Heckelei, Thomas; Hüttel, Silke; Odening, Martin; Rommel, Jens

Article

The p-Value Debate and Statistical (Mal)practice - Implications for the Agricultural and Food Economics Community

German Journal of Agricultural Economics (GJAE)

Provided in Cooperation with:
Gesellschaft für Wirtschafts- und Sozialwissenschaften des Landbaues e.V. (GEWISOLA)

Suggested Citation: Heckelei, Thomas; Hüttel, Silke; Odening, Martin; Rommel, Jens (2023) : The p-Value Debate and Statistical (Mal)practice - Implications for the Agricultural and Food Economics Community, German Journal of Agricultural Economics (GJAE), ISSN 2191-4028, Deutscher Fachverlag, Frankfurt a. M., Vol. 72, Iss. 1, pp. 47-67, https://doi.org/10.30430/gjae.2023.0231

The p-Value Debate and Statistical (Mal)practice – Implications for the Agricultural and Food Economics Community

Thomas Heckelei
University of Bonn

Silke Hüttel
University of Göttingen

Martin Odening
Humboldt-Universität zu Berlin

Jens Rommel
Swedish University of Agricultural Sciences, Uppsala, Sweden

Abstract

A vivid debate is ongoing in the scientific community about statistical malpractice and the related publication bias. No general consensus exists on the consequences and this is reflected in heterogeneous rules defined by scientific journals on the use and reporting of statistical inference. This paper aims at providing an overview on the debate, discussing how it is perceived by the agricultural economics community, and deriving implications for our roles as researchers, contributors to the scientific publication process, and teachers. Following a 'Mixed Methods Review', we start by summarizing the current state of the p-value debate in the context of the replication crisis and commonly applied statistical practices in our community. This is followed by motivation, design, results and discussion of an explorative and descriptive survey on statistical knowledge and practice among the researchers in the agricultural economics community in Austria, Germany and Switzerland. Instead of providing specific guidelines or rules, we derive implications for our roles in the scientific process to support a needed long-term cultural change regarding empirical scientific practices. Acceptance of scientific work should largely be based on the theoretical and methodological rigor and where the perceived relevance arises from the questions asked, the methodology employed, and the data used but not from the results generated. Revised and clear journal guidelines, the creation of resources for teaching and research, and public recognition of good practice are suggested measures to move forward.

Keywords

statistical inference; p-hacking; publication bias; replication crisis; pre-registration

1 Introduction

Replicability of research results is at the core of scientific credibility. The discussion of a "replication crisis" in science has intensified over the last years (SCHOOLER, 2014; LOKEN and GELMAN, 2017) and also reached the community of environmental and resource economics (FERRARO and SHUKLA, 2020, 2022) and agricultural economics (FERRARO and SHUKLA, 2022). Practices like selective reporting of results, incentives to find "significant" effects in statistical analysis and the underrepresentation of null results (MERVIS, 2014) are discussed as core issues in the debate.

A more specific but strongly related issue is the use and interpretation of p-values and "p-hacking" in the context of statistical hypothesis tests. "Mindless statistics" (GIGERENZER, 2004) and "The cult of statistical significance" (ZILIAK and MCCLOSKEY, 2008) are terms to describe the widespread misuse and misinterpretation of p-values and statistical significance in reporting results of statistical and econometric analyses. The American Statistical Association has published a statement (WASSERSTEIN and Lazar, 2016), and several researchers signed a call to "retire statistical significance" (AMRHEIN et al., 2019). However, this is countered by others who acknowledge existing problems but nevertheless defend p-values, basically saying that nothing is wrong with p-values if they are used correctly (IMBENS, 2021). Currently, no consensus across the scientific community exists on the consequences of publication bias and malpractices, and this is reflected in heterogeneous rules defined by scientific journals on the use and reporting of statistical inference.

The agricultural economics community in Germany joined the debate by the fundamental work of

DOI: 10.30430/gjae.2023.0231

HIRSCHAUER et al. (2019) who suggest changes of rules for using p-values and statistical inference. After the first discussion in an organized session at the annual meeting of the German agricultural economics association (GEWISOLA) in 2019, the association created a working group with the task to assess how "p-hacking" and the misuse of statistical hypothesis tests in our scientific publications can be best avoided. In addition to the discussion of specific rules and best practices, the incentives leading to p-hacking and misinterpretations in the publication process were of interest. Ultimately, the working group targeted giving recommendations to the members of the association on how we can improve upon the current practice by changing relevant aspects of teaching, research and the scientific publication process.

This paper presents results of the working group and discusses implications of the debate on p-values and statistical inference for our roles as researchers, contributors to the scientific publication process, and teachers. The discussion aims at raising awareness and changing practices, thereby supporting the cultural change needed to improve upon the quality of statistical reporting in the scientific output of the community in the long run. This paper can be categorized as a 'Mixed Methods Review' (GRANT and BOOTH, 2009) that combines a literature review with a survey and stakeholder consultations. Specifically, we first offer some background knowledge on the current state of the p-value debate and statistical practices more generally in the literature. This is followed by motivation, design, results, and discussion of an explorative and descriptive survey on statistical knowledge and practice among the researchers in the agricultural economics community in Austria, Germany and Switzerland. Based on this background and additional input from external experts and participants of two GEWISOLA events in 2020 and 2021, implications for the community are developed.

2 The P-Value Debate and Related Statistical Practice

The "p-value debate" has many facets. We argue that it is useful to distinguish two main problem areas: first, misinterpretations and wrong conclusions from statistical inference, particularly significance tests and p-values, second, intentional or unintentional malpractices when applying statistical test procedures. This distinction is useful, as we believe each calls for distinct responses from the community.

2.1 Misunderstanding and Common Flaws when Applying p-Values and Statistical Hypothesis Testing

2.1.1 Wrong Interpretations of p-Values and Significance Tests

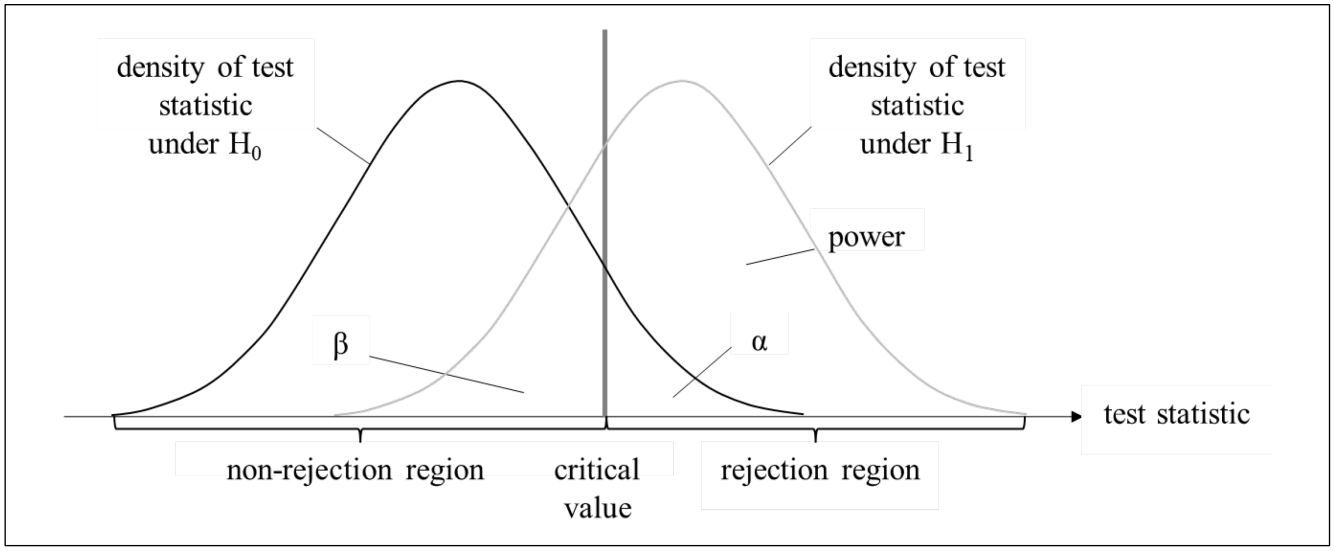
Before we turn to common misinterpretations of p-values and statistical hypothesis testing, we briefly reiterate their meaning. The purpose of a statistical test is to infer how compatible observed data D are with a null hypothesis H_0, which is specified in the framework of a statistical model, e.g. a regression model. The null hypothesis can be a statement about the size of a model parameter, e.g. the assumption that an unknown regression coefficient belonging to an economic variable has the value zero.[1] A statistical test requires (i) the derivation of a test statistic T, e.g. a z-score, a t-value or an F-value, for which the probability distribution is known, when the null hypothesis is true and some other distribution when the null hypothesis is false, and given that the set of model assumptions A are true, e.g. independence of the model's error terms; and (ii) a rejection rule, such if the value of the test statistic is an extreme one that would rarely be encountered by chance under the null hypothesis, then the test provides evidence against the null hypothesis.

In this setting, FISHER (1925) defines the p-value as the conditional probability of the test distribution that refers to the observed value of the test statistic, t, i.e. $Prob\,(T < t \mid H_0, A)$ for a one-sided test. Since it is often desired to arrive at a decision about the presence of an economic effect, the observed p-value is compared with a predetermined cut off-rate α, the "significance level", usually 0.05. If the observed p-value is smaller than the significance level, the null hypothesis is rejected, otherwise not.[2] The significance level reflects the type-I-error, i.e. $Prob\,(reject\ H_0 \mid H_0\ is\ true)$. The p-values are also called marginal significance levels as it relates to the respective test statistic's greatest level for which the

[1] It is important to note that H_0 need not to be a "nil hypothesis", as it is often the case in economic applications. In fact, the choice of a meaningless null hypothesis as a "strawman hypothesis", that can easily be rejected, has been blamed by ZILIAK and MCCLOSKEY (2008) as being part of the "cult of null hypothesis significance testing" (NHST).

[2] Some authors prefer to speak of a "non-rejection" and avoid "acceptance" following the approach of falsification, and also to avoid the wrong conclusion that H_0 is actually true.

Figure 1. Statistical hypothesis testing



Source: NEYMAN and PEARSON (1933)

test based on the test statistic fails to reject the null hypothesis.

The complete decision-theoretic framework as proposed by NEYMAN and PEARSON (1933) further involves the definition of an alternative hypothesis H_A and the determination of the test statistic's distribution under H_A. The distribution of the test statistic under H_A is used to determine the type-II-error $\beta = Prob\,(not\text{-}reject\ H_0 \mid H_A\ is\ true)$ and the power of the test $1-\beta = Prob\,(reject\ H_0 \mid H_A\ is\ true)$ (see Figure 1)[3]. In econometric applications, however, alternative hypotheses are often not explicitly spelled out, which renders the determination of β-errors and power calculations impossible.

Even stern critics of the concept of statistical hypothesis testing do not deny that p-values contain some useful information. Loosely speaking, the p-value informs how compatible data are with a null hypothesis (WASSERSTEIN and LAZAR, 2016). Thus, they are a quantitative tool to challenge our initial belief and can be considered as a "first defense line against being fooled by randomness" (BENJAMINI, 2016). However, one should not get confused by this statement. From the above definition of a p-value it follows that they are derived from the sample data and thus observed p-values are random themselves. They vary from sample to sample, a characteristic, that is sometimes labelled as "p-value dance" (GREENLAND, 2019).

