Why Too Many Political Science Findings Cannot be Trusted and What We Can Do About it: A Review of Meta-scientific Research and a Call for Institutional Reform

Alexander Wuttke, University of Mannheim

2019, Politische Vierteljahresschrift / German Political Science Quarterly 1, 60: 1-22, DOI: 10.1007/s11615-018-0131-7.
This is an **uncorrected pre-print**. Please cite the original article.

Witnessing the ongoing "credibility revolutions" in other disciplines, also political science should engage in meta-scientific introspection. Theoretically, this commentary describes why scientists in academia's current incentive system work against their self-interest if they prioritize research credibility. Empirically, a comprehensive review of meta-scientific research with a focus on quantitative political science demonstrates that threats to the credibility of political science findings are systematic and real. Yet, the review also shows the discipline's recent progress toward more credible research. The commentary proposes specific institutional changes to better align individual researcher rationality with the collective good of verifiable, robust, and valid scientific results.

Open Science; publication bias; replication crisis; replicability; transparency; meta-science

Forschung als soziales Dilemma: Eine meta-wissenschaftliche Bestandsaufnahme zur Glaubwürdigkeit politikwissenschaftlicher Befunde und ein Appell zur Veränderung akademischer Anreizstrukturen

Angesichts der "Glaubwürdigkeitsrevolution" in anderen Sozialwissenschaften liegen Fragen nach der Verlässlichkeit institutioneller Wissensproduktion auch in der Politikwissenschaft nahe. Dieser Kommentar beschreibt warum Wissenschaftler entgegen ihrem Eigeninteresse handeln, wenn sie Forschungsvalidität priorisieren. Ein umfassender Überblick der meta-wissenschaftlichen Literatur mit Fokus auf die quantitative weist einerseits auf jüngst eingeleitete Reformen zur Sicherung reliabler Forschung hin. Andererseits offenbar dieser Überblicksartikel systematische Probleme in der Glaubwürdigkeit veröffentlichter Forschungsbefunde. Dieser Kommentar schlägt konkrete Maßnahmen vor individuelle Forscheranreize in Einklang zu bringen mit dem gemeinschaftlichen Ziel verlässlicher Forschung.

Offene Wissenschaft; Replikationskrise; Transparenz; Replizierbarkeit, Reproduzierbarkeit;

Introduction: Open Science – A Contemporary Movement toward Scientific Self-Reflection

One of science's great virtues is its self-correcting ability. Reflecting upon scientific practices is conducive and inherent to the scientific enterprise. Currently, a new stream of meta-scientific debates is taking roots in many research disciplines, loosely connected by the umbrella term "Open Science". Open Science refers to a meta-scientific, inter-disciplinary reform movement, which raises epistemological, methodological, bibliographical, and practical concerns about the contemporary mode of knowledge production and advocates openness as a process and outcome goal of social inquiry. Open Science discusses the implications of various social, economic, and technological changes for the academic system. Since these changes are so broad, the Open Science movement encompasses vastly different ideas.[1] In many instances, Open Science scholars point to legacy gaps in academia, in which the traditions and conventions of institutional knowledge production reflect earlier historical conditions. Hence, Open Science advocates examine how to harness the potential of new tools and ideas for the scientific enterprise. In other words, Open Science's common denominator is the conviction that the current academic system has flaws and inefficiencies that hinder the construction and dissemination of knowledge. The movement therefore seeks to improve contemporary means of institutional knowledge creation.

One line of thought within the Open Science movement concerns the output of the scientific process, the epistemological status of academic research findings. An influential wake-up call was a paper by epidemiologist John Ioannidis, which argued that "most published research findings are false" (Ioannidis 2005). While seeming outrageous at the time, a growing

[1] Different lines of thought in the Open Science movement comprise "the infrastructure school (which is concerned with the technological architecture), the public school (which is concerned with the accessibility of knowledge creation), the measurement school (which is concerned with alternative impact measurement), the democratic school (which is concerned with access to knowledge) and the pragmatic school (which is concerned with collaborative research)" Fecher/Friesike (2014: 17).

body of meta-scientific research in the behavioral and social sciences has substantiated this claim ever since (for historical overviews, see Elman et al. 2018; Freese/Peterson 2017). Psychology in particular has experienced a lasting shake-up of disciplinary self-confidence. After a systematic assessment of the evidential value of the discipline's findings, psychological science has undergone a rocky but deeply transformative phase, which has turned the discipline's production and evaluation of scientific findings upside down (Nelson et al. 2018). Within a few years, psychology underwent fundamental changes in hiring practices (Schönbrodt/Mellor 2018), publishing formats (Hardwicke/Ioannidis 2018; Nosek et al. 2018), and statistical practices (Simmons et al. 2011; Motyl et al. 2017), and has rewritten standards for scientific publication (Nosek et al. 2015). Hence, after "psychology's renaissance" (Nelson et al. 2018) and recent "credibility revolutions" (Angrist/Pischke 2010) in other disciplines, one question naturally arises: Is political science more immune to the factors that undermined the credibility of other scientific disciplines or is the partial inertia in our field due to the fact that we have not yet received the news? This review shows that both perspectives have merits: Political science actually performs better than others disciplines in various criteria of research credibility, which is in part due to institutional reforms that were already undertaken. At the same time, however, meta-scientific evidence shows that severe systematic disincentives undermine the credibility of the discipline's evidence base to an extent that warrants closer attention of the members of this academic community.

In light of the ongoing change toward greater openness and trustworthiness around and within political science, this commentary examines the state of political science with respect to credibility of its quantitative empirical findings. Specifically, this article has three goals: First, it analyzes the relationship between academia's contemporary reward structures and the credibility of research findings. Employing social dilemmas as a guiding concept (Scheliga/Friesike 2014), this article identifies an incentive gap between the common values science seeks to achieve and the incentive system that penalizes the individual scientist who

prioritizes research credibility. Second, assessing whether and to what extent these disincentives undermine the trustworthiness of political science research, this article is the first to conduct a comprehensive review of the meta-scientific literature on the evidential value of contemporary findings in quantitative political science. Third, learning from the transformation in other disciplines, this article discusses options for improving the credibility of political science findings. Altogether, by reviewing the past few years of disciplinary self-reflection and by proposing measures the discipline might take in the next few years, this commentary hopes to serve as a primer for contemporary debates on research credibility in quantitative political science.[2]

Assessing the Credibility of Quantitative Political Science Findings

As authors, reviewers or readers, political scientists frequently judge the credibility of single studies. That is, scholars assess the "extent to which the research methods and data underlying findings can be considered reliable and valid representations of reality" (Cook et al. 2018: 2). Yet, even though single studies are regularly subject to credibility evaluations, political science has only recently begun to assess the credibility of a larger body of studies. Since these efforts are in their infancy, the discipline lacks criteria for systematic credibility evaluations. Thinking about the feasibility and the potential content of such criteria raises fundamental epistemological questions about what makes scientific studies more or less credible.

