ANNUAL REVIEWS

Annual Review of Sociology

Reproducibility in the Social Sciences

James W. Moody,[1,2] Lisa A. Keister,[1,2,3] and Maria C. Ramos[4]

[1]Department of Sociology, Duke University, Durham, North Carolina, USA; email: jmoody77@duke.edu

[2]Duke Network Analysis Center, Duke University, Durham, North Carolina, USA

[3]Sanford School of Public Policy, Duke University, Durham, North Carolina, USA

[4]Interdisciplinary Social Science Program, Florida State University, Tallahassee, Florida, USA

Annu. Rev. Sociol. 2022. 48:65–85

First published as a Review in Advance on April 26, 2022

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-090221-035954

Keywords

data replication, reproducibility

Abstract

Concern over social scientists' inability to reproduce empirical research has spawned a vast and rapidly growing literature. The size and growth of this literature make it difficult for newly interested academics to come up to speed. Here, we provide a formal text modeling approach to characterize the entirety of the field, which allows us to summarize the breadth of this literature and identify core themes. We construct and analyze text networks built from 1,947 articles to reveal differences across social science disciplines within the body of reproducibility publications and to discuss the diversity of subtopics addressed in the literature. This field-wide view suggests that reproducibility is a heterogeneous problem with multiple sources for errors and strategies for solutions, a finding that is somewhat at odds with calls for largely passive remedies reliant on open science. We propose an alternative rigor and reproducibility model that takes an active approach to rigor prior to publication, which may overcome some of the shortfalls of the postpublication model.

INTRODUCTION

The literature on replication, rigor, and reproducibility in the social sciences has grown dramatically in recent years, with hundreds of papers on the topic published annually in a diverse set of journals across the social sciences. Although sociologists have contributed to this literature with standalone articles (e.g., Lucas et al. 2013, Mack 1951, Wilson et al. 1973), special issues (Liu & Salganik 2019, Winship 2007) and comment and reply sequences (e.g., Herring 2017, Nelson 2019, Peterson 1996, Stojmenovska et al. 2017, Weitzman 1996), the debate has largely taken place in other fields. Efforts to improve replication and reproducibility standards have been discussed and debated in both substantive and methods journals (Freese 2007, Huey & Bennell 2017). This literature is vast and growing quickly, making it difficult to get a sense of the whole, particularly as the issue often comes to most people's attention only when splashy scandals break into the science news cycle. This sort of journalistic attention favors contact with advocacy voices (such as the Center for Open Science) that tend to repeatedly highlight the same issues while ignoring the wider field of research and efforts that go unnoticed.

Here, we use a formal review method (Moody & Light 2006) to help sociologists see past the scandals and advocate voices and to provide a general sense of the reproducibility research landscape. This big-picture approach complements previous reviews that focus on in-depth readings of the most influential works (Freese & Peterson 2017) by situating them in a wider field. The target audience for this review is quantitative sociologists interested in understanding the issues at play and the history of the debates invoked in this field. The goal of the review is to broaden awareness of the richness of issues that fall under the heading of reproducibility and the corresponding depth of challenges to doing rigorous and reproducible research. Two themes emerge as a result of this overview. First, sociologists have had only a minor voice in this literature despite the fact that there is every reason to expect our work falls victim to many of the same problems that other fields face. We show that the literature has been driven mostly by experimental psychologists and political scientists: Sociologists could contribute to the conversation if more of them actively engaged with the literature. Second, it seems clear (to us) that the most commonly invoked solution to the reproducibility crisis—to increase access to data and models for public scrutiny—is laudable but incomplete given the breadth of issues that underlie reproducibility failures. The effectiveness of open science models as a policy intervention is difficult to test as it would require auditing many papers that do and do not share data (see Wicherts et al. 2011, for example). But the approach fundamentally rests on a passive model of correction after the error rather than stopping the problem beforehand. However, there is good evidence that nonreproducible research is still cited despite being found false (von Hippel 2022), which suggests a need to prevent the publication of inaccurate results in the first place.

Our work proceeds in two phases. First, we present the results of the formal text modeling to identify topics within the reproducibility field, which highlights the breadth of work being done. This work reveals a rich intellectual landscape on reproducibility in the social sciences. Second, we attempt to integrate work across the field overall, asking how its themes and foci are relevant to contemporary sociological research practices.

An important conclusion that this broad review suggests is that social scientists can and should do better, particularly with respect to model robustness and error detection, but the rhetoric around crisis is probably not helpful. The vibrant discussion on robustness and replication certainly highlights disturbing patterns that need correction, including deep issues about career incentives that drive publication bias. But the literature contains many works aimed at solving these issues creatively, signaling that the intellectual debate is working as it should. Our reading suggests that it is important to distinguish exploratory and discovery work from policy intervention or well-defined theory testing, as these two types of scientific activity require different evidentiary

stances, though both are best served by a focus on accuracy. Finally, if we want authors to fully participate in best practices regarding rigor and reproducibility, we need to work with their career incentives. These considerations point to the importance of error reduction and robustness as issues particularly pertinent to sociological concerns over replication. We conclude this review with a discussion of the promise of prepublication review as one possible solution to these issues.

