide a clear rationale or even formal calculus for statistical inference in terms of post-study probabilities of scientific propositions (Goodman 2008). When trying to assess the evidence from a particular data set regarding a hypothesis, statistical power calculations (or more generally, the consideration of the p-value's sample-to-sample variability) would be a helpful complement to the conventional p-value approach. However, since the size of the true effect is unknown, we are limited to power calculations for varying but plausible effect sizes. Assuming such plausible effect sizes, in turn, requires some degree of prior knowledge.

Furthermore, mixing up the p-value concept ("null hypothesis significance testing") and the estimation of effect size in the same step is problematic since many of the best estimation procedures are based on the concept of unbiasedness. Claiming the effect size to be real that we happened to estimate out of a sample where the effect showed up as "significant" bears the risk of overestimating the effect. Unfortunately, with decreasing p-values this risk seems to increase. When samples are large enough, one might thence think about splitting the data at hand and using one part for the testing part of the analysis and the other one for the estimation part. This might in further perspective lead to cross-validation-like approaches or to resampling procedures. But at this point we can only postpone such thoughts to further research.

4. Conclusion and outlook

Spotting and avoiding misinterpretations and misuses of statistical significance tests is important because, as empirical social scientists and economists, we are interested in learning from observations and statistical inference; i.e., we want to draw inductive conclusions and make general propositions about regularities in social and economic life given the evidence from the data. We are also interested in assessing the trustworthiness of these inductive conclusions by assigning post-study probabilities to our propositions. Unfortunately, even though the label "hypothesis testing" is commonly attached to statistical significance testing, p-values cannot be used to test the trustworthiness of hypotheses in terms of assigning post-study probabilities to hypotheses. Furthermore, p-values and associated effect size estimates may exhibit a wide variability over replications. However, we rarely start from scratch. Besides the evidence in our own data, we would need to summarize the knowledge from prior studies and use Bayesian statistics to assess the trustworthiness of scientific propositions in terms of probabilities.

We must realize that statistical inference is not as trivial as the dichotomous "significance" vs. "nonsignificance" declarations suggest at first glance. The corresponding either-or interpretations regarding the trustworthiness and importance of estimated effects are neat and seemingly plausible but wrong. What is more, misuses such as disregarding multiplicities and p-hacking as well as selective publishing may inflate statistical significance claims and distort the published body of evidence. Many reform measures have been suggested in the literature to mitigate these problems. The most obvious one is the call for an increase of statistical literacy through better teaching. Others include changes of research standards and incentives that would reduce the publish-statistically-significant-results-or-perish pressure that dominate many disciplinary cultures. The most notable examples of reform on the institutional level (journals, scientific associations, etc.) are policies of pre-registration, sharing (of data and analytical protocols), replication, and unbiased publishing of both positive and negative results.

While some of these policies are being slowly introduced into economics, there are a number of specific concerns that need to be addressed before approaches from other fields can be successfully transplanted to (non-experimental) economic research. These concerns stem from the prevalence of regression-based observational study and the data-driven specification search that characterize many econometric analyses. A primary question is whether a given research context provides a probabilistic justification for using the p-value at all. Because the nature of the inference varies from one research context to the other, researchers should explicitly state whether their inferential argument is based on randomization in experiments, random sampling, or the assumption of having a random realization of an unseen parent population. Especially in the analysis of non-experimental data, another important question is whether and where we can draw a dividing line between confirmatory study (hypothesis testing) and data-driven exploratory modeling (hypothesis generation). Even though the labels "theory-based" and hypothesis testing" are frequently attached to econometric models, they are often data-based; i.e., some variable selection, data transformation, and other model specification search procedures are carried out *after* seeing the data. Considering this appropriate for confirmatory research gives rise to two more questions: first, where is the dividing line between adequate model specification and p-hacking? Second, if researchers declare a study confirmatory but nonetheless engage to some degree in specification search, (how) would they need to adjust for multiple testing?

Even widely applauded measures aimed at improving research practices need to be scrutinized regarding their viability and effectiveness in econometric research. It is not clear, for example, how measures such as the pre-registration of analytical designs, the replication of studies, and the correction for multiple testing would have to look like within a scientific culture where it is common practice to specify statistical models that fit the data. Related to that, the question arises of how to meet the requirement of considering the body of evidence instead of focusing on single studies. How can we carry out a meta-analysis regarding a specific scientific issue when comparability across studies is impeded because there are often (nearly) as many data-dependent model specifications as there are studies? Another question is how, in view of the p-value's inferential limitations, we can provide an interpretative orientation vis-à-vis the often large numbers of regression coefficient estimates. This is an especially urgent issue since data-based economic models are often heavily populated, besides the original variables of interest, by interaction terms, (log)transformed variables, lagged variables, instrumental variables, higher-order polynomials, and control variables. Can Bayesian approaches provide a practically feasibly solution in such a context? And if so, how can we specify Bayesian priors given the frequent lack of comparability among studies and the often large variable sets?