Here, political science can learn from psychology, in particular from LeBel et al. (2018) who recently proposed a meta-scientific framework for the quantification of research credibility. Despite the fact that their framework does not fully acknowledge the specific

[2] Even if similar discussions are gaining traction in other research cultures Kern/Gleditsch (2017) (Kern/Gleditsch 2017; Monroe 2018; Elman et al. 2018; Janz 2018), this commentary focuses on quantitative political science as published in English-language peer-reviewed journals, which has attracted most meta-scientific attention in recent years. Although it is a debate worth having, it is beyond the scope of this commentary to discuss how the evidence and arguments presented here can be applied to other research cultures and publication formats in political science.

epistemological challenges of the social sciences, it serves as a useful starting point to get a sense of the credibility of political science findings, judged by criteria that are valued across disciplinary boundaries. Specifically, the framework departs from the premise that credible findings are those that repeatedly survived high-quality and risky attempts to prove them wrong. Thus, based on the basic principle of falsification and falsifiability, the framework suggests four dimensions for assessing research credibility.

The first criterion for research credibility refers to pre-producibility (Stark 2018) whereas the other three criteria demand different forms of re-producibility (or replicability). Specifically, the first criterion assesses whether studies make themselves accessible to falsification attempts by providing *method and data transparency*. The second criterion entails the most basic form of empirical falsification, *analytical reproducibility*, i.e. the verification of results by repeating the same data processing and statistical analyses on the original data.[3] The third criterion, *analytical robustness*, also presupposes the use of old data but assesses results' sensitivity to different data-processing and data-analytic decisions. *Effect replicability,* finally, refers to what Freese/Peterson (2017) call repeatability and generalizability: whether an effect is observed when the analysis is repeated with new data.[4]

This framework will guide the following investigation on the credibility of the current body of published political science findings. These criteria do not suffice for definite judgements on a study's validity, because each assessment of the truth value of a study would require in-depth examinations of a study's research design. However, considering that valid knowledge is evidence-based and rule-bound (Elman et al. 2018), failing these criteria casts doubts on whether the reported findings were properly derived from the evidence and in accordance with the academic community's rules of institutional knowledge productions. In

[3] Freese/Peterson (2017) call this type of replication verifiability.
[4] These criteria can be ordered hierarchically in the sense that the latter are more likely fulfilled when the former are met.

other words, all else equal, we have more reasons to trust studies that withstood repeated falsification attempts, and we should be worried if political science studies systematically failed these criteria of credible research findings.

The remainder of the article explains how each criterion constitutes a collective good of the scientific enterprise and how the incentive structure in the academic system is either conducive or detrimental to the individual rationality of researchers to work toward its attainment. These analytical discussions are intertwined with a review of meta-scientific evidence about the credibility of political science's current evidence base.

Method and Data Transparency

Regardless of the individual research methods, "all evidence-based approaches to social inquiry in political science have a set of common characteristics that allow them to benefit from transparency" (Elman et al. 2018: 31). This is because members of research communities share common beliefs and norms about the adequate production of valid knowledge claims. Therefore, in line with the tenet that science is about "show me", not "trust me" (cf. Stark 2018), data and production transparency enable the mutual reassurance or validation that these rules were complied with. According to guidelines of the American Political Science Association, thus, "researchers have an ethical obligation to facilitate the evaluation of their evidence based knowledge claims through data access, production transparency, and analytic transparency" (Lupia/Elman 2014: 21). Moreover, data transparency in particular enables the accumulation of knowledge by building upon prior work (secondary data analysis). Consequently, it is easy to see how open data and open materials represent a collective good that may benefit all members of the scientific community and the scientific enterprise itself.

However, despite data and production transparency being valuable goals, it is a costly good to produce from the perspective of the individual researcher because it requires the investment of resources. These costs comprise the researcher's time but also potential losses

researchers might fear as by-products of transparency (being scooped, having one's errors exposed, etc.; see Washburn et al. 2018; Tenopir et al. 2011).

Because collective and individual rationality do not align, it is no surprise that many political science studies fall short of meeting the method and data transparency criterion. First, studies rarely provide the information necessary to repeat or fully assess a study. Humphreys (2018) highlighted the difficulty he and coauthors had in declaring designs for articles published in major political science journals. Very commonly, studies did not provide sufficient information to conduct a comprehensive diagnosis of the study's designs. In addition to insufficient method transparency, many studies to not provide data transparency either. When the publishing journal does not require the provision of data and code, a majority of studies in the discipline's top outlets do not provide such materials to the community (Key 2016; Stockemer et al. 2018). What is more, even when researchers make efforts to provide materials, they are often practically unavailable because three or four out of 10 links to the data are broken—even in very recently published studies (Key 2016; Gertler/Bullock 2017). However, when study authors are approached directly to share the underlying materials, about one out of two researchers do make the data available (Stockemer et al. 2018).[5] In line with the idea of data transparency as a social dilemma, these numbers show that political scientists are more inclined to share data and code when asked but that they do not invest many resources in transparency when not asked. Considering the fierce competition researchers face in their academic career, under the current incentive structure researchers may be well advised to spend their time on publishing yet another study than on publishing the material for their previous article.

[5] Note that making one's work accessible to inter-subjective assessment goes beyond data transparency and includes disclosure of data-analytical and processing procedures. Stockemer et al. (2018) discuss cases in which attempts to replicate prior results failed because neither the syntax nor the published article provided sufficient information to repeat the authors' analytical steps.

Luckily, having studied for decades how to solve collective action problems, political science should be in a good situation to overcome the ongoing transparency problems in its discipline. Indeed, few social science disciplines have been more dynamic in creating incentives, rules, and infrastructures for the promotion of research transparency than political science in recent years (Elman et al. 2018). Political scientists have helped establish data repositories, thereby lowering the cost for researchers to easily store their material. Others have participated in the PRO initiative[6], in which reviewers call for data transparency during the review process, thereby incentivizing researchers to comply with transparency standards and putting pressure on editors and journals to adapt data-sharing policies. Most influential, however, were the discussions on DA-RT (Lupia/Elman 2014; Monroe 2018; Janz 2018). As a consequence of these intra-disciplinary debates on research transparency, awareness of these issues has greatly increased: Within only a few years, transparency-related policies have greatly expanded. Among the first journals to require data-sharing, *Politische Vierteljahresschrift* adopted transparency policies years ago (Ghergina/Katsanidou 2013: 341). Other journals have followed suit and, by now, the majority of prominent journals in the discipline has instated transparency requirements (Stockemer et al. 2018: 2; Ghergina/Katsanidou 2013), with many journals demanding scholars to publish full reproduction materials (Key 2016).