Another characteristic of p-values is that they merge information regarding the size of an effect (the difference between the estimate and the hypothesized value) and the precision of the estimate (the standard error of the estimate). This "confounding" of information is per se not a problem (GREENLAND, 2019), but it facilitates a common confusion of statistical significance and economic importance (GELMAN and CARLIN, 2017). If enough data are available, the standard error of the estimate becomes small and in turn, even a small difference between the estimated model parameter and its hypothesized value is classified as "significant" regardless of its practical relevance. Conversely, large effects may not become statistically significant in small samples. In response to this potential confusion, some authors suggest not to use the term "significant" in empirical applications anymore (HIRSCHAUER et al., 2019; WASSERSTEIN et al., 2019).

A common misunderstanding that has been deplored in the p-value debate, applies to the interpretation of the outcome of statistical tests as a proof that either the null hypothesis or the alternative hypothesis are true or wrong (GREENLAND et al., 2016). According to GIGERENZER (2018) researchers are driven by the desire to provide empirical evidence for or against a hypothesis and hence p-values are erroneously interpreted as $Prob\,(H_0 \mid data)$. P-values are related to this conditional probability via Bayes theorem, i.e. $Prob\,(H_0 \mid Data) \sim Prob\,(Data \mid H_0) * Prob\,(H_0)$,

[3] HIRSCHAUER et al. (2021) emphasize that the concepts suggested by FISHER (1925) and NEYMAN and PEARSON (1933) are "two different kettle of fish". While FISHER aimed at inductive inference, i.e. identifying the most rational belief given the available data, NEYMAN and PEARSON'S statistical decision theory provides behavioral rules for repeated decisions.

where $Prob\,(H_0)$ denotes the a-priori probability of the null hypothesis. Nevertheless, these probabilities are different entities and equating them would constitute a "fallacy of reverse inference" (KRUEGER and HECK, 2019). Thus, it would be incorrect to conclude from a p-value larger than 0.05 (or any other predefined threshold) that an economic effect is absent or in other words: "absence of evidence is not evidence of absence" (ALTMAN and BLAND, 1995). In real-world applications, this is especially relevant when considering very rare but very impactful events. Likewise, it would be wrong to infer from a small p-value that a specific alternate hypothesis is true. A small p-value merely reflects a misfit of the null hypothesis (under maintained model assumptions) to the data. A small p-value is compatible with many alternative hypotheses and might also be caused by a violation of other model assumptions, e.g. homoskedasticity of error terms in a regression model.

A related problem is the interpretation of p-vales or significance levels as false discovery rates (FDR) (HIRSCHAUER et al., 2016). A FDR defines the probability of rejecting the null hypothesis though it is true. It is an unconditional probability that depends on the significance level, the probabilities of H_0 and H_A being true as well as the power of the test (COLQUHOUN, 2014). Apparently, the significance level α captures only a part of the FDR, because it is the conditional probability of rejecting the null under the assumption that H_0 is true.

Finally, it has been stressed in the literature that $1-p$ does not measure the probability of replicating an observed result. GIGERENZER (2018) provides a simple example to illustrate this "replication fallacy". If H_0 and H_A reflect the hypotheses that a dice is fair or loaded, respectively, and two times "six" is observed, one would reject H_0, because the probability of this event under H_0 is $\frac{1}{36} = 0.03 < 0.05$. However, this does not imply that one can expect to observe two sixes in 97% of all future dice throws.

2.1.2 Erroneous Applications of Significance Tests

Even if the notion of p-values is well understood by applied econometricians, several problems prevail that may invalidate the calculation of p-values and undermine conclusions that are derived from a statistical test. Here we focus on three issues that are highlighted in the current p-value debate, namely multiple testing, inference with data that do not constitute a (random) sample and questionable research practices such as p-hacking, HARKing (**H**ypothesizing **A**fter the **R**esults are **K**nown), and selective reporting.

Multiple Testing

Multiple testing becomes an issue if several individual hypotheses are tested with the same data set (and not with different data sets) (ROMANO et al., 2010). If α is the desired significance level and m hypotheses H_i are tested, then the probability of getting at least one significant result by chance is:

$$Prob(\text{at least one significant result}) = 1 - Prob(\text{no significant result}) = 1 - (1-\alpha)^m \quad (1)$$

This probability, which depicts the familywise error rate (FWER), exceeds the significance level α considerably if m is large. Several proposals have been made to address this accumulation of type-I-error. These correction procedures control either the FWER (e.g. Bonferroni correction) or the FDR (e.g. Benjamini-Hochberg method). While a correction of significance levels is standard in biostatistics, particularly in genomic applications, it is often ignored in socioeconomics. This begs the question how relevant the consideration of multiple testing issues is in economic applications. HIRSCHAUER et al. (2018: 137) argue that "*multiple testing is evident in multiple regression analysis whenever researchers independently perform and interpret more than one test on one data set*". Multiple testing can definitely lead to an inflation of "significant" results in explorative studies, where regression models are fed ad hoc with available data and p-values are scanned a-posteriori. If, however, the specification of multiple regression models is guided by economic theory, which is reflected by a set of predetermined hypotheses about the sign and the size of specific model parameters, no adjustment of significance levels is required. This holds a fortiori in situations where a single hypothesis is of particular interest. Adjusting the significance level of the variable of interest would unnecessarily deteriorate statistical power (ALBERS, 2019).

Non-Random Samples

A fundamental objection against statistical inference is raised by HIRSCHAUER et al. (2020) in case of full population surveys. They argue that displaying p-values does not make sense, because there is nothing to infer, and sampling error does not exist. Obvious examples are studies that search for relationships among variables using data from all existing entities (e.g. individuals, states, countries) in a predefined

population. However, it is not that clear to which situations this "urn model" applies and to which not. For example, in price analyses often data of all (available) transactions can be accessed that occurred in a specific market in a certain time period. Is it inappropriate to conduct statistical inference and hypothesis testing regarding price determinants using a full sample? The answer is "no", at least if one can think of observed prices as an outcome of a data generating process. More data will be generated by this process in the future and even in the past more transactions could have been potentially observed. That means, the true population size is unknown and the "full sample" is still a random sample.

When inferring from observed realizations to the properties of the unknown data generation process by means of a statistical model, one has, of course, to consider selectivity issues and the fact that the data generating process can change over time – even though this can be rather challenging given the uncertainties involved. A related issue is the use of non-random samples for inferential reasoning. Non-random sampling techniques include convenience sampling, quota sampling or snowball sampling. These techniques became increasingly important and are nowadays quite common in survey-based social science. Several potential problems arise with non-random samples (cf. ELLIOTT and VALLIANT, 2017). Selection bias occurs if the sample differs from the non-sample part of the population such that the sample cannot be projected to the population of interest. Another problem is attrition, i.e. the systematic drop-out of participants in a panel. There is a controversial discussion whether or not non-random samples should be used for inferential statistics. HIRSCHAUER et al. (2019) argue that convenience sampling precludes the use of p-values because researchers run the risk of misestimating coefficients and standard errors, at least if selection bias is not adequately considered. In contrast, SMITH (1983) and ELLIOTT and VALLIANT (2017) show how quasi-randomization and super-population modeling can mitigate potential biases and under what assumptions non-random samples still can be used for statistical inference.

P-Hacking, HARKing, Selective Reporting

A couple of questionable research practices have been spotted that target at producing statistically significant results, which presumably have a higher likelihood of being published. P-hacking describes practices by which test procedures and model specifications are adjusted to attain statistically significant results with generally lower p-values. The same applies to the transformation of dependent or independent variables, the removal of influential observations or the definition of the eligible data set (see BRUNS and KALTHAUS, 2020 for an example from innovation research). P-hacking often comes along with selective reporting, i.e. only those results are reported that support a preferred hypothesis, while other are suppressed. This practice has also been labelled as "file drawer problem" (MERVIS, 2014). Finally, researchers could also explore the data and then retrofit theories, hypotheses, and narratives to findings after the results are known (HARKing). HARKing is problematic, because if one analyses data without preset hypotheses and *afterwards* finds something unusual or unexpected, that result could be a chance finding in view of the aforementioned multiple testing problem. Note that these malpractices can be either done intentionally (to deliberately manipulate statistical significance levels) or unintentionally (due to a lack of statistical education or through motivated reasoning or confirmation bias, cf. BASTARDI et al., 2011).

A large share of researchers in environmental economics has admitted questionable research practices in a recent survey (FERRARO and SHUKLA, 2020), and the economic literature in major general interest journals appears biased towards false positive findings, as indicated by an unusual hump in the distribution of p-values around p-value thresholds of 0.05 and 0.1 (BRODEUR et al., 2016; BRUNS et al., 2019). O'BOYLE et al. (2017) study PhD dissertations and subsequent research papers published from those dissertations and note that the "*the ratio of supported to unsupported hypotheses more than doubled*". While this may indicate *intentional* p-hacking or HARKing, HUNTINGTON-KLEIN et al. (2021) further demonstrate a large variation in results if different teams analyse the same data. Notably, BRODEUR et al. (2020) report that the extent of p-hacking and publication bias varies with the research design: They find instrumental variables (IV) and difference-in-difference (DID) techniques to be more prone to p-hacking and publication bias than randomized controlled trials (RCT) or regression discontinuity designs (RDD). However, in a recent reanalysis KRANZ and PÜTZ (2022) contest the robustness of these findings. Whereas questionable research practices happen at the level of a single study, publication bias is an outcome of the publication process and concerns multiple studies.

2.2 Proposed Remedies

While broad consensus about potential flaws of p-values seem to exist, opponents and proponents often disagree about remedies. In fact, proposals range from a complete ban of statistical hypotheses testing to a maintenance of current practice due to the lack of superior alternatives. In what follows, we structure these proposals and discuss their pros and cons. This section is restricted to the remedies directly tied to statistical practices. Wider implications for the agricultural economics community in supporting a cultural change and facilitating the implementation of promising remedies are discussed in Section 4.

Banning of Significance Testing and P-Values

In view of the aforementioned concerns some authors suggest not to display p-values or asterisks (HIRSCHAUER et al., 2019) or even to completely retire the concept of statistical significance (AMRHEIN et al., 2019; GIGERENZER, 2004), and some scientific journals followed these suggestions. For example, the American Economic Review and Econometrica discourage the use of asterisks to denote the significance of estimation results. The former also recommends to present standard errors[4]. This critical view, however, is also challenged: VERHULST (2016), GELMAN (2016) and BENJAMINI (2016) point out that most concerns about p-values also apply to alternative methods. FRICKER et al. (2019) try to assess the implications of a p-value ban empirically by analyzing the quality of 31 empirical papers published in "Basic and Applied Social Psychology" after this journal prohibited the use of the null hypothesis significance testing procedure (including p-values and statements about significance) in 2015. In their conclusions, the authors state "*we found multiple instances of results seemingly overstated beyond what data would support if p-values* […] *had been used. Thus, the ban seems to be allowing authors to make less substantiated claims* […]". At the time being, it appears unlikely that this radical approach will be copied by many scientific journals.