Acknowledgments

We would like to thank the Editor Professor David Banks and two anonymous reviewers for their comments, ideas, and criticism. In addition, we are very grateful for the financial support provided by the German Research Foundation (DFG).

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. PeerJ, doi: 10.7717/peerj.3544.

Armstrong, J.S. (2007): Significance tests harm progress in forecasting. International Journal of Forecasting 23(2): 321–327.

Auspurg, K., Hinz, T. (2011): What Fuels Publication Bias? Theoretical and Empirical Analyses of Risk Factors Using the Caliper Test. Journal of Economics and Statistics 231(5-6): 636–660.

Baker, M. (2016): Statisticians issue warning on P values. Nature 531(7593): 151.

Becker, B.J., Wu, M-J. (2007): The Synthesis of Regression Slopes in Meta-Analysis. Statistical Science 22(3): 414–429. MR2416817

Benjamini, Y., Hochberg, Y. (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57(1): 289–300. MR1325392

Bennett, D.A., Latham, N.K., Stretton, C., Anderson, C.S. (2004): Capture-recapture is a potentially useful method for assessing publication bias. Journal of Clinical Epidemiology 57(4): 349–357.

Berning, C., Weiß, B. (2016): Publication Bias in the German Social Sciences: An Application of the Caliper Test to Three Top-Tier German Social Science Journals. Quality & Quantity 50(2): 901–917.

Berry, D.A. (2016): P-Values Are Not What They're Cracked Up to Be. Online Discussion: ASA Statement on Statistical Significance and P-values. The American Statistician 70(2): 1–2.

Berry, D. (2017): A p-Value to Die For. Journal of the American Statistical Association 112(519): 895–897. MR3735344

Boos, D.D., Stefanski, L.A. (2011): P-Value Precision and Reproducibility. The American Statistician 65(4): 213–221. MR2867504

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R. (2009): Introduction to Meta-Analysis. Chichester: John Wiley & Sons.

Bretz, F., Hothorn, T., Westfall, P. (2010): Multiple comparisons using R. Boca Raton: CRC Press.

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

Bruns, S.B. (2017): Meta-Regression Models and Observational Research. Oxford Bulletin of Economics and Statistics 0305–9049, doi: 10.1111/obes.12172.

Card, D., Krueger, A. B. (1995): Time-series minimum-wage studies: A meta-analysis. American Economic Review (AEA Papers and Proceedings) 85: 238–243.

Card, N. A. (2012): Applied meta-analysis for social science research. New York: Guilford Press.

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

Cooper, D.J., Dutcher, E.G. (2011): The dynamics of responder behavior in ultimatum games: a meta-study. Experimental Economics 14(4): 519–546.

Cooper, H., Hedges, L., Valentine. J. (eds.) (2009): The handbook or research synthesis and meta-analysis. 2$^{\text{nd}}$ ed., Russell Sage Foundation, New York.

Crouch, G.I. (1995): A meta-analysis of tourism demand. Annals of Tourism Research 22(1): 103–118.

Cumming, G. (2008): Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science 3(4): 286–300.

Denton, F.T. (1985): Data Mining as an Industry. Review of Economics and Statistics 67(1): 124–127.

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.

Didelez, V., Pigeot, I., Walter, P. (2006): Modifications of the Bonferroni-Holm procedure for a multi-way ANOVA. Statistical Papers 47: 181–209. MR2236050

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: Papers & Proceedings 2017: 107(5): 46–51.

Egger, M., Smith, G.D., Schneider, M., Minder, C. (1997): Bias in meta-analysis detected by a simple, graphical test. British Medical Journal 315 (7109): 629–634.

Engel, C. (2011): Dictator games: a meta study. Experimental Economics 14(4): 583–610.

Evanschitzky, H., Armstrong, J.S. (2010): Replications of forecasting research. International Journal of Forecasting 26: 4–8.

Fanelli, D. (2010): Positive" results increase down the hierarchy of the sciences. PLoS One 5(4): e10068.

Fanelli, D. (2011): Negative results are disappearing from most disciplines and countries. Scientometrics 90(3): 891–904.

Fisher, R.A. (1925): Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd.

Fisher, R.A. (1935): The design of experiments. Edinburgh: Oliver & Boyd.