SYSTEMATIC REVIEW: THE REPRODUCIBILITY INTELLECTUAL LANDSCAPE

Text network methods (Bail 2016, Light 2014) treat each paper in a corpus as a node and the similarity between papers as a link connecting them (see the **Supplemental Appendix** for details). Briefly, our process is to do an exhaustive search of scholarly indexes, filter papers for relevance (resulting in 1,947 papers), construct a network of papers based on overlapping terms, cluster the network to identify topics, and then map the network to visually represent the intellectual landscape.

Supplemental Material >

Brief Overview

Figure 1 provides a contour sociogram illustrating the global map of replication studies. In this figure, orientation is irrelevant; however, clusters near each other in this space generally share significant content whereas wide gaps in the topographic coloring indicate lack of connection. The labels refer to the 28 distinct clusters identified in the corpus.

This broad overview of the field underscores the sheer diversity of topics that comprise the landscape: We found 28 unique clusters spanning methods, misconduct and substantive topics, which range from health and policy to crime and language. At the top-center of the diagram, we find a set of papers related to data snooping, which refers to using the same dataset multiple times with different combinations of variables or coding strategies to test a given hypothesis (White 2000). Moving clockwise around the landscape, we find four clusters related to open science (OS) issues. The right side and bottom of the map reveal clusters dealing with issues of reproducibility and replication attempts in particular substantive fields. The center-left region of the map shows a common theme of statistical issues (SI), containing clusters about significance testing, statistical power, and the file drawer problem, among others. Finally, we encounter a set of clusters at the top-left corner of the diagram that consider misconduct broadly as well as specific scandals. We review the content of each cluster in detail below.

Cluster Details

The largest substantive cluster in our corpus, anchoring the bottom of **Figure 1**, contains work in applied clinical psychology ($n = 177$ papers), which emphasizes the study of alcohol, personality, and substance use disorders. In addition to alcohol-related terms, the most common terms are related to clinical assessment and item scoring tools. Representative papers[1] are on within-person replication/stability (Hopwood & Zanarini 2010),[2] are on replicating empirical findings (Bernecker et al. 2016), or are open calls for further replication and research (Shorey et al. 2016).

[1]A representative paper is one with high centrality, indicating that it is most similar to other papers in the cluster, using distinctive terms frequently. It is important to note that this is not necessarily the most cited paper on this topic.

[2]In deference to *Annual Review of Sociology* citation limits, we typically give only one example of each point. There are usually many. The table in the **Supplemental Appendix** provides a more detailed summary of each cluster.