Emphasizing Economic Significance and Relevance with a Clear Distinction from Statistical Significance

In two empirical studies investigating the statistical practice in the American Economic Review in the 1980s (MCCLOSKEY and ZILIAK, 1996) and the 1990s (ZILIAK and MCCLOSKEY, 2004). MCCLOSKEY and ZILIAK highlight the importance of interpreting research results in light of their real-world substance. They argue that economists do a poor job in distinguishing statistical significance (the uncertainty of an estimate) and economic significance (the size of an estimate). Among other things, they propose to use confidence intervals to gauge the plausibility of an estimate and to use simulations to explore a distribution of possible economic outcomes. In addition, they emphasize the role of power analysis and considering the implications of type II errors and their associated costs rather than solely focusing on type I error. Although the two authors witness some improvements over time (ZILIAK and MCCLOSKEY, 2004) many problems prevail. As discussed by ROMMEL and WELTIN (2021), similar problems are present in major agricultural economics journals.

Replacing P-Values and Use of Bayesian Methods

The desire of researchers "*to turn a p-value into a statement about the truth of a null hypothesis*" (WASSERSTEIN and LAZAR, 2016) has prompted the promotion of the Bayesian approach, which, in principle, is capable to combine a data likelihood and a prior probability to derive a posterior probability. This Bayesian posterior inference offers the intuitive interpretation of a probability that a parameter of interest falls into a certain range (conditional on model assumptions), alleviating the troubles with interpreting p-values and confidence intervals under the frequentist paradigm.[5] It also provides the possibility to leave the dichotomous world of classical hypothesis testing with all its problems laid out above by rather comparing hypotheses in a probabilistic manner[6] (BENDTSEN, 2018).

[4] See submission guidelines (https://www.aeaweb.org/journals/aer/submissions/guidelines; last accessed October 17, 2022)

[5] It is interesting to note that IONIDES et al. (2017) interpret the ASA statement on p-values (WASSERSTEIN and LAZAR, 2016) as an attempt to advocate the Bayesian paradigm and to discourage researchers from using frequentist inference and deductive reasoning.

[6] For issues debated among those using Bayesian statistical inference see ACZEL et al. (2020).

There are probably two main reasons why the Bayesian approach has not yet overtaken the frequentist statistical inference despite an increasing use in recent times (GEWEKE et al., 2011). First, the derivation of posterior distributions of model parameters has long been a tedious and case-specific challenge, requiring to derive posteriors via probability calculus or simulation-based analysis. Recent advances in automated Bayesian inference ("probabilistic programming", see VAN DE MEENT et al., 2018 and BINGHAM et al., 2019) may offer a general solution in the medium-term for conventional and "big data" but this will also require a change in educating applied (ag-) economists. The second reason is the need to specify a prior distribution for all hypotheses and many scientists are reluctant to do so (KRUEGER and HECK, 2019), even though one could argue that frequentist approaches do this implicitly (e.g. BENDTSEN, 2018). Against this backdrop, HARVEY (2017) suggests the use of the minimum Bayes factor as a compromise that takes advantage of the Bayesian paradigm but bypasses the need to specify a particular alternative hypothesis and a full prior distribution. The Bayes factor is the ratio of the likelihood under H_0 and H_A, respectively. The minimum Bayes factor utilizes a special choice for the likelihood under H_A, namely the maximum likelihood given the data. If one is willing to express prior information as an odds ratio of the two hypotheses, one can derive a posterior odds ratio using Bayes' theorem:

$$\underbrace{\frac{p(H_0|Data)}{p(H_A|Data)}}_{\substack{posterior\\ odds\ ratio}} = \underbrace{\frac{p(Data|H_0)}{p(Data|H_A)}}_{\substack{(minimum)\\ Bayes\ factor}} \cdot \underbrace{\frac{p(H_0)}{p(H_A)}}_{\substack{prior\\ odds\ ratio}} \qquad (2)$$

Based on this expression, GOODMAN (2001) and HARVEY (2017) show how to derive "Bayesianized" p-values from the minimum Bayes factor, which provide the (a posteriori) probability that a hypothesis is true.[7]

Complementing P-Values by Additional Information

Instead of banning or replacing p-values, a couple of proposals have been made to complement them while maintaining the general framework of statistical significance testing. This is in line with AMRHEIN et al. (2017), who conclude that "*apparently, bashing or banning p-values does not work. We need a smaller incremental step*…". GREENLAND et al. (2016) emphasize that a statistical test should be interpreted carefully by examining effect sizes and confidence intervals instead of focusing just on p-values. Confidence intervals have the advantage of disentangling the size and the precision of an estimate that are merged in a p-value. Moreover, GIGERENZER (2018) reminds us that the design of insightful economic experiments requires sufficient statistical power. While power, effect sizes, loss functions, and type-II-errors are an integral part of the statistical testing approach based on Neyman-Pearson (1933, see also Section 2.1.1), they are typically ignored in the NHST ritual. BUTTON et al. (2013) show that in low powered studies the replicability of significant results is low. Furthermore, the positive predictive value (PPV), i.e. the probability that a "positive" research finding reflects a true effect, is positively linked to the statistical power of the study. Unfortunately, a meta-analysis conducted by IOANNIDIS et al. (2017) reveals that empirical economics research is often severely underpowered. However, at least in some research areas, particularly in experimental economics, power calculations started becoming a norm. Power calculations have to be performed before data collection, and there are different approaches, typically focusing either on the sample size, the acceptable alpha and beta errors, or the minimum detectable effect size (see KANG, 2021, for a practical introduction).

Multiverse Analysis

Different research teams can come up with fundamentally different conclusions even when they are working with the same data and research questions (HUNTINGTON-KLEIN et al., 2021). Another problem is that researchers may strategically report robustness tests if they support a preferred narrative (YOUNG and HOLSTEEN, 2017). Specification curves acknowledge this problem by running a wide range of plausible models that could for instance include different sets of covariates (STEEGEN et al., 2016). The outcome is not a single p-value linked to a single estimate, but a distribution of plausible estimates and p-values that define a distribution of plausible results for a reasonable set of models (see Chapter 7 of CHRISTENSEN et al., 2019, for more details).

Replication Studies and Meta-Analysis

Statistically significant research results may be the outcome of chance. To detect false positive findings,

[7] We refer to Harvey (2017: 1424) for an example that illustrates how an a-posteriori probability for a hypothesis can be derived from the minimum Bayes factor and to what extent it differs from a p-value.

researchers have advocated replication studies. Replication can take different forms (see Chapter 9 of CHRISTENSEN et al., 2019 or CLEMENS, 2017). It may involve reanalyzing the original data of a study with the same (verification) or different (reanalysis) methods. It can also involve new data collection applying the same methods (direct replication) or different methods (extension). Direct replications of economic experiments show that the rate of false positives is substantially higher than expected from pure chance alone (e.g. CAMERER et al., 2016). Replication studies and the aggregation of studies for meta-analysis can increase the confidence in research findings. A recent study has shown that studies that did not replicate are more widely cited than those that replicate (SERRA-GARCIA and GNEEZY, 2021), calling into question the use of citations as an indicator of scientific quality, the power of the research community to self-correct more generally, and highlighting the need to provide incentives and resources for replication studies.

Pre-Registration, Registered Reports, and Results-Blind Review

Other remedies target both the individual study level and the scientific publication process. Pre-analysis plans are written commitments to a specific data analysis before the data are obtained or collected (see OLKEN, 2015, or BANERJEE et al., 2020, for a detailed discussion). In a pre-registration, researchers also submit this plan to a repository, thereby increasing the commitment by making it publicly available and referring to the pre-registration in publications. Although these two instruments substantially limit researcher degrees of freedoms and may successfully safeguard against p-hacking, they only address the producers of research findings, whereas editors and reviewers could still exhibit bias against non-significant findings. Registered reports or results-blind reviews have been proposed as a solution to this problem. In a registered report, a study design and analysis plan are submitted to a journal and reviewed by peers in a "first stage report" before data collection. If the authors pass this stage, the journal and publisher commit to a publication irrespective of the results. Results-blind review mimics this process, by suppressing results from the manuscript, hence allowing reviewers and editors to focus on research questions and methodological rigor. For a practical guide on how to write a registered report the reader is referred to ARPINON and ESPINOSA (2022).

3 Views of the Community

We conducted a survey among agricultural economists and social scientists in Germany, Austria, and Switzerland to explore the views of the respective agricultural economics community in 2020. The survey was administered in English to address non-German speakers in the three countries. The general objective of the survey was to get an overview on the problem perceptions, knowledge, practices, and attitudes regarding econometrics and statistics. The survey was not meant to be the basis for a rigorous empirical analysis for answering a specific research question or testing hypotheses.[8] The survey started with a short introduction, data use, and contact information. Consent to participate was obtained. The first part of the survey covered perceptions of the debate on statistical practices and knowledge on the topic. The second part dealt with practices and preferred remedies. Finally, respondents had to provide some personal details. We refer to the appendix/online appendix to see the full survey, data, and code.

3.1 Survey Design and Respondent Characteristics

The survey was distributed in the summer and fall 2020 to all members of the German Association of Agricultural Economists (GEWISOLA), members of the Swiss and Austrian associations of agricultural economists, an e-mail list of early career researchers in agricultural economics in Germany, and doctoral students enrolled in the Doctoral Certificate Program in Agricultural Economics. There is some overlap between these groups. We estimate that approximately one thousand people have been invited to participate in the survey by mail.

Different distribution links for these channels indicate that approximately 34% have entered the survey from the GEWISOLA invitation, 31% from the doctoral certificate program and 25% from the early career researchers e-mail lists. The remaining re-

[8] One reviewer criticized survey design and sample selection, the lack of clearly stated hypotheses and the missed opportunity to show a good practice example for statistical inference. We acknowledge these limitations. Data and code used for the analysis are shared by HECKELEI et al. (2023), https://doi.org/10.25625/WOBCK0. We like to actively invite comments and rigorous follow-up studies on statistical practices in the community.

spondents came from the Austrian and Swiss societies or other sources.

In total, 305 respondents opened the link, but there was a high drop out on the first screens. We removed one response due to highly inconsistent responses. For the analysis, we use all respondents who completed the survey at least until the second last screen, leaving us with a total of 108 respondents. Note that there are still missing observations for some of the variables which could lead to a lower number of observations for some of the recorded items, as we did not force answers on any of the questions (i.e. all responses were voluntary). All presented analysis is descriptive and must be viewed as explorative, as it stems from a self-selected sample.

The median response time in the survey was approximately 14 minutes. Most respondents either had a PhD (55%) or were in the process of obtaining one (39%). Participants indicated their gender as male (61%), female (35%), or did not indicate a gender (4%). The average age was 38 years (with a range from 25 to 72 and a median of 34; SD = 10.4). Approximately 39% stated that they were permanently employed. More than half of the respondents had five years or less of research experience. A little more than half of the respondents stated that they had published three or less research peer-reviewed research articles over the past five years.

3.2 Problem Perception and Knowledge

The survey started with several general questions on the perception of the problem. We openly asked whether generally speaking respondents "think there are problems with the way the scientific community represented by the Austrian, German, Swiss, and European associations of agricultural economists (GEWISOLA, ÖGA, SGA, and EAAE) deals with statistics and econometrics in research and teaching?" Respondents were asked to use a ten-point scale to differentiate their responses (1 = no problems at all to 10 = a lot of problems). The mean response was 5.13 (SD = 2.14; median = 5). We also asked people to assess their own statistical and econometric knowledge on a ten-point scale where higher values indicate better knowledge (mean = 6.38; SD = 1.59; median = 7). Finally, we asked for an assessment in which percentile respondents would place themselves in terms of knowledge relative to the target community. The median respondent placed themselves in the top 50%.