Accordingly, political science has recently made significant steps to align collective goods with individual researcher rationality by making access to a private good (publication in leading journals) dependent on its contribution to the common good of method and data transparency. Yet, even though prominent journals acted as vanguards in promoting transparency within the discipline, we must acknowledge that most less-known political science journals have not yet established binding transparency policies and that, among those who did, the quality of transparency varies tremendously (Key 2016). Still, albeit not yet standard

[6] See https://opennessinitiative.org/

practice, data and method transparency are about to become commonplace. Therefore, the broad attention to the new transparency rules in leading journals may facilitate adjustments of descriptive and injunctive norms within the community, thereby creating second-order effects that help overcome the social dilemma of transparent political science research.

Analytical Reproducibility

Analytical reproducibility is a necessary requirement for the credibility of empirical studies because reported findings that do not follow properly from the study's data cannot be regarded as valid knowledge claims (cf. Freese/Peterson 2017: 152). In other words, a study's credibility diminishes when the reported evidence is distorted by blunt mistakes or misspecifications in statistical procedures.

Although no researcher wants to publish flawed results, ensuring analytical reproducibility requires significant amounts of work and therefore presents a costly investment from a researcher's perspective. At the same time, considering contemporary academic culture and publishing practices, there is a good chance that a lack of analytical reproducibility will not be (widely) noticed. Despite the famous call for "Replication, Replication" (King 1995) more than 20 years ago, replications continue to be rare (Freese/Peterson 2017). Despite calls for a visible outlet for replication studies (Ishiyama 2014), such a replication journal does not exist yet.

It is telling that initially not a single article accepted for publication could be flawlessly verified when the American Journal of Political Science started to assess the analytical reproducibility of newly accepted articles in 2015.[7] In political science journals without verification policies, Stockemer et al. (2018) showed that one out of three studies could not be analytically reproduced even when the authors made data accessible. In some of the tests for

[7] See https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/

analytical reproducibility, the reproduced results deviated from the reported results. In other cases, reproducibility was impossible due to poor organization of data or code. Hence, for a significant portion of political science studies, we cannot be sure to obtain the same results if we re-ran the same analyses on the same data.

Presumably, all political scientists can agree on the goal of avoiding the publication of findings that evidently do not follow from the data. Hence, it is encouraging to observe that the repeated call for greater emphasis on reproducibility is now materializing in the discipline's actual practices. Recently, journals have established replication article formats (JEPS, JJPS), and even top journals begin to publish analytical reproductions of previous articles (APSR, see Cingranelli/Filippov 2018).

Actually, a large step toward the collective good of analytical reproducibility could be made by rather simple means. The American Journal of Political Science has shown that publication outlets can achieve a verifiability rate of 100% (Lenz/Sahn 2017: 3) as opposed to the afore-mentioned verifiability rate of one third of other journals (Stockemer et al. 2018). What is needed is to ensure adherence to transparency policies with dedicated verification procedures. Verifying analytical reproducibility requires staff and is therefore costly.[8] For several journals, this is not a suitable option. This is different, however, in the case of many prominent journals that are the official publication outlets of academic associations, for which licensing trademarks to publishers often carries significant revenues to the associations. In these cases, analytical reproducibility also becomes a question of prioritizing: How important is it for us as members of an academic community to ensure the evidential value of the findings reported in our journals and how much are we willing to invest in securing the credibility of these results?

[8] For details on the costs of AJPS's verification processes, see https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility

Analytical Robustness

Analytical robustness derives its importance from the fact that hundreds or thousands of analytical decisions are necessary before a finding can be reported in an academic study. Even though all of these decisions may affect the final estimate, the reader usually sees only a small subset of all potential outcomes. Hence, a credible study is a study in which the reported findings are a good representation of all reasonable analytical choices the researcher could make.

Explaining how structural factors in the academic system undermine analytical robustness requires a broader perspective that includes the reward system of contemporary academic publishing. Psychology's replication crisis illustrates how flawed incentive structures may fundamentally undermine a discipline's scientific credibility. After a thorough re-examination, psychological science recently needed to revise seminal textbook findings such as ego-depletion, which is now seen as lacking credible empirical support, while decades of previous research had seemingly provided repeated and concordant evidence of their presence (Vadillo et al. 2018; Nelson et al. 2018). The true population estimate and the average estimate reported in scientific literatures may diverge significantly when the published pile of evidence represents a biased collection of false–positive findings from an even larger body of unpublished studies whose true population estimate denotes the absence of the hypothesized effect. Psychology's replication crisis illustrates how flawed reward systems may lead to a body of literature that comprises dozens or even hundreds of studies that support the existence of an effect although the effect is truly absent or not meaningful in size.

One structural cause behind these biases is rooted in the specific public preferences of academic journals. When deciding whether to publish a manuscript, scientific outlets not only consider the merits of theory and research design but also have an eye on the study's findings, preferring studies with novel, positive, and clean results (Nosek et al. 2018). Systematic

preferences to publish studies with positive and clean findings also exist in political science, particularly among the discipline's leading journals (Gerber et al. 2001; Gerber et al. 2010). Here the hypothesis of no publication bias was rejected at a probability of 1 to 32 billion (Gerber 2008). What is more, the likelihood of publication increases when effect sizes are larger (Esarey/Wu 2016).[9] To quantify the extent of this bias, Franco et al. (2014) traced the fortunes of a clearly specified and known set of conducted research projects in political science and examined whether a conducted research project was more likely to be published when the study's outcome coincided with the preferences of academic journals. The authors showed that projects with positive findings were forty percentage points (or three times) more likely to be published and sixty percentage points more likely to be written up. In other words, when researchers test a hypothesis and the results do not provide confirming evidence, there is a good chance that the academic community will never know about these results.

Hence, there is ample evidence that study results affect whether the study is published, which is a fundamental problem in gauging the true population estimate from prior literature. Imagine one social phenomenon was examined in multiple studies but only the positive studies were published. In this case, only the positive studies would be accessible to the interested reader, only the positive studies would be accessible to the interested reader, but studies with negative or more ambiguous findings —which differ systematically from the accessible studies—would remain in the file-drawer. Publication bias, thus, narrows the reader's field of vision to a small fragment of the empirical reality, and—what is worse—the fragment we see differs from the portion we cannot see.