Fitzpatrick, L., Parmeter, C.F., Agar, J. (2017): Threshold Effects in Meta-Analyses With Application to Benefit Transfer for Coral Reef Valuation. Ecological Economics 133: 74–85.

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. MR3735346

Gerber, A. S., N. Malhotra (2008): Publication Bias in Empirical Sociological Research. Do Arbitrary Significance Levels Distort Published Results? Sociological Methods & Research 37(1): 3–30. MR2522170

Gerber, A.S., Malhotra, N., Dowling, C.M., Doherty, D. (2010): Publication Bias in Two Political Behavior Literatures. American Politics Research 38(4): 591–613.

Gigerenzer, G., Krauss, S., Vitouch, O. (2004): The null ritual: what you always wanted to know about significance testing but were afraid to ask. In: Kaplan, D. (ed.): The SAGE handbook of quantitative methodology for the social sciences (Chapter 21). Thousand Oaks: Sage.

Gigerenzer, G., Marewski, J.N. (2015): Surrogate Science: The Idol of a Universal Method for Statistical Inference. Bayesian Probability and Statistics in Management Research, Special Issue of the 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. (1992): A Comment of Replication, P-Values and Evidence. Statistics in Medicine 11: 875–879.

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.

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

Haller, H., Krauss, S. (2002): Misinterpretations of Significance: A Problem Students Share with Their Teachers? Methods of Psychological Research Online 7(1): 1–20.

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.

Hartung, J., Knapp, G., Sinha, B.K. (2008): Statistical Meta-Analysis with Applications. Hoboken: John Wiley & Sons. MR2435836

Head, M.L, Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D. (2015): The Extent and Consequences of P-Hacking in Science. PLoS Biology 13(3): e1002106, doi: 10.1371/journal.pbio.1002106.

Hirschauer, N., Mußhoff, O., Grüner, S., Frey, U., Theesfeld, I., Wagner, P. (2016): Inferential misconceptions and replication crisis. Journal of Epidemiology, Biostatistics, and Public Health 13(4): e12066-1–e12066-16.

Hochberg, Y., Tamhane, A.C. (1987). Multiple comparison procedures. New York: Wiley. MR0914493

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2): 65–70. MR0538597

Howard, G.S., Maxwell, S.E., Fleming, K.J. (2000): The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. Psychological Methods 5: 315–332.

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. MR2216666

Joober, R., Schmitz, N., Dipstat, L.A., Boksa, P. (2012): Publication bias: What are the challenges and can they be overcome? Journal of Psychiatry & Neuroscience 37(3): 149–152.

Kerr, N.L. (1998): HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review 2(3): 196–217.

Kicinski, M, Springate, D.A., Kontopantelis, E. (2015): Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. Statistics in Medicine 34: 2781–2793. MR3375981

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

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.

Lange, T. (2016): Discrimination in the laboratory: A meta-analysis of economics experiments. European Economic Review 90: 375–402.

Leamer, E.E. (1978): Specification Searches: Ad Hoc Inference with Nonexperimental Data. New York: Wiley. MR0471118

Lecoutre, B., Poitevineau, J. (2014): The Significance Test Controversy Revisited. The Fiducial Bayesian Alternative. Heidelberg: Springer. MR3308437

Light, R.J., Pillemer, D.B. (1984): Summing Up: The Science of Reviewing Research. Cambridge: Harvard University Press.

List, J.A., Shaikh, A.M., Xu, Y. (2016): Multiple Hypothesis Testing in Experimental Economics. No. w21875. National Bureau of Economic Research, Working Paper No. 21875.

Loomis, J.B., White, D.S. (1996): Economic benefits of rare and endangered species: summary and meta-analysis. Ecological Economics 18(3): 197–206.

Lovell, M.C. (1983): Data Mining. Review of Economics and Statistics 65(1): 1–12. MR0763105

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

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.

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., du Sert, N.P., Simonsohn, U., Wagenmakers, E-J., Ware, J.J., Ioannidis, J.P.A. (2017): A manifesto for reproducible science. Nature Human Behaviour 1(0021): 1–8.

Nickerson, R.S. (2000): Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods 5(2): 241–301.

Nosek, B.A., Ebersole, C.R., DeHaven, A.C., Mellor, D.T. (2018): The preregistration revolution. Proceedings of the National Academy of Sciences of the United States of America 115(11): 2600–2606.

Nuzzo, R. (2014): Statistical Errors. *P*-values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506(7487): 150–152.

Oakes, M. (1986): Statistical inference: A commentary for the social and behavioural sciences. New York: Wiley.