We used the six survey items developed by OAKS (1986) to get an overview on knowledge of the correct interpretation of a p-value. These items have been widely applied to different samples of researchers (see GIGERENZER, 2018, for an overview of studies in different academic communities). Respondents were presented with the following scenario:

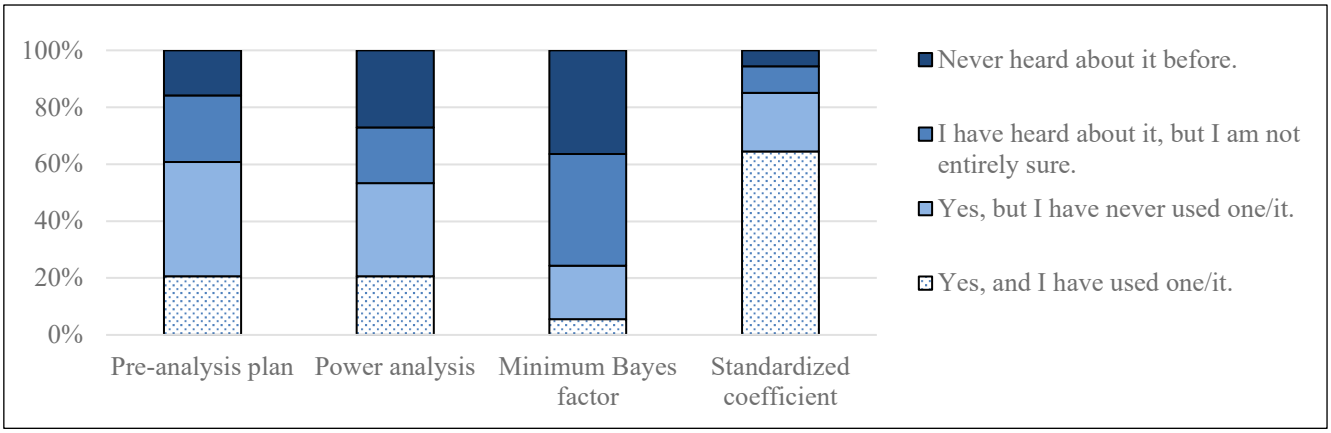
"Suppose you have an exogenous variation that you suspect may alter the outcome you are interested in for a certain task or behavior in a given population. You compare the means of your control and treatment groups (say 20 randomly selected subjects in each sample). Further, suppose you use a simple independent means t-test and your result is (t = 2.7, d.f. = 18, p = 0.01). Please mark each of the statements below as "true" or "false". "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct."

Table 1. Overview on endorsement of statements

Statement	Percentage of respondents wrongly endorsing the statement as true
• You have absolutely disproved the null hypothesis that there is no difference between the population means.	26.3%
• You have absolutely proved your alternative hypothesis that there is a difference between the population means.	18.3%
• You have found the probability of the null hypothesis being true.	21.4%
• You can deduce the probability of the alternative hypothesis being true.	48.4%
• You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.	57.6%
• You have a reliable finding in the sense that if, hypothetically, the study was repeated a great number of times, you would obtain a significant result on 99% of occasions.	51.2%
Percentage of respondents wrongly endorsing at least one statement (among those who responded to all statements)	81.3%

Notes: adapted from GIGERENZER (2018)
Source: own calculations

Figure 2. Knowledge and applications of remedies (n = 107)



Source: own calculations

Table 1 presents the six statements and displays responses. All of the statements are false and represent different delusions regarding the meaning of a p-value (GIGERENZER, 2018). Hence, the percentage of respondents endorsing a statement as true may be viewed as an indicator of knowledge. Approximately 80% of respondents who have responded to all six items endorse at least one of the delusions. Note that the number of correctly answered statements only weakly correlated with the item of self-assessed knowledge above (Spearman's rho = 0.1; n = 75).

We also asked about knowledge of and experience with some of the remedies/practices outlined above (Figure 2).

3.3 Practices and Attitudes

We asked people for their agreement with several survey items to understand attitudes and practices regarding statistics and econometrics (Figure 3). There were high levels of agreement with the importance of economic significance (as traded off against statistical significance) and data sharing practices. At the same time, respondents stated that they feel pressured to produce statistically significant findings. Many stated they have committed or witnessed p-hacking.

3.4 Views on Remedies and Suggested Fields of Future Action

We asked about an assessment of how useful remedies were perceived (Figure 4). There was a high perceived usefulness of confidence intervals, display of effect sizes and standardized effect size, as well as summary statistics. There was little perceived usefulness in the overall ban of p-values or asterisks/stars from research results or publications.

To identify target areas and fields of action we asked respondents to state who would have the largest impact on one's statistical and econometric practice (Figure 5). Respondents assigned high importance to colleagues, teachers and educators, as well as reviewers as drivers for their own statistical practice.

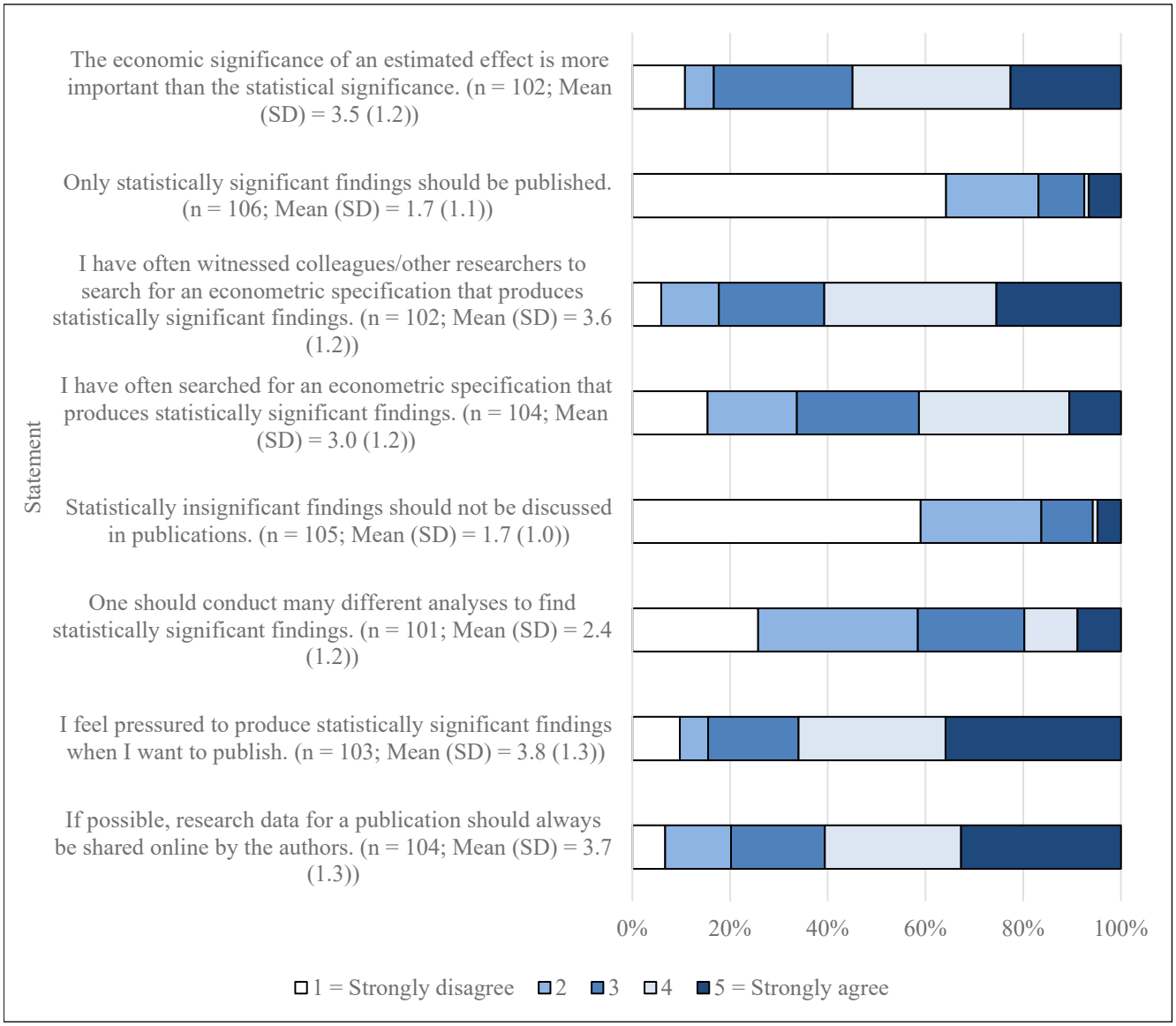
3.5 Discussion of the Survey Results

The survey results reveal that knowledge on the interpretation of statistical hypothesis testing and p-values, and the potential remedies of current malpractices may still need a substantial educational push at various levels. At the same time, the community feels fairly strong about not abandoning p-values altogether (50% consider this remedy "not at all useful"). The dichotomous nature of the current practice in hypothesis testing is seen somewhat more critical (a clear majority considers abandoning the use of stars/asterisks at least "somewhat useful").

At least a 70% majority of respondents view certain practices offering information beyond the pure outcome of hypotheses tests and that are not yet widely applied at least as "fairly useful". These include those that allow better understanding or visualizing uncertainty of statistical results (display of confidence intervals) and understanding better the (relative) economic importance of the determinants considered (standardized coefficients and economic effect sizes).

To the extent that these remedies are known, respondents consider power analysis (n = 62) and preregistration plans (n = 93) at least "fairly useful" with a majority larger than 60%. Hidden behind these responses might be a differentiated view on the question for what type of analysis such remedies are useful. They are discussed and implemented in the context of controlled experiments, where sample size and treat