Examining the literature on democratic innovations, for example, reveals that less than five percent of published research reports on failed interventions and no top journal ever published studies about failures or unintended consequences in democratic innovations

[9] Esarey/Wu (2016) estimate that the true value of published relationship is on average 40% smaller than their published value.

(Spada/Ryan 2017). However, anecdotal evidence and basic human experience suggest that, in fact, not all past experiments on democratic innovations were successful. Hence, making judgements about the success of democratic innovations from the published literature alone presumably leads to wrong estimates about the effectiveness of such tools. Naturally, publication bias is not specific to the literature of democratic innovation, but the subfield is only one example of distortions in the political science literature more generally.

Unfortunately, the incongruence between published and unpublished findings is not the only and not the most problematic consequence of the journals' preferences for novel, clean, and significant results. The graver consequence is that publication biases constitute a social dilemma. Considering the importance of publications as a currency for scientific success, publication decisions based on study outcomes incentivize scientists to get a hold of the one element of the research process they actually should not control: the findings of a study.

Publication bias incentivizes researchers to deviate from the orderly process of conducting research that philosophers of science have proposed for ensuring credible research (cf. Chambers 2017). According to one stream of thought in the philosophy of science, research should follow the hypothetico-deductive model of knowledge creation, according to which research begins with the theoretical deduction of a hypothesis (Chambers 2017; Pearl/Mackenzie 2018). Next, the researcher designs a decisive empirical test for the hypothesis, the results of which are used to inform further theorizing by rejecting or revising the theory. In other words, the researcher makes a theoretical prediction about the world and then tests if the hypothesized phenomenon is observed in practice. Even though this description is overly simplistic and alternative ways for knowledge production exist, this idealized model describes the process that most studies in quantitative political science try to mimic, which is visible in the formal structure of most published articles. However, a scientific culture that privileges clean and positive findings provides incentives not to adhere to this model of

knowledge production, despite promising credible evidence, but instead to adhere to a model of publication production that promises academic careers.

According to the idealized model, researchers have clearly specified the empirical test before it is conducted. In practice, however, extensive analytical discretion allows researchers to run multiple tests. Hence, models and corresponding results may be selected according to the presumed likelihood of achieving publication success with the respective findings. For instance, modifying the analytical strategies until significant results appear (p-hacking) is a sound research strategy if it is transparently disclosed; otherwise, such research practices severely threaten the validity of any empirical study (see https://projects.fivethirtyeight.com/phacking/ for an interactive demonstration; for a simulation of the consequences of p-hacking see Humphreys et al. 2013). Simmons et al. (2011) showed that the false–positive error rate increases from the usual five percent to an astonishing 60 percent when analytical discretion is exploited on only three dimensions of researcher flexibility. Moreover, employing real data, the authors managed to provide seeming evidence for the evidently false hypothesis that participant age diminished after listening to "When I'm Sixty-Four" by the Beatles. Put differently, they showed that exploiting researcher degrees of freedom allows for the generation of seeming evidence even for the most implausible hypotheses.

Apart from fitting the data to the hypothesis, researchers may also fit the hypothesis to the data when striving for clean results. Hypothesizing after the results are known, HARKing (HARKing, Kerr 1998) reverses the order of scientific steps as originally proposed by the hypothetico-deductive model. HARKing is compatible with an exploratory research process. Yet, most quantitative political science studies purport adherence to a hypothesis-driven confirmatory research process, in which HARKing resembles the proverbial peasant who first

shoots the hole in the fence and then paints the bullseye around it (cf. Shweder/Fiske 1986: 6).[10] Hence, HARKing is yet another form of analytical flexibility that undermines the credibility of the research process, but it may serve as a gateway to produce clean and positive results.

A growing body of meta-scientific studies provides indicative evidence for the questionable use of such research practices in the political science literature. Examining the race discrimination literature, Zigerell (2017) demonstrated several cases of selective reporting in which alternative model specifications remained unreported when they led to different results. In a large-scale examination of 249 political science studies from a research competition for which data and research questions were known, Franco et al. (2015) showed that only one out of five published studies reported all experimental conditions and outcome variables. The average study left 0.5 experimental conditions and 3.1 experimental outcomes undisclosed. Hence, for a typical political science study, there is a good chance that the reader is unaware of the existence of evidence, which is relevant to the assessment of the phenomenon under investigation.

Studies that investigate researcher degrees of freedoms in model specification (Simonsohn et al. 2015; Rohrer et al. 2017; Montgomery/Nyhan 2010; Steegen et al. 2016; Abel Brodeur, Nikolai Cook, Anthony Heyes 2018) provide additional insights into the practice of analytical flexibility. For instance, Lenz/Sahn (2017) showed that political science studies using observational data frequently report bivariate relationships when they are statistically significant. Yet, when the associations of interest become significant only after the inclusion of additional covariates, bivariate statistics are rarely reported and usually only the significant multivariate model is disclosed.

[10] Note that HARKing and confirmatory analysis are perfectly reconcilable if the new data is collected before both analytical steps but not without the collection of new data: "Just as conspiracy theories are never falsified by the facts that they were designed to explain, a hypothesis that is developed on the basis of exploration of a data set is unlikely to be refuted by that same data. Thus, one always needs a fresh data set for testing one's hypothesis." Wagenmakers et al. (2012).

Because meta-scientific research on the credibility of political science findings is still rare, these studies do not provide conclusive evidence about intentions and causes of researchers' use of their analytical discretion. However, two pieces of evidence suggest a significant role of strategic publication considerations in the use of researcher degrees of freedom. First, the above-cited findings by Franco et al. (2014), according to which researchers consider the likelihood of publication success in the decision whether to write up results, suggest the existence of similar publication-oriented considerations in the decision how to write up certain findings. Second, meta-scientific evidence from other disciplines quantifies the extent to which analytical flexibility is exploited to maximize publication success. Researcher surveys reveal that a majority of scientists in other behavioral disciplines openly report past uses of questionable research practices (Fiedler/Schwarz 2016; Fox et al. 2018). Persuasive evidence on the misuse of analytical discretion also derives from clinical trials, which were recently required by law to pre-specify their analytical strategy before data collection. As a consequence of transparent pre-registration, analytical freedom diminished, and the rate of allegedly successful clinical trials decreased dramatically. The sudden drop in successful trial may have two reasons: Either drug researchers suddenly became much less successful in inventing new medicine after pre-registration was mandated or they suddenly became much less successful in making ineffective drugs appear to be effective (Kaplan/Irvin 2015).

Public pre-registration also allows tracing the outcome of clinical studies, showing for depression drugs that all trials with positive findings were published but almost no negative study made it into publications that would acknowledge the absence of the hypothesized effect. Instead, the majority of originally negative trials was either published with a new spin in narrative or 'became positive' by switching or omitting outcome variables (Vries et al. 2018). Hence, although we cannot assess if and to what extent researcher degrees of freedom are misused in political science, other disciplines show that the reported findings reflect the true

population estimate. However, they also reflect strategic researcher considerations to make a study visible and interesting.