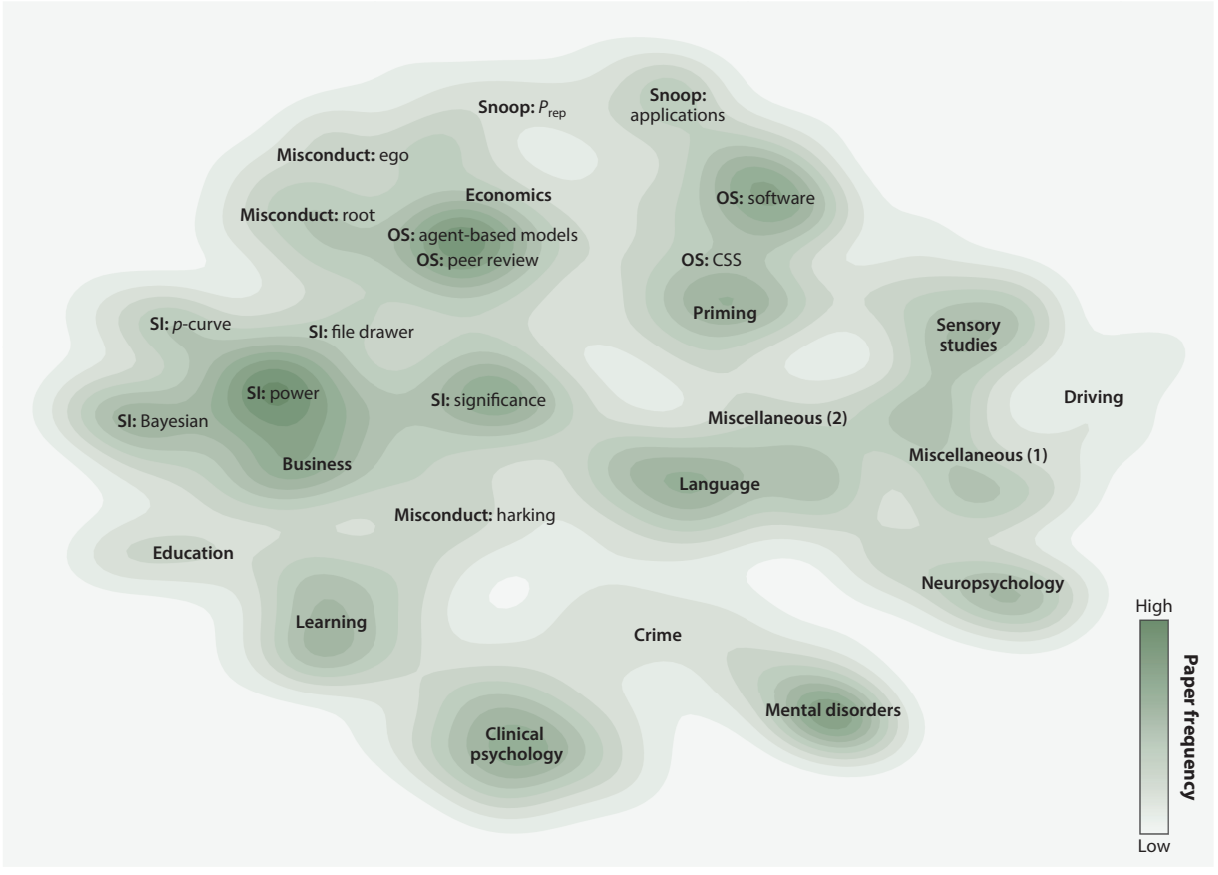

Figure 1

Replication research intellectual landscape. This is a graphical representation mapping the topic similarity network of 1,947 papers on replication in the social sciences. The space is constructed to maximize co-location of topically similar papers while placing dissimilar papers far apart, which results in topical clusters being close together. For visual clarity, we then fit a two-dimensional kernel-density estimate for the number of papers at each point in the space, so clusters emerge as hills in the map. Abbreviations: CSS, computational social science; OS, open science; SI, statistical issues.

There are few general calls to the overall problem of replication per se; rather, it seems to have been fully folded into the set of concerns guiding the everyday work of this applied field (13% of the papers in this cluster have a variant of "a replication and extension" in the paper title).

Closely related to this cluster along the bottom of **Figure 1** are topics on mental disorders ($n = 92$), learning ($n = 91$), neuropsychology ($n = 78$) and crime ($n = 52$). The most common terms in the mental disorders cluster are related to particular disorders, including "eat," "ADHD," and "attention-deficit." As above, many papers focus on item-response and scale-building efforts (Machado et al. 2018), empirical replications of prior work (Wilson et al. 2013), or internal replication of a particular association on multiple samples (Wernicke et al. 2019). Studies in the learning cluster follow a similar pattern, with fully 25% of the papers having some variant of "a replication study" in their title. Representative papers in the neuropsychology cluster again focus primarily on either replicating prior work (Powell et al. 2018) or building robust measurement tools (Dörrenberg et al. 2018). Psychology's penchant for explicit replication is high here as well, with 20% of papers self-identifying as "a replication study" in the title.

Sociologists might expect the crime cluster to include mainly criminology work, and while that work is included here, most of the work is psychological, focusing on topics at the intersection of crime, violence, and aggression. Of the 39 journals represented in this cluster, 15 have some variant of "psych" in the title and many of the remainder are violence or mental health specific outlets. The most common terms used are "offender," "abuse," "suicide," "recidivism," "maltreatment," "violence," and "homicide." The pattern here is similar to those discussed above, with many explicit replications of prior work (31% have "a replication" or similar in title).

In addition to the empirical replication attempts or internal validity replication calls for specific studies, there are a handful of methods-specific calls in these five clusters. For example, Rubio-Fernández (2018) proposes a set of methods standards aimed at improving replication, but large-scale generalized reviews of replicability results or standards are rare (but see Gelman & Geurts 2017). Overall, however, it seems that the fields dominated by psychology have taken replication to be a key methodological goal, if not an explicit standard, in their applied work. Perhaps in response to the widespread public attention to the replication crisis, these authors report that their work is internally replicated or represents an explicit failure to replicate prior work. While the tone of much of this work addresses replication as a matter of course, since having some focus on replication is a prerequisite for being in our sample, it is impossible to know from these data if this is generalizable to all studies in these areas.

The priming ($n = 80$) cluster consists mainly of psychology papers on the practice of priming subjects' mental states in experimental studies. Priming has been used widely in research on multiple social-psychological topics where some subtle treatment is used to predispose subjects toward a particular mental state or point of view. For example, one study has participants draw two points where closeness is randomly varied and finds that participants given distant points rate their psychological closeness to family lower than those given proximate points, suggesting that geometric closeness primes social closeness (Pashler et al. 2012, Williams & Bargh 2008). Priming studies have been central in the replication crisis as the practice is widely used and often substantively interesting in quirky and creative ways, but the work has often proven difficult to replicate (Caruso et al. 2017). Many of the papers in this cluster are either direct or meta-analytic studies of specific priming effects, but a handful focus on why the priming effect has proven so difficult to replicate (Cesario 2014), with many blaming some variant of not publishing insignificant results (file drawer) or reporting ex post significant results as if a priori (harking, explained below). Multiple arguments have been made for the baseline value of replication (Simons 2014) and the difficulty of doing it well (Klein et al. 2012). There is a sense from the reflective pieces in this cluster that the difficulties associated with replication have encouraged the field to be more careful in explaining elements of the experimental conditions as well as more creative in identifying contingent effects and moderators that signal how nuanced some of these effects may be.

The education ($n = 46$), business ($n = 66$), and economics ($n = 65$) clusters are internally coherent, with key terms that clearly indicate the subfield. We describe these clusters together because they are located somewhat close together in the overall field but also, more importantly, because they are social science fields that share a concern with broad issues (i.e., meta-issues) related to replication rather than relaying the results of specific replication efforts. For example, the papers in the economics cluster have titles such as "Replication Studies in Economics" (Mueller-Langer et al. 2019) and "To Replicate or Not to Replicate?" (Maniadis et al. 2017). The most central papers in the business and education clusters are equally general with titles such as "Replication in Advertising Research" (Park et al. 2015), "Replication in Strategic Management" (Hubbard et al. 1998), and "Facts are More Important than Novelty" (Makel & Plucker 2014). An important point on which these clusters differ from other clusters is the age of their most central papers: The ten most central papers in the business cluster span the past 40 years (Madden et al. 1995,

Schultz et al. 2019, Weinstein et al. 1984). By contrast, the oldest, central paper in the economics cluster is from 2008 (Stanley et al. 2008), and the oldest in education is from 2013 (Grant et al. 2013, Montgomery et al. 2013). Importantly, even the earliest central paper in business (Reid et al. 1981) deals directly with basic issues in replication of research findings, suggesting that the field had at least a foundational concern with the subject well before the current crisis.