Figure 3. Overview on attitudes and practices



Source: own calculations

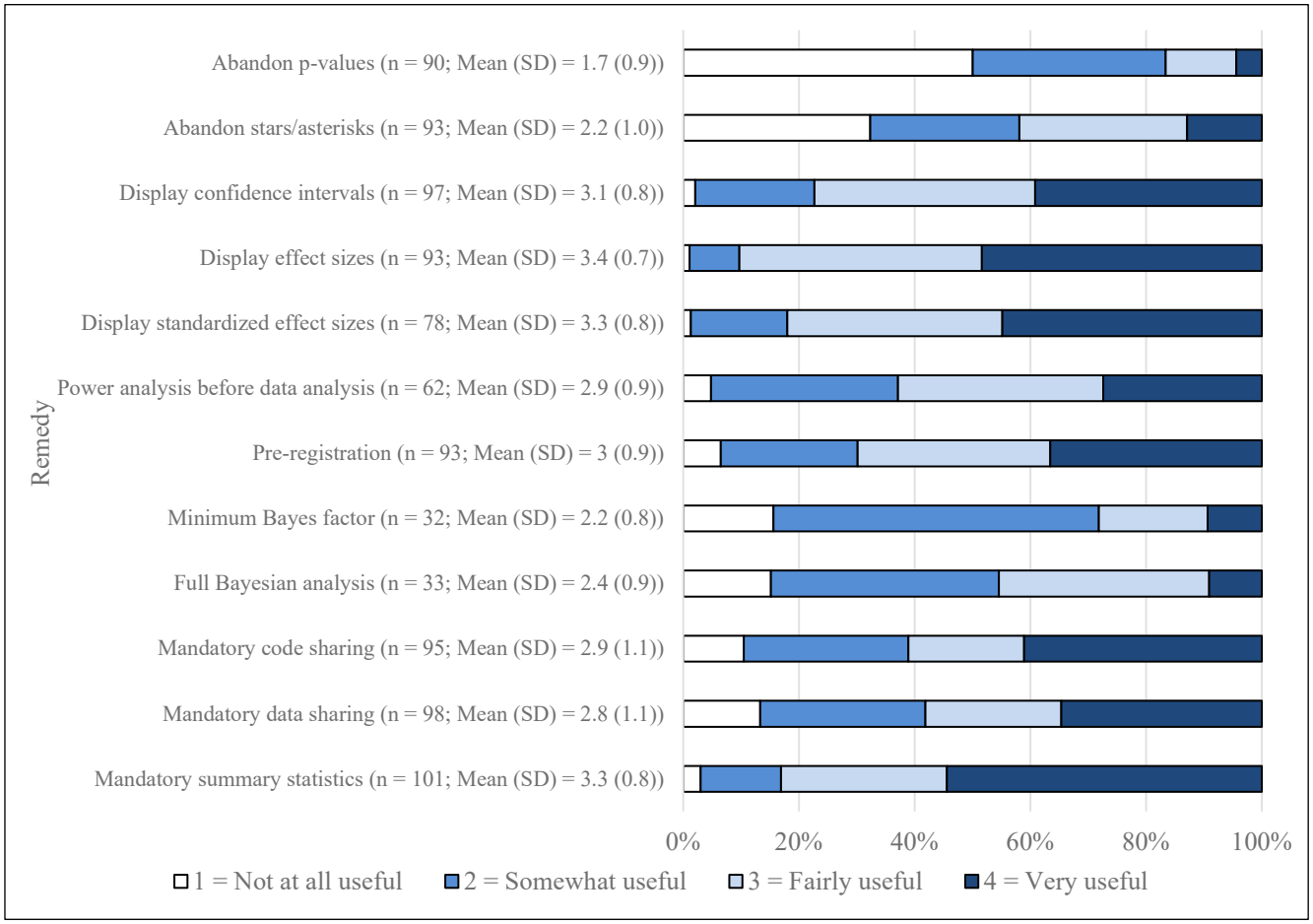
ments are often part of the deductive analytical design decided upon before the data collection. Also, in these cases good priors are often available. Pre-registration may in principle also be considered for observational or even explorative studies to prevent that the research design is driven by initial results in a not fully reflected empiricist manner (cf. HAVEN and VAN GROOTEL, 2019; OLKEN, 2015).

Less than one third of survey participants feels comfortable to judge Bayesian remedies (Minimum Bayes Factor and full Bayesian analysis with n = 32 and 33, respectively), but those who do are moderately positive about them (above 70% consider them somewhat or fairly useful). The low level of participation in these questions may indicate a limited amount of training and experience with Bayesian analysis. The only moderately positive perception could be related to the effort still needed to develop case-specific Bayesian approaches discussed above.

The community feels quite strongly about the usefulness of mandatory data sharing, code sharing and summary statistics with majorities larger than 80% considering them at least "somewhat useful". More than a third consider data and code sharing and more than half summary statistics "very useful". The comparatively moderate responses regarding the data sharing may reflect the not uncommon situation that confidentiality requirements of individual firm and consumer level data often restrict the possibilities to share data. The reason why data summary statistics do not even have a stronger support may lie in the view that it alone does not help to solve the statistical

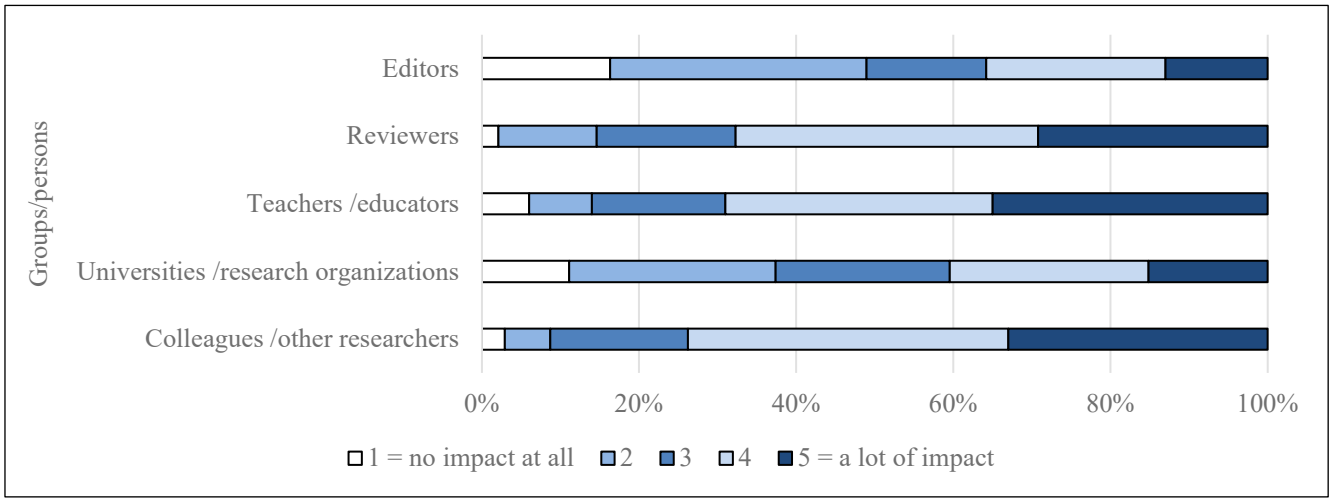
inference issues of the framing even if more respondents may view it as a key ingredient to a transparent and "data-aware" empirical economic analysis. Working towards well-documented code and data can benefit from integrating open practices foresightedly with research project work flows. Senior colleagues may lead the way to facilitate and enforce good practices.

Figure 4. Attitudes on remedies



Source: own calculations

Figure 5. Who affects statistical practice the most



Source: own calculations

Respondents clearly consider teachers/educators to have the largest impact on statistical practices (almost 70% in the top two impact categories). This coincides with the in parts limited knowledge on some statistical misinterpretations and malpractices found above and points at the longer-term effort needed to fundamentally change practices through a revision of curricula and teaching methods.

It is quite interesting to note that respondents view reviewers as affecting statistical practices more (almost 60% in the top two impact categories) than editors (less than 40%). It raises the question if editors of the journals relevant for the community are rather passive with respect to setting and guarding editorial policies on statistical practices and/or often shy away from evaluating/adjusting/weighing/complementing reviewer comments with respect to the editorial standards in this respect. Editors could have a crucial role in changing statistical inference practices if they took an active stance on it.

4 Implications for the Community

4.1 Research Process

The most important implication resulting from the p-value debate from the viewpoint of researchers is to avoid what GIGERENZER (2004) blamed as "mindless statistics". Many fallacies may arise when applying the statistical hypothesis testing framework; we argue that the merit of the p-value debate is to recall (at least some) potential fallacies to researchers' minds. Being aware of these problems is in fact the most undisputable implication. Yet we are reluctant to recommend a ban of p-values or the use of asterisks in general. This view is shared by the majority of our survey participants, and at least in parts our community (MARGARIAN, 2022). Correctly calculated and interpreted p-values contain useful information about the underlying statistical hypotheses that otherwise would be neglected. In a recent paper, Nobel Prize laureate Guido Imbens characterizes economic applications where p-values are dispensable and where they contain relevant information (IMBENS, 2021). Testing a null hypothesis versus an alternative hypothesis is meaningful in some situations and examples include the efficient market hypothesis, market integration, or the existence of speculative bubbles. Moreover, it is often necessary to test whether data show certain statistical properties, such as stationarity, variance homogeneity or spatial and temporal independence. In these situations, a decision shall be made based on a statistical decision rule. This then necessarily includes a threshold determining what the decision will be.

In many economic applications, however, testing against a null hypothesis of "no effect" is not of particular interest. For example, it is generally not very insightful to test whether farmers' education increases farm income or not, whether a gender pay gap exists or not, or whether invest-ment aid stimulates investment demand or not.[9] Here the magnitude of the (treatment) effect is what matters and the causal mechanism, e.g. how investment aid stimulates investments. We believe that in situations, where no specific decision on a hypothesis has to be made, it suffices to display standard errors or to interpret p-values as indicators of the general compatibility of the data with the corresponding hypothesis. In these cases, specific thresholds have no defendable meaning beyond a long-practiced ritual. Given the documented publication bias around these thresholds (e.g. BRODEUR et al., 2016), avoiding the use of asterisks potentially reduces incentives for p-hacking. Whether with or without specific significance levels, the important point is, however, that we as researchers derive hypotheses based on logical thinking and theories, and apply statistical analysis "mindfully" in light of an underlying theoretical concept, and to avoid extreme forms of empiricism.

In situations where statistical hypothesis testing makes sense, the following aspects deserve attention when designing, conducting and interpreting statistical tests. Perhaps the most basic question is whether observed data can be considered as a random sample, i.e. as an outcome of a random data generating process, because this is a prerequisite for inferential statistics. If, in contrast, data fully describe the entire population, there is no need for statistical testing. If in this instance, inferential reasoning is based on the notion of a superpopulation, this should be clearly labelled and defined. Moreover, if data come from convenience sample, any source of potential bias regarding estimates of regression coefficients and standard errors should be carefully considered and discussed. Non-proportional stratified samples like the Farm Accountancy Data Network (FADN) data with over-

[9] Of course, one can imagine research settings in which testing a nil hypothesis related to education is not trivial, for example whether the marginal effect of an additional year of schooling is zero or not. This shows that a clear classification of applications in which NHST is (in)appropriate, is hardly possible.

and underrepresented groups can lead to a bias when estimating a population mean without an appropriate correction by sampling weights (see NEUENFELDT and GOCHT, 2014, or BARKASZI et al., 2009, for details of sampling in the FADN). However, weights may also be used to reduce estimator variance in such contexts making their choice a non-trivial task depending on model specification. For a deeper discussion of correction possibilities we refer to DUMOUCHEL and DUNCAN (1983).

From the stated objectives of an empirical analysis, we recommend authors to be clear about whether their study is explorative or whether they aim at testing of hypotheses that are derived from theory. This distinction is important, because in explorative studies that try to identify potential relationships among dependent and explanatory variables, a multiple testing problem is immanent that calls for an adjustment of significance levels to avoid false rejections of null hypotheses. Unfortunately, this distinction is not always straightforward in applied economics, because theoretical predictions do not cover all aspects of econometric model specification. That is, even if theory suggests a positive or negative relationship among economic variables, it might be necessary to "explore" the appropriate functional form in a regression model or the number of lags in a time series model (OLKEN, 2015). We do not consider this search for a data fitting model specification per se as "p-hacking". It is important to use objective and appropriate criteria (e.g. F-tests, RESET tests, Breusch-Pagan tests, non-parametric specification tests) for deciding on the functional form, the number of lags or appropriate econometric estimators. The crucial point is to describe this process in a transparent manner and to report the results of alternative model specifications instead of presenting only selected results. In these instances, careful documentation of data and code, as well as tools such as multiverse analysis, may address selective reporting more appropriately (STEEGEN et al., 2016).