Pigeot, I. (2000): Basic concepts of multiple tests – A survey. Invited paper. Statistical Papers 41: 3–36. MR1746085

Pitchforth, J.O., Mengersen, K.L. (2013): Bayesian Meta-Analysis. In: Alston, C.L., Mengersen, K.L., Pettitt, A.N. (eds.): Case Studies in Bayesian Statistical Modelling and Analysis. Chichester: John Wiley & Sons, Ltd.: 118–140.

Poorolajal, J., Haghdoost, A.A., Mahmoodi, M., Majdzadeh, R., Nasseri-Moghaddam, S., Fotouhi, A. (2010): Capture-recapture method for assessing publication bias. Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences 15(2): 107–115.

Roberts, C.J. (2005): Issues in meta-regression analysis: An overview. Journal of Economic Surveys 19(3): 295–298.

Romano, J.P., Shaikh, A.M., Wolf, M. (2010): Multiple Testing. In: Palgrave Macmillan (eds.) The New Palgrave Dictionary of Economics. London: Palgrave Macmillan, doi: 10.1057/978-1-349-95121-5_2914-1.

Rosenberg, M.S. (2005): The File-drawer Problem Revisited: A General Weighted Method for Calculating Fail-Safe Numbers in Meta-Analysis. Evolution 59(2): 464–468, Rosenthal, R. (1979): The file drawer problem and tolerance for null results. Psychological Bulletin 86(3): 638–641.

Rothstein, H., Sutton, A.J., Borenstein, M. (2005): Publication Bias in Meta-Analysis. Prevention, Assessment and Adjustments. Sussex: Wiley. MR2238828

Schmidt, F.L., Hunter, J.E. (2014): Methods of meta-analysis: Correcting error and bias in research findings. Los Angeles: Sage publications.

Silliman, N. (1997): Hierarchical selection models with applications in meta-analysis. Journal of American Statistical Association 92(439): 926–936. MR1482123

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.

Simonsohn, U., Nelson, L.D., Simmons, J.P. (2014): *P*-Curve: A Key to the File-Drawer. Journal of Experimental Psychology 143(2): 534–547.

Smith, M.L. (1980): Publication bias and meta-analysis. Evaluation in Education 4: 22–24.

Song, F., Eastwood, A.J., Gilbody, S., Duley, L., Sutton, A.J. (2000): Publication and related biases. Southampton: The National Coordinating Centre for Health Technology Assessment.

Song, F., Hooper, L., Loke, Y.K. (2013): Publication bias: what is it? How do we measure it? How do we avoid it? Open Access Journal of Clinical Trials 5: 71–81.

Stanley, T.D., Jarrell, S. B. (1989): Meta-regression analysis: A quantitative method of literature surveys. Journal of Economic Surveys 3(2): 161–170.

Stanley, T.D., Doucouliagos, H. (2012): Meta-Regression Analysis in Economics and Business. London: Routledge.

Sterling, T.D. (1959): Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance – Or Vice Versa. Journal of the American Statistical Association 54(285): 30–34.

Sterne, J.A.C., Egger, M. (2005): Regression Methods to Detect Publication and Other Bias in Meta-Analysis. In: Rothstein, H.R., Sutton, A.J., Borenstein, M. (eds.): Publication Bias in Meta-Analysis. Prevention, Assessment and Adjustments. Chichester: Wiley: 99–110. MR2238828

Sterne, J.A.C., Egger, M., Moher, D. (2008): Addressing reporting biases. In: Higgins, J.P.T., Green, S. (eds.): Cochrane handbook for systematic reviews of interventions: 297–333. Chichester: Wiley.

Trafimow, D. et al. (2017): Manipulating the alpha level cannot cure significance testing. Frontiers in Psychology 9: 699, doi: 10.3389/fpsyg.2018.00699.

Van Houtven, G.L., Pattanayak, S.K., Usmani, F., Yang, J.C. (2017): What are Households Willing to Pay for Improved Water Access? Results from a Meta-Analysis. Ecological Economics 136: 126–135.

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. MR3511040

Weiß, B., Wagner, M. (2011): The identification and prevention of publication bias in the social sciences and economics. Jahrbücher für Nationalökonomie und Statistik 231(5-6): 661–684.

Westfall, P., Tobias, R., Wolfinger, R. (2011): Multiple comparisons and multiple testing using SAS. Cary: SAS Institute.

Zelmer, J. (2003): Linear public goods experiments: A meta-analysis. Experimental Economics 6(3): 299–310.

Ziliak, S.T., McCloskey, D.N. (2008): The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives. Ann Arbor: The University of Michigan Press. MR2730043

Zyphur, M.J., Oswald, F.L. (2015): Bayesian Estimation and Inference: A User's Guide. Bayesian Probability and Statistics in Management Research, Special Issue of the Journal of Management 41(2): 390–420.