The preceding discussion illustrated the scope of researcher degrees of freedom, showing how structural incentives invite misuses of analytical discretion. When reported findings do not resemble reasonable analytical choices of the researcher but instead are determined by other considerations than approximating the true population estimate then analytical robustness decreases. Hence, misuses of analytical flexibility undermines the credibility of scientific findings. Every attempt to confine the credibility-undermining potential of researcher degrees of freedom faces the challenge that analytical flexibility is inherent to the research process and ambiguous in nature. Its misuse cannot be prevented with abstract rules, because the distinction between reasonable and questionable uses of analytical flexibility often requires case-by-case, in-depths examinations. The adequate response to this challenge is making the research process more transparent for those who judge its outcomes. For instance, judging whether it is reasonable to gloss over an experiment's second experimental condition requires awareness of its existence. Yet, political science lacks behind in maintaining institutional incentives to enhance the transparency of the research process.

Political science can make a tremendous step toward analytical robustness if it follows other disciplines and starts to pre-register prospective studies (Nosek et al. 2018). Publishing research intentions and analytical procedures before collecting data has become a rapidly expanding practice in other behavioral sciences, in which the number of pre-registered studies doubles each year and now exceeds 20,000 pre-analysis plans published on one hosting repository alone (Hardwicke/Ioannidis 2018).[11] Pre-registration allows for adapting analytical decisions after observing the data, but deviations from the pre-analysis plans are reported transparently, thereby establishing a clear distinction between predictive confirmatory analyses

[11] https://www.theatlantic.com/science/archive/2018/08/scientists-can-collectively-sense-which-psychology-studies-are-weak/568630/, last accessed August 28, 2018.

and postdictive exploratory analyses (Nosek et al. 2018). Political scientists may learn from experiences in other disciplines and built on existing advice in order to apply pre-registrations even on observational (Burlig 2018), secondary (Weston et al. 2018a), and qualitative (Mellor et al. 2018) data.

Because pre-registration requires researchers to think clearly about research goals and strategies before the data is collected, pre-registration may foster meaningful theory-driven research. However, it purposefully also reduces researcher degrees of freedoms to polish their results. Hence, a wide-spread use of pre-registration may enhance scientific credibility. Nevertheless, but as long as publication bias exists, researchers' self-interest to retain control over the results remains, and this is at odds with the purpose of pre-registration. The most promising institutional reform to overcome this social dilemma is the adoption of pre-registration in combination with result-blind reviews (called registered reports, see www.cos.io/rr/). A steadily growing number of journals (currently 132) in the behavioral sciences has introduced new submission formats in which articles are reviewed and potentially accepted prior to data collection. Although slowly, the idea of pre-registration in combination with result-blind reviewing is gaining traction in political science. Some journals conducted trials with these article formats (CPS, Findley et al. 2016), and first journals started offering registered reports as a regular option for submissions (JJPS, Nature Human Behavior).

Pre-registration and result-blind review are not the panacea for all problems regarding analytical robustness. For instance, political scientists often analyze pre-existing large-scale datasets for which pre-registration is sometimes possible (Weston/Bakker 2018; Weston et al. 2018b; Burlig 2018; Nosek et al. 2018) but not always suitable. Since such datasets offer an incredible number of reasonable model specifications, in these cases it is particularly important to know that significant effects provide more informational value than "merely [demonstrating] that it is possible to find a specification that fits the author's favorite hypothesis" (Ho et al. 2007: 199). Methods that are less model-dependent (Ho et al. 2007; Wilcox 2017) or that transparently

report model-sensitivity (Montgomery/Nyhan 2010; Simonsohn et al. 2015; Steegen et al. 2016) present methodological solutions to provide readers with the necessary information to assess the analytical robustness of research findings in such cases.

Effect Replicability

Effect replicability is an important yardstick for assessing the credibility of findings. This is because findings are less likely to replicate with newly collected data if the original findings resulted from extensive p-hacking, HARKing, or from mistakes in data processing. Hence, all previously discussed structural problems feed into low rates of effect replicability.

Because political science investigates fluid social systems that are undergoing constant change, most political science findings are context-dependent and, thus, do not necessarily generalize across time and space (Shweder/Fiske 1986). Therefore, not each previous finding is expected to replicate if the study was conducted again under today's circumstances. Yet, replications of previous studies still are valuable tools to assess scientific credibility because theoretically sound studies clearly specify the boundary conditions for which authors expect an effect.

Unfortunately, political science currently does not invest much effort on theoretical discussions of boundary conditions or on empirical replicability tests. As discussed with regard to analytical reproducibility above, replications are a rare species in the political science ecosystem (Freese/Peterson 2017). The low appreciation for replication is lamentable because, without replications, it is impossible to know whether findings are either unconfirmed genuine discoveries or unchallenged fallacies (cf. Chambers 2017: 50). If conducted in a theoretically informed way, replications function as the scientific immune system that takes the principle of falsification seriously and eliminates untrue findings from the literature (cf. Chambers 2017: 46).

In recent years, replications have become more widespread, helping to answer the fundamental question to which extent previous findings have informational value in predicting the occurrence of phenomena in the contemporary context. The Social Science Replication Project (SSRP) recently conducted highly powered replication attempts of 21 influential social science experiments (Camerer et al. 2018). Although all of the original studies were conducted very recently and the replication design was developed with and approved by the original authors, only 13 replication studies found significant effects in the same direction as the original study. This replication rate is larger than the estimate obtained in similar replication efforts in psychology (Open Science Collaboration 2015) and similar to that in economics (Camerer et al. 2016).

While some failures to replicate may result from changes in the social or political context (Sparrow 2018), the SSRP also provides indicative evidence about the credibility of the original results. Among the non-replicated studies, there was essentially no evidence for the original findings, leading some original authors in conjecture of additional evidence to declare their loss of confidence in the relevance of the effects they originally described (Gervais/Norenzayan 2018). Moreover, the replicated effect sizes are on average only half as large as the original effect sizes. Importantly, chance alone cannot explain these differences, because the effect size discrepancies follow a clear pattern—that is they are consistently lower in the replications than in the original studies even among studies which successfully replicated. Finally, the SSRP corroborated previous evidence that a study's replication likelihood is associated with study-level characteristics (e.g., the original study's p-value), which partly explains why independent researchers in many cases correctly predicted which studies would replicate.