There are two large but heterogeneous clusters that are likely largely residual in nature, labeled miscellaneous (1) ($n = 100$) and miscellaneous (2) ($n = 72$), pulled together more by their distinctiveness from the other clusters than by internal cohesion (see the **Supplemental Appendix** for details). In many ways, these papers offer a nice cross-section of substantive topics, from smoking to policing to classroom education. Unlike the other field-based clusters, there is a greater emphasis on general review and programmatic statements. For example, "Pseudoreplication is a Pseudoproblem" (Schank & Koehnle 2009) tackles a genre of data resampling used in experimental studies. Köhler & Cortina (2019) attempt to reconceptualize replication generally to focus on constructive replications. Shiffrin et al. (2018) note the importance of taking a big-picture look at the progress of science despite the inability to replicate particular findings.

Supplemental Material >

Methodological Clusters

The upper-left section of the landscape represented in **Figure 1** is dominated by general methodological and statistical issues implicated in the replication crisis, with 880 papers (~45% of the corpus). These break into four broad areas, each with topic-specific subsections. Although the clusters cover distinct topics, they overlap heavily, given that they are all based in either identifying issues that contribute to overrepresentation of nonreproducible research in the publication process or providing statistical solutions to identify these sorts of problems.

Statistical Issues

Perhaps the most common critique raised here relates to significance testing and *p*-values. The significance cluster has 92 papers, and the most representative papers are on how significance testing leads to a biased view of scientific investigations (see Sterling et al. 1995) or ruminating on the epistemological foundations for statistical testing (Rodgers 2010). This is not a new problem, of course, and many representative papers predate the current crisis (Rozeboom 1960, Sterling 1959). The fundamental difficulty is well known: Substantive effect-size arguments are often subjective, so *p*-values less than 0.05 have become a substitute decision rule over substantive value. This leads investigators to rely too heavily on statistical significance tests, either gaming to get the value below 0.05 or not considering the substantive meaning of the effect.

The cluster also includes proposals of alternatives to *p*-values. For example, many researchers propose reporting confidence intervals (Schmidt 1996), using Bayesian approaches (see below), and placing greater emphasis on the interpretation of effect sizes rather than focusing solely on whether the effect passed the significance threshold (Valentine et al. 2015). Other authors put forth broader viewpoints of how science should be done, including calls to abandon hypothesis testing entirely in favor of prediction of future outcomes (Billheimer 2019).

The importance of significance testing as a contributing factor for publication bias has led to three closely related topics: problems associated with statistical power ($n = 107$), *p*-curve analysis ($n = 33$), and Bayesian solutions ($n = 78$). There are two types of problems discussed in the power cluster. On the one hand, larger sample sizes inflate power and lead to precise, and thus statistically significant, tests for trivial effect sizes. For example, Kühberger et al. (2014) find a consistent negative correlation between effect size and sample size in psychological studies (see also Levine et al. 2009). On the other hand, small and underpowered studies will sometimes generate statistically

significant results by chance, but due to selective reporting bias, these low-powered but significant results are likely to fail replication attempts. Other work points out that multiple testing, even in the context of attempts to self-replicate internal to a study, can lead to overestimates of statistical significance (Schimmack 2012).

The *p*-curve cluster is generally about problems associated with *p*-hacking—the practice of selectively publishing only significant analyses by adjusting analyses until the effect becomes significant—and with evaluating replications more generally. The *p*-curve solution to the *p*-hacking problem assesses whether reported results are likely due to selective reporting (Simonsohn et al. 2014). The core idea is that if the results are true, we should see comparatively more *p*-values around 0.001 than 0.04, while those that have been selectively reported are more likely to have borderline *p*-value distributions. This study is well cited, and there is much discussion of when the model is applicable (Bruns & Ioannidis 2016). Similar papers argue that 0.05 should be replaced with stricter limits (Benjamin et al. 2018). This cluster also includes papers on the general issue of how to evaluate replication studies, with emphases on error in reporting and selective publication. A key point in this discussion is that the effect sizes in the original papers are stronger than those in the replication, due to either selective publication (a preference for novel and surprising results) or simple regression to the mean in replication (Fiedler & Prager 2018).

The Bayesian cluster approaches the problem of statistical significance as one that really needs approaches that sidestep binary decision rules (e.g., $p \leq 0.05$) in favor of providing more nuanced information via credible intervals, distributions, or Bayes factors. The most central paper in this cluster, Wetzels et al. (2011), evaluated 855 *t*-tests and compared different results, showing that the Bayesian models are preferred in borderline cases as they encourage more cautious conclusions. Many other papers focus on didactic instruction or attempts to persuade people to switch from frequentist to Bayesian approaches (Kruschke & Liddell 2018, Wagenmakers et al. 2018). Although there are many papers focusing on the advantages of Bayesian models, others argue that this approach simply replaces the theoretical problem of statistical decision rules with other more complicated decision rules, and as such, we cannot get around the problem of determining substantive importance (Gigerenzer & Marewski 2015).