The need for flexibility during the model specification process in many forms of analysis limits the scope of instruments that have been proposed to prevent p-hacking, e.g. pre-registration. Pre-registration is useful in particular for deductive work, which involves primary data collection. However, even for these cases, recent studies show that researchers who use pre-registration rarely specify pre-analysis in sufficient detail (BAKKER et al., 2020). In other instances, there may be a risk that pre-analysis plans limit the reporting of relevant findings (BANERJEE et al., 2020). The effectiveness of pre-registration is also sensitive to the platform used (BAKKER et al., 2020). Yet, perceived benefits from pre-registration outweigh the costs in many instances, and major benefits emerge from thinking about analysis before the data are collected (LOGG and DORISON, 2021). Therefore, when applicable, pre-analysis plans should, for instance, enter PhD- and third-party funded project plans and output/performance measures as milestones to account for the required resources and to provide a structure for monitoring and enforcement. Current schemes of performance measures of universities and researchers seem to not sufficiently value such efforts. In conclusion, pre-registration and pre-analysis plans are a useful tool in many fields that involve primary data collection, while the risk that pre-registration becomes ritualized and a form of virtue signalling if not complemented by more fundamental cultural change remains (BUCK, 2021).

Another important insight of the discussion about verification, re-analysis, and aggregation of scientific research is the need to pay more attention to adequate power of statistical tests. This is important to avoid "false negatives" but also to ensure a high positive predictive value, i.e. the likelihood that a claimed relationship is actually true (CHRISTENSEN and MIGUEL, 2018). Researchers have at least two options to control statistical power. First, via sample size which can be determined for a desired power level in an a priori power analysis, given that information about the effect size is available, e.g. from pilot projects or similar studies (IOANNIDIS et al., 2017). Computational software is available that supports this calculation for many research designs, e.g. G*Power (FAUL et al., 2007). The second option is the choice of the statistical test. In time series analyses, for example, the use of panel unit root tests can help improving power compared with standard unit roots tests, which are known to have low power.

Regarding the interpretation of statistical test results, two recommendations appear unchallenged. First, presentation of statistical results should include effect sizes, and the interpretation should involve the economic relevance of variables rather than focusing solely on their statistical significance. Plots of coefficients or marginal effects along with their confidence intervals to illustrate related uncertainty (in the notation of AMRHEIN et al., 2019, "compatibility intervals") may support this. Second, p-values should be interpreted as what they are, the likelihood for ob-

served data given a null hypothesis, though it is tempting to consider them incorrectly as likelihood of a hypothesis as noted in 2.1.1. HIRSCHAUER et al. (2016) provide an illustrative example of how the use of sloppy language turns a statistically correct statement into a wrong one. We thus strongly recommend to use precise wording when interpreting the results of statistical hypotheses tests, along with careful documentation of the test procedure. Our survey showed strong support for confidence intervals and descriptive statistics, and authors and journals may consider them even more. Although confidence intervals are easily calculated from standard errors and coefficient estimates, displaying them may change the reader's perspective.

4.2 Publication Process

Journals can achieve a lot through submission guidelines, which should be up to date and enforced. For instance, clear editorial statements (BLANCO-PEREZ and BRODEUR, 2020) and check lists on how to report statistics and results of statistical testing may be useful and can have an impact (see GIOFRÈ et al., 2017); some authors even call for "statistical co-editors" (WEHRDEN et al., 2015). A prerequisite for any change to the better is, however, that all involved stakeholders are clear in their communication, reach their audience effectively and editors take responsibility to moderate reviews carefully and decide according to clearly communicated rules.

Our survey showed support for open data and methods and we recommend that the agricultural economics journals make code and data sharing mandatory.[10] Data and code sharing do not only increase transparency of results, they also make it easier to discover data manipulations and in turn, researchers will become more reluctant to violate good research practice. This in turn, will improve quality and reproducibility of the results.[11] As shown by a recent study of leading economics journals, excess statistical significance (i.e. inflated effect sizes) is substantially reduced if data sharing is mandated (ASKAROV et al., 2022). Clearly, relevant journals in a field should pursue similar policies in this regard to avoid a selection of authors into journals with less restrictive policies. In some cases, sharing of raw data may be hampered by data protection regulations. This applies to farm level data, such as data from the FADN. As a minimum, authors should document in this case how they got access to the data to allow the reader to pursue the same path (if possible) and replicate the analysis using the code provided. However, other ways of reproducing the results are preferable, for instance, by remote access. Moreover, if data are bought from and owned by third parties, researchers cannot easily share them, yet also here, replicability can be made available by remote solutions together with third parties; owners or providers of data sets are expected to be interested in most reliable results produced with the data. These additional efforts are again resource-consuming and could be alleviated if raw data collected by the public (e.g. FADN data) would be generally accessible in anonymized form for scientific research institutions. In turn, all scientific institutions should commit to FAIR principles[12] for research data, and universities should collaborate for efficient research data management processes which would benefit the whole community.

Researchers have highlighted problems with the direct replicability of research results especially in experimental economics and business economics' studies, and the sensitivity of research results to context (CAMERER et al., 2016; RAHWAN et al., 2019). When engaging in a replication, authors bear major publication risks when editors predominately select manuscripts on novelty. New publication formats could lower these risks. In a recent call for papers in the journal *Applied Economic Perspectives and Policy* (AAEA, 2021), the editors invite replications in a two-stage format. Replication protocols are reviewed *before* the bulk of the work is done, and the journal and editors commit to a conditional acceptance for publication for the selected proposals (or reject proposals). Adopting this format on a regular basis either in the form of special issues or regular sections could give rise to more replication attempts. Authors can substantially lower their risks of engaging in replication, and

[10] As an anonymous reviewer pointed out, this data and code sharing should also include the 'raw' data and the code for preparing the data for the final statistical analysis.

[11] As one reviewer pointed out, PÜTZ and BRUNS (2021) find that data and code availability policies can help to reduce reporting errors.

[12] FAIR principles for research data target at a sustainable data collection, processing and use. F stands for findable, where meta data should be made available, A indicates accessible, where meta data must be available, I stands for interoperable, i.e. clearly documented and applicable language, and lastly, R means re-usable, i.e. a clear data use agreement/license is required. For Germany, more details can be found for instance here: https://www.forschungsdaten.org.

the publication will not depend on whether the results are deemed interesting by the reviewers and editor.

Registered reports - a two-stage publication format where the study design is reviewed *before* the data collection (see LEMKEN, 2021, for a recent example of a first stage report in agribusiness consumer research) - could be embraced by more journals in the agricultural and food economics domain. Whereas a pre-registration only involves the authors, a registered report is integrated with the peer review and journal publication process. Hence, with a registered report, several important steps of the research and publication process are front-loaded, potentially reducing risks for authors and the research community in several important ways. Authors will benefit from feedback on their work already in the design stage. Other researchers become aware of what others are working on earlier, facilitating collaboration and innovation. Editors and reviewers evaluate studies on novelty and a sound research design, rather than results. In the future, research funders may even condition grants and research funds on the acceptance of registered reports for studies that involve primary data collection. As of today, only a few journals in which agricultural economists publish offer registered reports (PLOS ONE, Nature Human Behavior, Journal of Development Economics, Q Open), and more journals and editors may want to consider opening up for the format. Pre-registration can also be applied to some types of explorative and qualitative research, but it will be critical to adjust platforms such as the open science framework to the specific needs of the qualitative research community (HAVEN and VAN GROOTEL, 2019). We encourage editorial boards and scholarly associations to discuss data sharing policies, registered reports, and pre-registrationn and to communicate their conclusions (e.g. as in BARREIRO-HURLÉ, 2021).

4.3 Teaching

Based on our review of the p-value debate and statistical practices, and the discussion of respective sources and remedies, we see implications for higher education and teaching in the field of applied agricultural, resource and food economics. We qualitatively enrich our presentation of teaching implications by the perspective of survey participants in the German-speaking community. In fact, we conclude that from a perspective of today, the way of teaching research methods and statistics/econometrics contributed to the problem of a misuse of p-values, non-mindful statistics, and a more ritual-based use of empirical methods for hypothesis testing. We therefore discuss five implications for teaching at all levels of higher education, i.e. Bachelor, Master and PhD level, and for researchers as teachers. The statements in the survey that teachers/educators have the largest impact on respondents' statistical practices corroborate our motivation and demonstrate that teaching is a key part of the needed cultural change.

As a base, we clarify learning objectives: achieving statistical thinking rests on the ability to understand, apply and evaluate statistical (and other) methods for empirical research. The ability of critical reflection plays an important role to overcome the rituals-orientised use of methods and get students and researchers sensitized that strategic use of methods has been observed and can pose a problem. This way, from our perspective, teaching in the subject "methods for empirical research" shall combine statistics, data (science), and scientific working.

To reach these objectives and considering that higher education in agricultural economics typically rests on an interdisciplinary curriculum, we first suggest to generate a specific strand in the curriculum to impart a sound understanding of empirical research, including hypothesis testing. This strand covers modules on empirical research, specific methods, and scientific working. Modules for quantitative methods must provide a clear understanding of empirical methods and different ways of hypothesis testing, including statistical inference. Modules covering good scientific practice and scholarship ideally are based on Philosophy of Science, cover theoretical and behavioural models as base for hypothesis formation, and include ethics and transparency as key parts of existing code of conducts[13] in the community. Linked to the core modules covering research methods, appropriate research design, sampling and data collection, pre-registration, documentation of data and coding (research data management) set the base for teaching empirical models and methods for identification of effects, and how to distinguish between mindful and not so mindful empirical work.

Second, we see module structures that go beyond classical lectures with practical hands-on parts. Practi-

[13] Several examples exist, we refer to the German Research Foundation https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf and the European Association of Agricultural Economists https://www.wecr.wur.nl/EAAEUploads/Other/CodeOfProfessionalConduct.pdf.

cal training could be based on simulated data sets (BEKKERMAN, 2015) and hands-on inference, followed by contextual empirical applications. A main part of the module must be active practising, from the data preparation over the data analytics to the hypothesis testing and the interpretation of the results. This offers a more holistic understanding of statistical inference based on data. We also see here potential to improve links to typical subjects of study by including topics and examples from the agricultural, food and environmental economics domain. Vice versa, topic-related modules that use empirical findings as material could ideally integrate discussing the ways the empirical findings were generated. This could mean for instance, presenting and discussing research designs, data sets, used hypotheses, and methods for testing. Using interesting and timely examples can raise attention also for methods and stimulate critical reflection.

Third, the goal to enable students to strengthen their ability to critically reflect on their choice of method for their own research but also for future reviews calls for experience-based learning with a strong interactive component. We see the idea of experiential learning as a fruitful guide. For instance, opportunities for thinking, acting and reflecting can be created if besides pure assignments with applications, (poster) presentations and short reports about the data work are part of modules; at higher levels (advanced Master studies, pre-doc and PhD level), this part may include critical reflection of existing research as well as finding and presenting best practice examples. To foster interaction, digital platforms for exchange and interactive problem-oriented discussions in Wikis and forums can provide additional incentives and would offer alternative grading opportunities.

Fourth, we suggest that replication should become an important component in modules, but also in seminars and graduation theses. Recent initiatives by journals in presenting replication studies can be helpful and serve as reference for graduation theses in this direction. Clearly, we are not suggesting that pure replication can replace a thesis in its core, we suggest considering combinations, for instance, replication and literature review, another period or new data set but same question as a way to make replication studies more valuable.