The insight that the likelihood of replicating a recent and prominent social science experiment is not much larger than the flip of a coin may shake prior beliefs about the predictive power of existing social scientific studies on future occurrences of social phenomena. Since the

social sciences are not used to large-scale replications, there is still a lot to learn about the proper assessment of such replication results. In this vein, the SSRP may trigger debates about how much the limited replicability is attributable to contextual changes or whether biases in the estimates of the original studies are reasonable causes. In order to better understand the former and to reduce the latter, the SSRP will hopefully be a stepping stone toward more frequent and systematic replication efforts of previous observational and experimental studies in the political sciences.

Table 1: The credibility of political science findings as a social dilemma

	Collective good	Disincentives	Status in political science	Fix
Method & data transparency	Mutual validation, data re-usage	Researcher resources, loss of competitive advantage	Many or most studies do not provide data and materials, Significant recent progress	Binding transparency policies, shift in community norms
Analytical reproducibility	Findings follow from the data	Researcher resources, low awareness	1 in 3 studies not verifiable; progress by avant-garde journals	Dedicated verification procedures
Analytical robustness	Findings resemble reasonable analytical choices; published studies resemble entire body of knowledge	Publication bias, incentivizes misuse of analytical flexibility	Strong publication bias: Positive findings three times more likely to be published; substantial underreporting: On average 0.5 experimental conditions undisclosed; potential misuse of analytical flexibility in observational studies	Study pre-registration before data collection to emphasize theory and design and to confine analytical flexibility; result-blind peer review to reduce publication bias
Effect replicability	Falsification	All of the above	1 in 3 experimental social science studies do not replicable	All of the above

Note: Status estimates refer to small and unrepresentative sets of studies. See text for more information.

Conclusion

Political science helps understand the intricacies of social life and informs politicians as well as citizens in their efforts to change it to the better. However, political science faces increasing scrutiny by the public and other stakeholders who question the discipline's capacity to meet these goals (Elman et al. 2018). Responding to such skepticism along the principles of the scientific enterprise means to respond with a sober evaluation of the practices and outcomes of academic knowledge creation, that is, to employ the method of social inquiry on ourselves.

Based on a framework to assess the credibility of scientific findings, this article reviewed the meta-scientific evidence with a focus on the quantitative political science literature. The main result of these meta-scientific inquiries is that a significant portion of examined studies do not meet one or several credibility criteria. Specifically, by not providing data and method transparency, many or most political science studies make themselves inaccessible to inter-subjective validity assessments, and—when put to a test—the empirical findings of many studies cannot be verified. Moreover, findings in published and unpublished research diverge strongly and systematically, and the published studies show evidence of substantial underreporting. Altogether, meta-scientific evidence indicates deficiencies in the credibility of political science studies and suggests that the body of published political science findings is not an unbiased representation of the entire evidence base (see Table 1 above for an overview).

Even though it remains an open question to which extent the published estimates deviate from true population estimates, this article illustrates how the academic reward system incentivizes novel and clean results at the expense of the findings' validity. Because "scientists are meant to be detectives searching for the truth rather than lawyers cherry-picking the evidence to fit an argument" (Chambers/Etchells 2018), we should overcome scientific cultures that require researchers to polish and re-interpret their findings until they fit the preferences of

academic publishing. Concrete institutional reforms (e.g., dedicated syntax verification procedures and pre-registered studies in combination with result-blind reviewing) could be implemented to overcome social dilemmas in which researchers hurt their own interests when prioritizing research credibility.

Although institutional changes are central to overcoming social dilemmas, the academic community is small enough for each individual to make a difference toward a more transparent, reproducible valid political science: The reviewer who does not consider the eye-catching results but instead asks for public reproduction material to evaluate the accuracy of the findings; the editor who provides journal space for replications of previous studies in the outlet; the conference discussant who praises the transparent disclosure of negative findings; the dean who gives a premium to job applicants who made their research transparent; the student who shares what she has learned about open science practices and the scientist who is willing to avoid questionable research practices many of us have employed in the past. Altogether, when top-down leadership occurs in concert with bottom-up enthusiasm (Gernsbacher 2018: 3), it is feasible to establish those norms and rules that are yet missing for a more open and reliable scientific culture.

Even if the intention of this commentary was to highlight the necessity and potential of intra-disciplinary reform, several notes of balance are warranted. First, when we report above that many studies do not meet certain credibility criteria, then it should also be noted that the remaining studies do meet the standards of scientific credibility. Second, the reviewed meta-scientific evidence relies on small and biased samples of studies. Hence, it is desirable that more meta-scientific research will be conducted to assess the credibility of the discipline's current evidence base. Third, in addition to incentive-induced threats to scientific credibility discussed in this commentary, other systemic shortcomings may undermine the validity of research findings, such as prevailing methodological misconceptions (Rinke/Schneider 2015). Fourth, as all political scientists operate within the prevailing norms and incentive structures of their

academic communities, pointing to lacks in credibility is not to criticize individual scientists but is a call for changing these norms: to align academic rewards with scientific values.

Literatur

Abel Brodeur, Nikolai Cook, Anthony Heyes. 2018. *Methods Matter: P-Hacking and Causal Inference in Economics.*

Angrist, Joshua D. und Jörn-Steffen Pischke. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives* 24: 3-30.

Burlig, Fiona. 2018. Improving transparency in observational social science research: A pre-analysis plan approach. *Economics Letters* 168: 56-60.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen und Hang Wu. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351: 1433-1436.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers und Hang Wu. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*. https://www.nature.com/articles/s41562-018-0399-z.pdf.

Chambers, Chris. 2017. *The seven deadly sins of psychology. A manifesto for reforming the culture of scientific practice*. Princeton, Oxford: Princeton University Press.

Chambers, Chris und Pete Etchells. 2018. *Open science is now the only way forward for psychology*. https://www.theguardian.com/science/head-quarters/2018/aug/23/open-science-is-now-the-only-way-forward-for-psychology. 31.08.2018.

Cingranelli, David und Mikhail Filippov. 2018. Are Human Rights Practices Improving? *American Political Science Review*: 1-7. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/972B4F471231AFC7908A4DA951020340/S0003055418000254a.pdf/div-class-title-are-human-rights-practices-improving-div.pdf.

Cook, Bryan G., John Wills Lloyd, David Mellor, Brian A. Nosek und William J. Therrien. 2018. Promoting Open Science to Increase the Trustworthiness of Evidence in Special Education. *Exceptional Children* 7: 001440291879313.

Elman, Colin, Diana Kapiszewski und Arthur Lupia. 2018. Transparent Social Inquiry: Implications for Political Science. *Annual Review of Political Science* 21: 29-47. https://www.annualreviews.org/doi/pdf/10.1146/annurev-polisci-091515-025429.