A common theme across the SI clusters is publication bias, the lower likelihood of publishing null results, and this issue is prominent in the file drawer cluster. This is an old and well-known problem (Rosenthal 1979), and several papers attempt to test for it empirically. Gerber & Malhotra (2008) review papers published in the *American Sociological Review*, the *American Journal of Sociology*, and *Social Problems* and find there is strong evidence for general publication bias—that the distribution of *p*-values and effect sizes observed in the published record is unlikely to have been generated by a nonbiased selection process (see also Ferguson & Heene 2012). Much of the concern in this cluster, and across the publication bias themes in general, is a trade-off between novelty and accuracy (Cropley 2017), but this is a false dichotomy.

A related set of papers focus on data preparation and model building, represented in the snoop clusters. The term "data snoop" is used most frequently by financial and economics authors and refers to fitting models to the same data repeatedly. The main concern in this context is that good predictive ability within sample (or used to select the sample) could have little out-of-sample predictive validity. In financial performance models, where the goal is to identify profitable investment strategies that are necessarily out of sample, this is obviously a serious problem. The central papers in this cluster focus on proposed solutions to the problem, particularly variants on the superior predictive ability tests (Hansen 2005) that do not include data-snooping bias (Hsu & Kuan 2005). The broader general set of issues raised in this cluster relates to the problems associated with multiple testing and how it leads to bias as authors tailor model specifications to heighten *p*-values

(Young 2009, 2018).[3] A small but distinct subset of papers in this cluster are on the P_{rep} coefficient (Killeen 2005). P_{rep} was proposed as an alternative to standard null-hypothesis significance testing and garnered much attention when first introduced, though later work debated the model's validity (Iverson et al. 2009, Maraun & Gabriel 2010). While it was promising initially, the approach has not had significant uptake in practice.

Misconduct and Fraud

There are 126 papers in our corpus that fall into three clusters that focus broadly on fraud and misconduct. The most common informative terms are "fraud," "misconduct," "ethics," and "standards." Central papers are on retractions (Fang et al. 2012), how policies affect misconduct (Fanelli et al. 2015), how careful practice should be promoted (Sijtsma 2016, Waldman & Lilienfeld 2016), and theoretical models for why one might cheat (Lacetera & Zirulia 2011; the answers are high stakes and low likelihood of detection). Many of the general papers here are about identifying questionable practices and attempt to distinguish deliberate fraud from uncareful work or well-intentioned but ultimately biased practices in data analysis, preparation, or reporting (Bedeian et al. 2010, Butler et al. 2017, Rubin 2017). There are also papers on how replication issues—particularly the often messy, public nature of misconduct cases—affect trust in science and scientists (Anvari & Lakens 2018, Pickett & Roche 2018). Finally, there are papers on ethical issues in science raised by nonreproducible work and/or misconduct (DuBois & Antes 2018), as well as on general practices and standards (Mathur et al. 2019).

It is noteworthy that, in the papers related to fraud and misconduct, surveys of editors and reviewers suggest that deliberate malfeasance appears to be uncommon and not a primary driver of the inability to reproduce prior work (Hopp & Hoover 2019), though it is a common reason for retractions (Fang et al. 2012). While it is obviously difficult to know for sure, given our inability to know the true extent of deliberate fraud, taken at face value, this work suggests a mismatch between the public scandal model for nonreproducible results, which often frames the issue around questionable practices or misconduct, and the substantive problem of reproducible research. It seems likely that treating nonreproducibility as an ethical violation unnecessarily raises the reputational stakes for reproducibility, leading authors to double down on the correctness of their results.

In addition to the broad issues discussed above, the misconduct cluster has two subclusters focused on particular literatures. The first, harking ($n = 49$), is broadly concerned with post hoc data analysis and particularly focused on harking—an acronym for "hypothesizing after results are known," which refers to the practice of treating a post hoc analysis as if it were a priori (Kerr 1998). The main argument against the practice is that it misrepresents the way the work was done in a manner that is inconsistent with null-hypothesis statistical tests and conveys an idealized scientific method model that is not, in practice, followed. There is considerable debate on the extent to which this should be counted as misconduct, since there are clear cases where exploratory analyses and robustness checks are useful and necessary (Leung 2011, Rubin 2017). This practice is related to *p*-hacking, though it is more reflective of reframing an unexpected result rather than adjusting the analysis or sample to achieve a particular *p*-value. Related work notes the value of splitting samples into exploratory and confirmation sets (when data are sufficient), the general value of descriptive and exploratory work for novelty and discovery, and the value of preregistering

[3] Snooping, harking, and *p*-hacking are substantively similar—the terms all refer to adjusting analysis in an effort to get an expected result. The appearance in different clusters reflects how different social science communities use these terms.

hypotheses (Prosperi et al. 2019). The second misconduct subcluster, ego, centers on a set of papers attempting to replicate the ego depletion effect in psychology ($n = 16$). This is a now-infamous controversy over whether having a psychological store of willpower is beneficial. The inability to reproduce a finding that, until recently, was seen as a long-standing and well-known result has helped shape the tenor of the replication crisis in psychology.