Fifth, thesis writing needs systematic support, where a dynamic guideline that covers dos and don'ts in a sense of a checklist for orientation could be helpful. The dynamic nature would mean that this guide is never fixed and in parts under students' responsibility for continuous update but monitoring by teachers. Material for standards for empirical work, data handling/management and ethics must be provided and pre-selected by senior staff, while students discuss the material and prepare/develop and update their checklists. At higher levels, the student work may include contrasting examples based on empirical papers as well as critically assessing and discussing the procedures. Recently established asynchronous teaching with prepared videos to be viewed flexibly by the students could increase quality contact time with students focusing on specific problem sets, critical reflections, and presentation of own work. Here we see that respective associations could support substantially by offering platforms for exchange and share points for examples of the checklists mentioned above, also aimed at harmonizing implicit rules and support learning throughout the academic life.

Gigerenzer's idea of mindful statistics and statistical thinking seems to be inspired by observing ritual-type behavior. To infer on the behavioral reasons in our community, however, is not possible based on the literature review and the survey and would go beyond this review paper. Proper statistical education denotes an important piece in overcoming these problems that appear at first glance inside academia; however, we also like to note the "economic" relevance outside academia. For instance, students as future decision-makers in companies, ministries, international organizations, or other non-university employers, will decide under uncertainty and rely on empirical research to inform the decision-process. Correctly interpreting p-values and results of statistical inference will therefore be important also for society.

4.4 Research Culture

Changing statistical practices is a challenge as they develop in a complex dynamic interplay of what we have been taught and what experiences we make interacting with our peers in publication processes and collaborative research as we build our careers. Developed rituals are not easily changed, and such change requires a new consciousness to slowly penetrate all our academic activities. A long-term, cultural change of knowledge and norms is needed with complementary changes in teaching, research, and publishing activities that go beyond the definition of rules for use and interpretation of specific statistical tools. This is all the more difficult, as change must come from within a community with heterogeneous knowledge of statistics and diverse viewpoints on how to address challenges in statistical practice. Knowledge and the perception of problems are most likely not independent

of each other. That is, the support of remedies and change may oftentimes require good knowledge which can create a situation where a poorly trained community is locked in an inferior equilibrium of poor practice.

New rules and recommendations to use some statistical tools and not others will not alone ensure that research is conducted and papers written to primarily generate replicable scientific knowledge. The currently observed misuse relates to a considerable extent to the explicit or implicit expectation that certain findings are more interesting than others. What is required is a culture of acceptance of scientific work that is largely based on the theoretical and methodological rigor and where the perceived relevance arises from the questions asked, the methodology employed, and the data used but not from the results generated.

The quality of statistical analysis in economics falls and rises with the careful argumentation backed up by theories from the field of economics, social sciences and psychology, and related subjects that govern agents' decisions and respective results. A discussion of most likely mechanisms underlying the data generation process guards against pure empiricist interpretations of statistical results and the confusion of correlation in the data with "true" effects or causation (ANGRIST and PISCHKE, 2008). With a sound theoretical foundation, the conditionality of statistical results on the model employed in the analysis becomes transparent and thereby creates an inherent caution with respect to the interpretation of results. Some even argue that "both statistical foundations and basic statistics can and should be taught using formal causal models" (GREENLAND, 2020). Thinking carefully about "what matters" for economics actors will also help in recognizing that a dichotomous world of hypothesis testing is not sufficient to derive meaningful implications. The size of effects of policies or other determinants of economic behavior matter for stakeholders and should receive at least as much attention as the question whether there is an effect or not.

5 Concluding Remarks

We would like to conclude our paper with a few brief ideas on what could be done at the "policy level" to improve the situation in the short-term and to foster a cultural change of statistical inference and research practice in the long-term. Here we suggest a set of "top-down" measures that have some promise in bringing about the needed change jointly with the desirable "bottom-up" developments at the individual scientists' level.

Communication on best practices can clearly move forward right away. Here, scientific journals and connected learned societies can work together. Recent discussions and activities seem to lead towards a closer relationship between the GEWISOLA and the German Journal of Agricultural Economics (GJAE). A joint activity between the association and the journal can lead to setting standards of reporting statistical inference in journal articles that are then clearly communicated with the instructions to authors for the preparation of manuscripts and by the association to its members moving the community of reviewers. HIRSCHAUER (2021) suggests guidelines that might serve as a starting point for the discussion on the formulation of such standards.

Better recognition of the effort reviewers put into the publication process may go some way in alerting to the value of this resource scarce in quantity and quality. Some journals like the European Review of Agricultural Economics already have a best referee award, which also could be picked up by the association in collaboration with the GJAE. Choosing criteria for awarding these prices wisely and making them transparent may offer another piece to making community members more aware of some remedies for the current replicability crisis. One could additionally consider awards to authors for outstanding transparency and excellent communication regarding data and statistical analysis.

To allow for better statistical inference from a sample to a population, researchers should be put in a position where they can draw random samples from a population. Often this is not a simple task, especially if farmers are involved. Making registry data more widely and more openly available for research purposes would be an important task for the future. Alternatively, a farmer panel, similar to the socio-economic panel, could be maintained as a critical research infrastructure in the GEWISOLA field. The existing FADN could form a base. The FADN often remains limited when it comes to behavioural or specific land cover and management questions, particularly hindering evaluation of policy impact on sustainability other than the socio-economic pillar. Thus, enrichments of this data set by other existing data, for instance from the Integrated Administrative Control System (IACS), surveys and by information about selection of the farms into the sample would be helpful.

Support by the community for revising the teaching curricula and methods could be to establish a central pool of teaching examples for experiential learning and assignments in the domain of the community. Associations such as the GEWISOLA or the EAAE could provide the infrastructure and incentivize investments of teachers into such modules to foster sharing materials that offer clear guidance on good scientific practices, including hypothesis testing and mindful statistics, transparency in data, code, and writing. Replication studies could be incentivized also for teaching purposes by journals and publishers to overall foster a longer-term change of the social norms governing our practices.

The ideas mentioned here are certainly not exhaustive and may be complemented as we go along this process of change. Perhaps it would be helpful to have one agenda element on the issue of statistical and/or scientific practice in each annual meeting of the GEWISOLA in the coming years, actively solicited by those responsible for the program and nudged by the association. They can have different formats – presentation on current developments, workshop, organized session, best practice updates – depending on what currently concerns the members or more generally the scientific community. Perhaps a future stronger liaison between the GJAE and the association can help to identify a person responsible to keep this on the agenda.

References

AAEA (Agricultural & Applied Association) (2021): Call for Papers for a Special Issue on 'Replications in Agricultural Economics' in Applied Economic Perspectives and Policy. http://blog.aaea.org/2020/09/call-for-papers-for-special-issue-on.html. Call: 28.10.2021.

ACZEL, B., R. HOEKSTRA, A. GELMAN, E.-J. WAGENMAKERS, I.G. KLUGKIST, J.N. ROUDER, J. VANDEKERCKHOVE, M.D. LEE, R.D. MOREY, W. VANPAEMEL, Z. DIENES and D. VAN RAVENZWAAIJ (2020): Discussion points for Bayesian inference. In: Nature Human Behaviour 4 (6): 561-563.

ALBERS, C. (2019): The problem with unadjusted multiple and sequential statistical testing. In: Nature Communications 10 (1): 1921.

ALTMAN, D.G. and J.M. BLAND (1995): Absence of evidence is not evidence of absence. In: BMJ 311 (7003): 485.

AMRHEIN, V., S. GREENLAND and B. MCSHANE (2019): Scientists rise up against statistical significance. In: Nature 567 (7748): 305-307.

AMRHEIN, V., F. KORNER-NIEVERGELT and T. ROTH (2017): The earth is flat (p 0.05): significance thresholds and the crisis of unreplicable research. In: PeerJ 5: e3544.

ANGRIST, J. and J.-S. PISCHKE (2008): Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton.

ARPINON, T. and R. ESPINOSA (2022): A Practical Guide to Registered Reports for Economists. Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=

ASKAROV, Z., A. DOUCOULIAGOS, H. DOUCOULIAGOS and T.D. STANLEY (2022): The Significance of Data-Sharing Policy. In: Journal of the European Economic Association.

BAKKER, M., C.L.S. VELDKAMP, M.A.L.M. VAN ASSEN, E.A.V. CROMPVOETS, H.H. ONG, B.A. NOSEK, C.K. SODERBERG, D. MELLOR and J.M. WICHERTS (2020): Ensuring the quality and specificity of preregistrations. In: PLOS Biology 18 (12): e3000937.

BANERJEE, A., E. DUFLO, A. FINKELSTEIN, L. KATZ, B. OLKEN and A. SAUTMANN (2020): In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics. Working Paper 26993. National Bureau of Economic Researc.

BARKASZI, L., S. KESZTHELYI, E. K. CSATÁRI and C. PESTI (2009): FADN Accountancy Framework and Cost Definitions. FACEPA Deliverable No. D1.1.1 - July 2009. In: http://facepa.slu.se/documents/Deliverable_D1-1-1_LEI.pdf. Call: 9.3.2020.

BARREIRO-HURLÉ, J. (2021): Spanish Journal of Agricultural Research Editorial Policy Update: Pre-registration of submissions based on primary data. In: Spanish Journal of Agricultural Research 19 (4): e01105.

BASTARDI, A., E. L. UHLMANN and L. ROSS (2011): Wishful thinking: belief, desire, and the motivated evaluation of scientific evidence. In: Psychological science 22 (6): 731-732.

BEKKERMAN, A. (2015): The role of simulations in econometrics pedagogy. In: Wiley Interdisciplinary Reviews: Computational Statistics 7 (2): 160-165.

BENDTSEN, M. (2018): A Gentle Introduction to the Comparison Between Null Hypothesis Testing and Bayesian Analysis: Reanalysis of Two Randomized Controlled Trials. In: Journal of Medical Internet Research 20 (10): e10873.

BENJAMINI, Y. (2016): It's not the p-values' fault. In: The American Statistician, Online Discussion 70: 1-2.

BINGHAM, E., J.P. CHEN, M. JANKOWIAK, F. OBERMEYER, N. PRADHAN, T. KARALETSOS, R. SINGH, P. SZERLIP, P. HORSFALL and N.D. GOODMAN (2019): Pyro: Deep Universal Probabilistic Programming. In: Journal of Machine Learning Research 20 (1): 973-978.

BLANCO-PEREZ, C. and A. BRODEUR (2020): Publication Bias and Editorial Statement on Negative Findings. In: The Economic Journal 130 (629): 1226-1247.

BRODEUR, A., N. COOK and A. HEYES (2020): Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. In: American Economic Review 110 (11): 3634-3660.

BRODEUR, A., M. LÉ, M. SANGNIER and Y. ZYLBERBERG (2016): Star Wars: The Empirics Strike Back. In: American Economic Journal: Applied Economics 8 (1): 1-32.

BRUNS, S.B., I. ASANOV, R. BODE, M. DUNGER, C. FUNK, S.M. HASSAN, J. HAUSCHILDT, D. HEINISCH, K. KEMPA, J. KÖNIG, J. LIPS, M. VERBECK, E. WOLFSCHÜTZ and G. BUENSTORF (2019): Reporting errors and biases

in published empirical findings: Evidence from innovation research. In: Research Policy 48 (9): 103796.

BRUNS, S.B. and M. KALTHAUS (2020): Flexibility in the selection of patent counts: Implications for p-hacking and evidence-based policymaking. In: Research Policy 49 (1): 103877.

BUCK, S. (2021): Beware performative reproducibility. In: Nature 595 (7866): 151.