Esarey, Justin und Ahra Wu. 2016. Measuring the effects of publication bias in political science. *Research & Politics* 3: 205316801666585.

Fecher, Benedikt und Sascha Friesike. 2014. Open Science: One Term, Five Schools of Thought. In: Sascha Friesike und Sönke Bartling (Hrsg.), *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing.* s.l.: Springer, 17-47.

Fiedler, Klaus und Norbert Schwarz. 2016. Questionable Research Practices Revisited. *Social psychological and personality science* 7: 45-52. http://journals.sagepub.com/doi/pdf/10.1177/1948550615612150.

Findley, Michael G., Nathan M. Jensen, Edmund J. Malesky und Thomas B. Pepinsky. 2016. Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study. *Comparative Political Studies* 49: 1667-1703.

Fox, Nick, Nathan Honeycutt und Lee Jussim. 2018. *How Many Psychologists Use Questionable Research Practices? Estimating the Population Size of Current QRP Users*: PsyArXiv. https://psyarxiv.com/3v7hx/download?format=pdf.

Franco, Annie, Neil Malhotra und Gabor Simonovits. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science* 345: 1502-1505. http://science.sciencemag.org/content/345/6203/1502.full.pdf.

Franco, Annie, Neil Malhotra und Gabor Simonovits. 2015. Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results. *Political Analysis* 23: 306-312.

Freese, Jeremy und David Peterson. 2017. Replication in Social Science. *Annual Review of Sociology* 43: 147-165. https://www.annualreviews.org/doi/pdf/10.1146/annurev-soc-060116-053450.

Gerber, Alan. 2008. Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science* 3: 313-326.

Gerber, Alan S., Donald P. Green und David Nickerson. 2001. Testing for Publication Bias in Political Science. *Political Analysis* 9: 385-392. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/FF6FB46C4DA51606969E84A4E00B2215/S1047198700003910a.pdf/div-class-title-testing-for-publication-bias-in-political-science-div.pdf.

Gerber, Alan S., Neil Malhotra, Conor M. Dowling und David Doherty. 2010. Publication Bias in Two Political Behavior Literatures. *American Politics Research* 38: 591-613.

Gernsbacher, Morton A. 2018. Rewarding Research Transparency. *Trends in Cognitive Sciences*. http://www.sciencedirect.com/science/article/pii/S1364661318301621.

Gertler, Aaron L. und John G. Bullock. 2017. Reference Rot: An Emerging Threat to Transparency in Political Science. *PS: Political Science & Politics* 50: 166-171. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/54F56CFC2CBE05778130E40CABB2CCC5/S1049096516002353a.p

df/div-class-title-reference-rot-an-emerging-threat-to-transparency-in-political-science-div.pdf.

Gervais, Will M. und Ara Norenzayan. 2018. Analytic atheism revisited. *Nature Human Behaviour*. https://www.nature.com/articles/s41562-018-0426-0.pdf.

Ghergina, Sergiu und Alexia Katsanidou. 2013. Data Availability in Political Science Journals. *European Political Science* 12: 333-349. https://doi.org/10.1057/eps.2013.8.

Hardwicke, Tom E. und John P. A. Ioannidis. 2018. *Mapping the Universe of Registered Reports*: BITSS. https://osf.io/preprints/bitss/fzpcy/.

Ho, Daniel E., Kosuke Imai, Gary King und Elizabeth A. Stuart. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15: 199-236. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/4D7E6D07C9727F5A604E5C9FCCA2DD21/S1047198700006483a.pdf/div-class-title-matching-as-nonparametric-preprocessing-for-reducing-model-dependence-in-parametric-causal-inference-div.pdf.

Humphreys, Macartan. 2018. *Declare Design*. Mannheim.

Humphreys, Macartan, Sierra, Raul Sanchez de la und Peter van der Windt. 2013. Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration. *Political Analysis* 21: 1-20. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/BD935F7843BF07F338774DAB66E74E3C/S104719870001322Xa.pdf/div-class-title-fishing-commitment-and-communication-a-proposal-for-comprehensive-nonbinding-research-registration-div.pdf.

Ioannidis, John P. A. 2005. Why Most Published Research Findings Are False. *PLOS Medicine* 2: e124.

http://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.0020124&type=printable.

Ishiyama, John. 2014. Replication, Research Transparency, and Journal Publications: Individualism, Community Models, and the Future of Replication Studies. *PS: Political Science & Politics* 47: 78-83. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/71B35EBCE3E426A7B59ECD73A991FF48/S1049096513001765a.pdf/div-class-title-replication-research-transparency-and-journal-publications-individualism-community-models-and-the-future-of-replication-studies-div.pdf.

Janz, Nicole. 2018. *Replication and Transparency in Political Science - Did We Make Any Progress?* https://politicalsciencereplication.wordpress.com/2018/07/14/replication-and-transparency-in-political-science-did-we-make-any-progress/. 14.06.2018.

Kaplan, Robert M. und Veronica L. Irvin. 2015. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLOS ONE* 10: e0132382. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0132382&type=printable.

Kern, Florian G. und Kristian Skrede Gleditsch. 2017. *Exploring Pre-registration and Pre-analysis Plans for Qualitative Inference*: Unpublished.

Kerr, N. L. 1998. HARKing: hypothesizing after the results are known. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 2: 196-217.

Key, Ellen M. 2016. How Are We Doing? Data Access and Replication in Political Science. *PS: Political Science & Politics* 49: 268-272. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/8EE49C7FE45B46F20A50D9271BB3AAC4/S1049096516000184a.pdf/div-class-title-how-are-we-doing-data-access-and-replication-in-political-science-div.pdf.

King, Gary. 1995. Replication, Replication. *PS: Political Science & Politics* 28: 444-452. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/85C204B396C5060963589BDC1A8E7357/S1049096500057607a.pdf/div-class-title-replication-replication-div.pdf.

LeBel, Etienne P., Randy J. McCarthy, Brian D. Earp, Malte Elson und Wolf Vanpaemel. 2018. A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and Practices in Psychological Science* 276: 251524591878748.

Lenz, Gabriel und Alexander Sahn. 2017. *Achieving Statistical Significance with Covariates*: BITSS. https://osf.io/preprints/bitss/s42ba/download?format=pdf.

Lupia, Arthur und Colin Elman. 2014. Openness in Political Science: Data Access and Research Transparency. *PS: Political Science & Politics* 47: 19-42.

Mellor, David, Alexandra Hartman und Florian Kern. 2018. *Preregistration for Qualitative Research Template*: OSF. https://osf.io/j7ghv/.