Open Science Solutions

The final methodological clusters are about open science approaches framed as a generalized solution to the multiple problems of reproducibility and misconduct. The OS broad cluster has 246 papers distributed across 5 subclusters on peer review ($n = 109$), software ($n = 99$), computational social science (CSS) ($n = 20$), agent-based models ($n = 11$), and Twitter ($n = 7$; not labeled). Generally, the open science movement takes the position that current scientific publication practices overemphasize novelty at the expense of accuracy, opening the door for multiple biases that hamper reproducibility and scientific efficiency (see Munafò et al. 2017, Nosek & Bar-Anan 2012 for a concise summary of these issues). This line of work rightly points out that the incentive structure and everyday working constraints (limited resources and time) in social science promote publication of false-positive results while limiting the ability and incentives to correct such errors. For example, the Open Science Collaboration (2015) evaluated 100 empirical findings in psychology and found that over half of these were unreproducible, which has led to a broader evaluation of social science replication rates more generally with the SCORE (Systematizing Confidence in Open Research and Evidence) project from the Center for Open Science (**https://www.cos.io/score**). Most of the policies promoted within the open science framework are common-sense improvements on everyday practice. Increasing statistical training and improving transparency in reporting results are noncontroversial. But, the hallmark of the open science movement is to make results, data, programs, and workflows open to other investigators, as doing so "...offers field-wide advantages in terms of accountability, data longevity, efficiency and quality (for example, reanalyses may detect crucial mistakes or even data fabrication)" (Munafò et al. 2017, p. 6). This fully open approach—from data to publication—is a bit more controversial, as it puts new burdens on investigators with unclear or potentially negative returns, which results in generally low compliance in practice.

Papers in the peer review subcluster generally argue that open data should be part of peer review and criticize contemporary publishing in favor of models with greater transparency, data sharing, and open access (Byington & Felps 2017, Morey et al. 2016). There are multiple papers examining factors associated with data sharing (Fecher et al. 2015, Wicherts et al. 2011) and its impact (Eubank 2016, Piwowar et al. 2007). Many of the papers in this cluster focus on how difficult it is to review papers for simple reporting errors, which are a common source of nonreproducible results (Nuijten et al. 2017). For example, Nuijten et al. (2016) examined 30,717 papers in psychology between 1983 and 2015 and found that nearly half had *p*-values inconsistent with the standard errors, many of which are likely transcription errors.

The software cluster contains papers mainly on package features in the R language (Becker et al. 2017); database issues (Leeper 2014); and online, open-access tools for data sharing, reporting, and collaboration (Ram 2013). Other papers focus on often-neglected issues such as the effects of software and hardware platforms on the ability to reproduce prior work (Santana-Perez & Pérez-Hernández 2015) and open-source software tools for particular analysis problems (Warnholz & Schmid 2016) or fields (Ducke 2012, Strupler & Wilkinson 2017). The agent-based models and computational social science clusters are largely area-specific analogs to the general software cluster, focusing on issues of data sharing, code replication and building/archiving the elements of computational experiments or web searches (Eberlen et al. 2017, Nicholson 2000).

The open science papers vary topically between two rhetorical frames. First, there is a "sunlight is the best disinfectant" framing, arguing that by making data, code, and analysis tools public and shareable, errors and malfeasance will be both discouraged (because the odds of getting caught are higher) and discovered [because people will look, creating a positive feedback cycle (King 1995)]. This frame often notes that publicly funded work should be publicly accessible and that the principles underlying peer review imply actual data review. The second framing focuses on increasing efficiency and lowering barriers to research and collaboration by building shared tools. This is important for reproducibility in the sense that errors abound when people write their own code rather than build on the (presumed to be) well-vetted code of others.

IMPLICATIONS

The survey of the intellectual landscape above catalogs the scope of work on replication and reliability in the social sciences. We next step back to integrate common themes raised in the readings across topics. This aspect of our article is admittedly both more qualitative and more subjective than the systematic review, but it is critical for drawing conclusions from the review. Throughout this section, we clarify the points at which our evaluation outsteps the available evidence.

Where Are the Sociologists?

Perhaps the first takeaway from the systematic review of the literature for sociologists is just how rare it is to find sociological work represented. Sociology journals make up only about 2% of the journals in our corpus and published an even smaller percentage of papers. Indeed, only 6 of the 985 articles published in *American Journal of Sociology*, *American Sociological Review*, or *Social Forces* between 1970 and 2020 include "replication," "reproducibility," or "reanalysis" in the core search terms, based on a Web of Science search limited to these journals. Although we might expect a bias for novelty in the most prominent journals in the field, speaking as former editors of *Socius*, we note there were similarly very few submissions aimed directly at replicating prior work [excepting a special issue devoted to the topic, including the articles by Fisher (2019), Liu & Salganik (2019), and Lundberg et al. (2019)], and when such works were submitted, the authors typically had difficulty convincing reviewers that such activity was valuable. Thus, our first observation is that sociologists seem to favor novelty over replication to such a deep extent that evaluating the depth of replication success is difficult. If nobody sees value in replicating initial work, we are unlikely to find the cases that fail.