BUTTON, K.S., J.P.A. IOANNIDIS, C. MOKRYSZ, B.A. NOSEK, J. FLINT, E.S.J. ROBINSON and M.R. MUNAFÒ (2013): Power failure: why small sample size undermines the reliability of neuroscience. In: Nature Reviews Neuroscience 14 (5): 365-376.

CAMERER, C.F., A. DREBER, E. FORSELL, T.-H. HO, J. HUBER, M. JOHANNESSON, M. KIRCHLER, J. ALMENBERG, A. ALTMEJD, T. CHAN, E. HEIKENSTEN, F. HOLZMEISTER, T. IMAI, S. ISAKSSON, G. NAVE, T. PFEIFFER, M. RAZEN and H. WU (2016): Evaluating replicability of laboratory experiments in economics. In: Science 351 (6280): 1433-1436.

CHRISTENSEN, G., J. FREESE and E. MIGUEL (2019): Transparent and Reproducible Social Science Research. How to Do Open Science. University of California Press, Berkeley, California.

CHRISTENSEN, G. and E. MIGUEL (2018): Transparency, Reproducibility, and the Credibility of Economics Research. In: Journal of Economic Literature 56 (3): 920-980.

CLEMENS, M.A. (2017): The meaning of failed replications: A review and proposal. In: Journal of Economic Surveys 31 (1): 326-342.

COLQUHOUN, D. (2014): An investigation of the false discovery rate and the misinterpretation of p-values. In: Royal Society Open Science 1 (3): 140216.

DUMOUCHEL, W. and G.J. DUNCAN (1983): Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. In: Journal of the American Statistical Association 78 (383): 535-543.

ELLIOTT, M.R. and R. VALLIANT (2017): Inference for Nonprobability Samples. In: Statistical Science 32 (2): 249-264.

FAUL, F., E. ERDFELDER, A.-G. LANG and A. BUCHNER (2007): G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. In: Behavior Research Methods 39 (2): 175-191.

FERRARO, P. J. and P. SHUKLA (2020): Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics? In: Review of Environmental Economics and Policy 14 (2): 339-351.

FERRARO, P.J. and P. SHUKLA (2022): Credibility crisis in agricultural economics. In: Applied Economic Perspectives and Policy.

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

FRICKER, R.D., K. BURKE, X. HAN and W.H. WOODALL (2019): Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p -Value Ban. In: The American Statistician 73 (sup1): 374-384.

GELMAN, A. (2016): The Problems With P-Values are not Just With P-Values. In: The American Statistician, Online Discussion.

GELMAN, A. and J. CARLIN (2017): Some Natural Solutions to the p -Value Communication Problem—and Why They Won't Work. In: Journal of the American Statistical Association 112 (519): 899-901.

GEWEKE, J., G. KOOP and H. VAN DIJK (2011): Introduction. In: : The Oxford Handbook of Bayesian Econometrics: 1-8.

GIGERENZER, G. (2004): Mindless statistics. In: The Journal of Socio-Economics 33 (5): 587-606.

GIGERENZER, G. (2018): Statistical Rituals: The Replication Delusion and How We Got There. In: Advances in Methods and Practices in Psychological Science 1 (2): 198-218.

GIOFRÈ, D., G. CUMMING, L. FRESC, I. BOEDKER and P. TRESSOLDI (2017): The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. In: PLOS ONE 12 (4): e0175583.

GOODMAN, S.N. (2001): Of P-Values and Bayes: A Modest Proposal. In: Epidemiology 12 (3): 295.

GRANT, M.J. and A. BOOTH (2009): A typology of reviews: an analysis of 14 review types and associated methodologies. In: Health information and libraries journal 26 (2): 91-108.

GREENLAND, S. (2019): Valid P -Values Behave Exactly as They Should: Some Misleading Criticisms of P -Values and Their Resolution With S -Values. In: The American Statistician 73 (sup1): 106-114.

GREENLAND, S. (2020): The causal foundations of applied probability and statistics. https://arxiv.org/pdf/2011.02677.

GREENLAND, S., S.J. SENN, K.J. ROTHMAN, J.B. CARLIN, C. POOLE, S.N. GOODMAN and D.G. ALTMAN (2016): Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. In: European Journal of Epidemiology 31 (4): 337-350.

HARVEY, C. (2017): Presidential Address: The Scientific Outlook in Financial Economics. In: The Journal of Finance 72 (4): 1399-1440.

HAVEN, T.L. and L. VAN GROOTEL (2019): Preregistering qualitative research. In: Accountability in Research 26 (3): 229-244.

HECKELEI, T., S. HÜTTEL, M. ODENING and J. ROMMEL (2023): Replication Data for: The p-value debate and statistical (mal)practice - implications for the agricultural and food economics community. GRO.data.

HIRSCHAUER, N. (2021): The debate on p-values and statistical inference: What are the consequences for our community? Problems and solutions in statistical practice. GEWISOLA 2021 Pre-Conference Workshop.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF and C. BECKER (2021): A Primer on p-Value Thresholds and α-Levels - Two Different Kettles of Fish. In: German Journal of Agricultural Economics 70 (2): 123-133.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF, C. BECKER and A. JANTSCH (2019): Can p-values be meaningfully interpreted without random sampling? In: Statistics Surveys 14: 71-91.

HIRSCHAUER, N., G. SVEN, O. MUSSHOFF, F. ULRICH, T. INSA and W. PETER (2016): Die Interpretation des p-Wertes - Grundsätzliche Missverständnisse. In: Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik) 236 (5): 557-575.

HUNTINGTON-KLEIN, N., A. ARENAS, E. BEAM, M. BERTONI, J.R. BLOEM, P. BURLI, N. CHEN, P. GRIECO, G. EKPE,

T. PUGATCH, M. SAAVEDRA and Y. STOPNITZKY (2021): The influence of hidden researcher decisions in applied microeconomics. In: Economic Inquiry 59 (3): 944-960.

IMBENS, G.W. (2021): Statistical Significance, p -Values, and the Reporting of Uncertainty. In: Journal of Economic Perspectives 35 (3): 157-174.

IOANNIDIS, J.P.A., T.D. STANLEY and H. DOUCOULIAGOS (2017): The Power of Bias in Economics Research. In: The Economic Journal 127 (605): F236-F265.

IONIDES, E.L., A. GIESSING, Y. RITOV and S.E. PAGE (2017): Response to the ASA's Statement on p -Values: Context, Process, and Purpose. In: The American Statistician 71 (1): 88-89.

KANG, H. (2021): Sample size determination and power analysis using the G*Power software. In: Journal of Educational Evaluation for Health Professions 18.

KRANZ, S. and P. PÜTZ (2022): Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment. In: American Economic Review 112 (9): 3124-3136.

KRUEGER, J.I. and P.R. HECK (2019): Putting the P -Value in its Place. In: The American Statistician 73 (sup1): 122-128.

LEMKEN, D. (2021): The price penalty for red meat substitutes in popular dishes and the diversity in substitution. In: PLOS ONE 16 (6): e0252675.

LOGG, J.M. and C.A. DORISON (2021): Pre-registration: Weighing costs and benefits for researchers. In: Organizational Behavior and Human Decision Processes 167: 18-27.

LOKEN, E. and A. GELMAN (2017): Measurement error and the replication crisis. In: Science 355 (6325): 584-585.

MARGARIAN, A. (2022): Beyond P-Value-Obsession: When are Statistical Hypothesis Tests Required and Appropriate? In: German Journal of Agricultural Economics 71 (4): 213-226.

MCCLOSKEY, D.N. and S.T. ZILIAK (1996): The Standard Error of Regressions. In: Journal of Economic Literature 34 (1): 97-114.

MERVIS, J. (2014): Research Transparency. Why null results rarely see the light of day. In: Science 345 (6200): 992.

NEUENFELDT, S. and A. GOCHT (2014): Integrating Econometric and Mathematical Programming Models into an Amendable A Handbook on the use of FADN Database in Programming Models. Thünen Working Paper No. 35. https://literatur.thuenen.de/digbib_extern/dn054328.pdf.

NEYMAN, J. and E.S. PEARSON (1933): On the problem of the most efficient tests of statistical hypotheses. In: Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231 (694-706): 289-337.

O'BOYLE, E.H., G.C. BANKS and E. GONZALEZ-MULÉ (2017): The Chrysalis Effect. In: Journal of Management 43 (2): 376-399.

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

OLKEN, B.A. (2015): Promises and Perils of Pre-Analysis Plans. In: Journal of Economic Perspectives 29 (3): 61-80.

PÜTZ, P. and S.B. BRUNS (2021): The (Non-)Significance of Reporting Errors In Economics: Evidence from Three Top Journals. In: Journal of Economic Surveys 35 (1): 348-373.

RAHWAN, Z., E. YOELI and B. FASOLO (2019): Heterogeneity in banker culture and its influence on dishonesty. In: Nature 575 (7782): 345-349.

ROMANO, J.P., A.M. SHAIKH and M. WOLF (2010): Multiple Testing. In: The New Palgrave Dictionary of Economics 4.

ROMMEL, J. and M. WELTIN (2021): Is There a Cult of Statistical Significance in Agricultural Economics? In: Applied Economic Perspectives and Policy 43 (3): 1176-1191.

SCHOOLER, J.W. (2014): Metascience could rescue the 'replication crisis'. In: Nature 515 (7525): 9.

SERRA-GARCIA, M. and U. GNEEZY (2021): Nonreplicable publications are cited more than replicable ones. In: Science Advances 7 (21).

SMITH, T.M.F. (1983): On the Validity of Inferences from Non-random Sample. In: Journal of the Royal Statistical Society. Series A (General) 146 (4): 394.

STEEGEN, S., F. TUERLINCKX, A. GELMAN and W. VANPAEMEL (2016): Increasing Transparency Through a Multiverse Analysis. In: Perspectives on psychological science : a journal of the Association for Psychological Science 11 (5): 702-712.

VAN DE MEENT, J.-W., B. PAIGE, H. YANG and F. WOOD (2018): An Introduction to Probabilistic Programming. In: https://arxiv.org/pdf/1809.10756.

VERHULST, B. (2016): In Defense of P Values. In: AANA journal 84 (5): 305-308.

WASSERSTEIN, R.L. and N.A. LAZAR (2016): The ASA Statement on p -Values: Context, Process, and Purpose. In: The American Statistician 70 (2): 129-133.

WASSERSTEIN, R.L., A.L. SCHIRM and N.A. LAZAR (2019): Moving to a World Beyond " $p < 0.05$". In: The American Statistician 73 (sup1): 1-19.

WEHRDEN, H. von, J. SCHULTNER and D.J. ABSON (2015): A call for statistical editors in ecology. In: Trends in Ecology and Evolution 30 (6): 293-294.

YOUNG, C. and K. HOLSTEEN (2017): Model Uncertainty and Robustness. In: Sociological Methods & Research 46 (1): 3-40.

ZILIAK, S. and D. MCCLOSKEY (2008): The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor, MI.

ZILIAK, S.T. and D.N. MCCLOSKEY (2004): Size matters: the standard error of regressions in the American Economic Review. In: The Journal of Socio-Economics 33 (5): 527-546.

Contact author:
PROF. DR. THOMAS HECKELEI
University of Bonn
Institute for Food and Resource Economics
Nußallee 21, 53115 Bonn
e-mail: thomas.heckelei@ilr.uni-bonn.de