Monroe, Kristen R. 2018. The Rush to Transparency: DA-RT and the Potential Dangers for Qualitative Research. *Perspectives on Politics* 16: 141-148. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/A3874B1072E02CFFC00780232943F1FC/S153759271700336Xa.pdf/div-class-title-the-rush-to-transparency-da-rt-and-the-potential-dangers-for-qualitative-research-div.pdf.

Montgomery, Jacob M. und Brendan Nyhan. 2010. Bayesian Model Averaging: Theoretical Developments and Practical Applications. *Political Analysis* 18: 245-270. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/3179D92A3C9353DE7E4674987C33FD28/S1047198700012432a.pdf/bayesian_model_averaging_theoretical_developments_and_practical_applications.pdf.

Motyl, Matt, Alexander P. Demos, Timothy S. Carsel, Brittany E. Hanson, Zachary J. Melton, Allison B. Mueller, J. P. Prims, Jiaqing Sun, Anthony N. Washburn, Kendal M. Wong,

Caitlyn Yantis und Linda J. Skitka. 2017. The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of personality and social psychology* 113: 34-58.

Nelson, Leif D., Joseph Simmons und Uri Simonsohn. 2018. Psychology's Renaissance. *Annual review of psychology* 69: 511-534.

Nosek, Brian A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. A. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson und T. Yarkoni. 2015. Promoting an open research culture. Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science* 348: 1422-1425.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven und David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*: 201708274. http://www.pnas.org/content/early/2018/03/08/1708274114.full.pdf.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: aac4716. http://science.sciencemag.org/content/349/6251/aac4716.full.pdf.

Pearl, Judea und Dana Mackenzie. 2018. *The book of why. The new science of cause and effect*. New York: Basic Books.

Rinke, Eike M. und Frank M. Schneider. 2015. *Probabilistic misconceptions are pervasive among communication researchers*. San Juan, Puerto Rico.

Rohrer, Julia, Boris Egloff und Stefan C. Schmuckle. 2017. Probing Birth-Order Effects on Narrow Traits Using Specification Curve Analysis. *Psychological science*. http://home.uni-leipzig.de/diffdiag/pppd/wp-content/uploads/bop_specification-curve_preprint.pdf.

Scheliga, Kaja und Sascha Friesike. 2014. Putting open science into practice: A social dilemma? *First Monday* 19.

Schönbrodt, Felix und David Mellor. 2018. *Academic job offers that mentioned open science*: Open Science Framework. https://osf.io/7jbnt/.

Shweder, Richard A. und Donald Winslow Fiske. 1986. Uneasy Social Science. In: Donald W. Fiske und Richard A. Shweder (Hrsg.), *Metatheory in social science. Pluralisms and subjectivities.* Chicago: Univ. of Chicago Press, 1-18.

Simmons, Joseph P., Leif D. Nelson und Uri Simonsohn. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22: 1359-1366.

Simonsohn, Uri, Joseph P. Simmons und Leif D. Nelson. 2015. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal.*

Spada, Paolo und Matt Ryan. 2017. The Failure to Examine Failures in Democratic Innovation. *PS: Political Science & Politics* 50: 772-778. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/85193D488F11DD6C4A76BD9A22BC0089/S1049096517000579a.pdf/div-class-title-the-failure-to-examine-failures-in-democratic-innovation-div.pdf.

Sparrow, Betsy. 2018. The importance of contextual relevance. *Nature Human Behaviour*: 1. https://www.nature.com/articles/s41562-018-0411-7.pdf.

Stark, Philip B. 2018. Before reproducibility must come preproducibility. *Nature* 557: 613.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman und Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on psychological science : a journal of the Association for Psychological Science* 11: 702-712.

Stockemer, Daniel, Sebastian Koehler und Tobias Lenz. 2018. Data Access, Transparency, and Replication: New Insights from the Political Behavior Literature. *PS: Political Science &*

Politics: 1-5. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/64CA07CBA652E299079FF32BC5A6DCB3/S1049096518000926a.pdf/div-class-title-data-access-transparency-and-replication-new-insights-from-the-political-behavior-literature-div.pdf.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff und Mike Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE* 6: e21101. http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0021101&type=printable.

Vadillo, Miguel A., Natalie Gold und Magda Osman. 2018. Searching for the bottom of the ego well: failure to uncover ego depletion in Many Labs 3. *Royal Society Open Science* 5: 180390.

Vries, Y. A. d., A. M. Roest, P. de Jonge, P. Cuijpers, M. R. Munafò und J. A. Bastiaansen. 2018. The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. *Psychological Medicine* 48: 2453-2455. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/71D73CADE32C0D3D996DABEA3FCDBF57/S0033291718001873a.pdf/div-class-title-the-cumulative-effect-of-reporting-and-citation-biases-on-the-apparent-efficacy-of-treatments-the-case-of-depression-div.pdf.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas und Rogier A. Kievit. 2012. An Agenda for Purely Confirmatory Research. *Perspectives on psychological science : a journal of the Association for Psychological Science* 7: 632-638.

Washburn, Anthony N., Brittany E. Hanson, Matt Motyl, Linda J. Skitka, Caitlyn Yantis, Kendal M. Wong, Jiaqing Sun, J. P. Prims, Allison B. Mueller, Zachary J. Melton und Timothy S. Carsel. 2018. Why Do Some Psychology Researchers Resist Adopting Proposed Reforms to Research Practices? A Description of Researchers' Rationales.

Advances in Methods and Practices in Psychological Science 1: 251524591875742. http://journals.sagepub.com/doi/pdf/10.1177/2515245918757427.

Weston, Sara J. und Marjan Bakker. 2018. *Preregistration hack-a-shop*: Open Science Framework. https://osf.io/vjdwm/.

Weston, Sara J., David Mellor, Marjan Bakker, Olmo van den Akker, Lorne Campbell, Stuart J. Ritchie, William J. Chopik, Rodica I. Damian, Jessica Kosie und Courtney K. Soderberg. 2018a. *Secondary Data Preregistration*: OSF. https://osf.io/x4gzt/.

Weston, Sara J., Stuart J. Ritchie, Julia M. Rohrer und Andrew K. Przybylski. 2018b. *Recommendations for increasing the transparency of analysis of pre-existing datasets*: PsyArXiv. https://psyarxiv.com/zmt3q/download?format=pdf.

Wilcox, Rand R. 2017. *Introduction to robust estimation and hypothesis testing*. Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo: Elsevier Academic Press.

Zigerell, L. J. 2017. Reducing Political Bias in Political Science Estimates. *PS: Political Science & Politics* 50: 179-183. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/16E11C12726906123D8B7408E7FCBEF2/S1049096516002389a.pdf/div-class-title-reducing-political-bias-in-political-science-estimates-div.pdf.