Despite the dearth of explicit replication attempts, there are at least three good reasons to be suspicious that such tests might frequently fail. The first is the finding that significance tests reported in the sociological literature have distributions consistent with a publication bias favoring barely significant results (Gerber & Malhotra 2008). This is, in our opinion, sufficient smoke to suggest fire. Second, prominent comment and reply sequences suggest the sorts of mistakes typically uncovered in the absence of careful reproduction and, ultimately, replication. These exchanges usually focus on data selection (choice of cutoff dates, outliers, etc.), coding (missing data codes, top codes), or modeling issues (convergence checks, etc.) that are necessary to produce findings.[4] Finally, the lack of concern with replication in sociology is made clear by the contrast with overt replication concerns in psychology reports. Although we cannot evaluate changes in rates of

[4]There are ongoing systematic attempts at evaluating the reproducibility of social and behavioral science, such as the Center for Open Science's SCORE project. Results are yet to be released. Other work (Anderson & Dewald 1994, Hardwicke et al. 2018) evaluates aspects of open science quality, but evaluating the effectiveness of these policies for improving replication is challenging.

replication (or success) from this corpus as constructed, the mere existence of hundreds of papers explicitly attempting replication in psychology suggests that psychology has room for this sort of work that is largely missing in sociology.

Experimental Research and Best Practice for Nonexperimentalists?

The replication crisis has been most visible in experimental studies, as made clear by the contentious work on priming and ego depletion discussed above. This work highlights how subtle experimental conditions affect results and how (generally low) statistical power in these cases, combined with publication bias favoring reports of positive results, selects for results with weak evidentiary foundations. These sorts of concerns are directly applicable to experimental sociologists, but how relevant are they to those engaged in secondary data analysis or archival/public-records work?

Our sense is that the issues of publication bias, selectivity, and reporting accuracy laid out in the review above likely hold just as well for work based on secondary data analysis as for experiments, but that there are three additional considerations that merit careful thinking on issues of previewing results, robustness checks, and coding errors. First, much of the work on snooping and harking discussed above is framed around a model of testing clear theory against newly collected data without any prior examination of the data. For much sociological research, this sort of idealized model is impractical or even impossible, because our data sources are collective goods used by many people in prior work. The community has invested heavily in these datasets, and most work will build on results from others using the same data. Innovation involves slicing the data in new ways, adding new control variables, or building new measures on otherwise well-known cases. To ensure consistency and accuracy, most researchers will (and, we think, should) reproduce prior work using the same data before embarking on these new examinations, which means that people are almost always previewing similar results before running their core statistical tests of interest. If the question of interest is a test of a well-specified theory or is directly informative to policy, then preregistration might help (see the Center for Open Science preregistration website at **https://osf.io/prereg/**), but in many cases data exploration is necessary and valuable.

Second, much of the concern with publication bias in secondary data analysis depends on investigators making more of a marginally significant result than is warranted, by, for example, publishing significant results that hold only for very narrow model specifications. A good solution to these sorts of nonrobust specifications is to explicitly address robustness with sensitivity analysis (Frank et al. 2013, Young 2018). A good approach, data availability permitting, is to use split sample or data holdout models, where exploratory work is later confirmed on a held-out sample. Unfortunately, this necessarily trades a snooping problem for a power problem, but when the power suffices, it may be effective.

Finally, large-scale reviews of statistical reporting, both of articles after publication (Ferguson & Heene 2012, Gerber & Malhotra 2008, Wetzels et al. 2011) and of prepublication code review (Eubank 2016), suggest that simple mistakes are common. Investigators often transcribe numbers incorrectly in tables or make syntax errors in coding data files that lead to mistakes and, thus, irreproducible results. Anyone who has ever had a paper returned by a copyeditor knows how humbling a second pair of eyes can be on creative work, and we should not expect analysis syntax to do much better than prose. As with spell-check or grammar correction, tools are available that help ensure syntax is accurate and tables produced with fewer hand-edits (Hlavac 2018), and we should use these whenever possible. In the end, however, these sorts of automated solutions to writing cannot replace a good copyeditor, and we expect the same in data analysis: If we are serious as a discipline about ensuring accuracy then investments in substantive prepublication code review or other significant editorial adjustments will be necessary (Colaresi 2016, Maner 2014).

Significance Testing, Power, and Effect Sizes

Empirical attempts at replication and meta-studies of such attempts all point to the problems created by a simple $p \leq 0.05$ decision rule. The conundrum here is deep: Publication is essential for academic success, so the pressure to publish is intense. At the same time, journals are overwhelmed with submissions, which means that reviewers are similarly overburdened. *p*-Values provide a simple decision rule for authors and reviewers: If something is not statistically significant, then it is ignorable. Unfortunately, this decision rule fails in the face of stochastic data-generating processes because some number of results will be statistically significant by chance alone. Since authors need to publish, they emphasize significant results, and reviewers cannot see, and would not have time to review even if they could see, the large number of tests that go unreported. The substantive effect of false positives is even more precarious when statistical power is low. Arguments to do away with significance testing entirely are tempting but not entirely convincing—the pragmatic value of a first-pass decision rule is just too high and the history too deep, though arguments for increasing the threshold might be persuasive (Benjamin et al. 2018). This is an old problem with a similarly well-known solution: Significance tests should be part of a decision about substantive significance rather than the only determining factor. Effect sizes and confidence intervals should always be presented in a way that highlights substantive importance. Whether these take Bayesian or frequentist flavors seems, to us, less important than understanding substantively what an effect size means in the theoretical context under investigation.

Incentivizing Best Practice

A common theme running through many of the statistical issues and misconduct clusters is that current incentives are misaligned and encourage publishing nonrobust results. Since top journals favor novelty over robustness and outlets for null findings are few, authors in need of publications for promotion and tenure, raises, and other incentives are compelled to search for a statistically significant finding or reframe a theory around a surprising, and likely nonrobust, finding. Although we doubt the viability of calls to overhaul the entire academic incentive structure, models that align incentives with best practices seem promising. For example, the use of registered reports—where journals review the scientific merits of the question, design, and model specification prior to authors generating results—offers a route to publishing without biasing against null findings.

A common proposed solution (Nosek & Bar-Anan 2012, Munafò et al. 2017) is to make data and code open and available for inspection and future use by interested parties (there are, of course, other reasons to support open science). There are two complementary arguments behind this recommendation. The first is a deterrence argument: When authors know their work can be checked, they should be more careful in looking for errors themselves, and some evidence is consistent with this result (Wicherts et al. 2011). The second is a collective good argument: Making the original data and code readily available eases the ability of new investigators to reproduce earlier work, minimizing the introduction of new errors, and facilitates checking the data to help root out errors. This model works well in open-source coding platforms, where users identify bugs and contribute to the creation of common utilities. The logic is that collective action in science production would work much the same way and allow for larger, more ambitious projects (Uhlmann et al. 2019).

Unlike registered reports, open data models do not clearly align author and journal incentives. Although authors of articles published in open science outlets could see increased citations to their own work based on their contributions to the publication of an article, they rarely gain authorship credit, despite making what would otherwise generally be considered an authorship-level contribution (such as writing base code or guiding data collection). Under current academic credit systems, a publication is almost always worth more than citations, at least within the range of

citations one can reasonably expect. As a solution to replication problems, making data and code openly available without restriction may be more than is necessary and work against the (perceived) self-interest of authors (Longo & Drazen 2016). This misalignment of incentives likely contributes to poor compliance (see McCullough et al. 2006, Stockemer et al. 2018, Wicherts et al. 2006, Young & Horvath 2015). For example, one study of 500 articles published in 50 high-impact journals found that 59% of articles subject to data availability policies did not fully comply, and many compliers did so by stating their willingness to share their data rather than making it available online. Only 9% of all articles in the sample made their full primary raw data available online (Alsheikh-Ali et al. 2011). Similar work finds that when datasets and code are shared, they are often incomplete or so poorly documented as to make reproduction exceedingly difficult (Carp 2012, Garijo et al. 2013, Ioannidis et al. 2009). Postpublication data sharing policies rightly include exceptions for sensitive data, but this allows researchers with sensitive data to sidestep data sharing requirements and creates perverse incentives to claim data as sensitive to avoid sharing. In contrast, authors seem more willing to share their data and code for verification when their article is under review or has been accepted than they are after the article has been published (Dewald et al. 1986). This is again understandable from an incentive alignment perspective, as discovering mistakes prior to publication is face saving, but finding mistakes after the article is published is embarrassing. To be effective, open science policies should alleviate the risk that authors perceive to their reputations and intellectual labor.

Scientific Humility in the Face of Nonreplication

The final issue that emerges from our systematic review is whether there is utility in framing replication failure as a crisis that threatens the ultimate success of science. There are two facets to this issue: First, there are questions regarding scientific accuracy and the related role of fraud, misconduct, and error. The second facet is a larger "so what?" problem that, at its core, is about the extent to which nonreplication is a new problem representing a crisis or simply the recognition that scientific frontiers are messy and rely on long-term self-correction.

The cluster on misconduct and fraud is comparatively large (roughly 6% of our corpus), and discussions of questionable research practices are found throughout the corpus, often stated in ways that impugn the motivations, skill, or character of researchers. Many of the most notorious cases of nonreplicable results take the form of public outing of results that contain data irregularities, sometimes by anonymous web critics. True levels of malfeasance are impossible to ascertain, of course, but surveys of reviewers and editors suggest that simple errors are much more likely than deliberate fraud. Prepublication audits suggest that simple transcription errors are common (Eubank 2016). Our sense is that much of the attention paid to these sorts of issues is more voyeuristic than substantive: A hint of scandal, a fall from prominence, or a David-and-Goliath struggle between (unjustly?) prominent superstars and workaday researchers makes for good science-press copy, but it is unlikely to do much to shape the nature and progress of science itself and might actually cause more distrust in science (Anvari & Lakens 2018, Wingen et al. 2020). On the other hand, if a significant portion of statistical tests are incorrect as a result of data management, programming, or transcription mistakes, there are much deeper and more systematic problems (Brown et al. 2018, Smaldino & McElreath 2016). Importantly, these simple errors may not be as publicly intriguing as deliberate malfeasance, but such errors create, at a minimum, unnecessary noise in the system and likely also distort effects and waste effort. Unfortunately, the public outing model also raises the reputational stakes of error detection, since nobody wants to find an error in work that is already published. A better model would help authors prevent errors from ever appearing in the published record in the first place and avoid having incorrect findings influence future work (von Hippel 2022).

A small but consistent subtheme running throughout the corpus is asking how much damage, overall, the failure to replicate causes. For example, Jamieson (2018) explores the narrative structure of reports on the scientific process in the science press, noting three archetypical forms—an honorable quest for truth, a dishonorable quest for truth, and science as a broken system. The fraud and misconduct studies and public outings of particular replication failures typically follow the second, while indictments of base procedures focus on the third. In both cases that she investigates closely (in the fields of oncology and psychology), the case for concern is probably overstated: "that science is broken is a generalization unwarranted by the available evidence, including that which shows a failure to replicate key studies, a rising rate of retractions, and problems in widely accepted forms of statistical inference" (Jamieson 2018, p. 2622). She notes that for science to be self-correcting, we need to be able to identify (and presumably discuss) errors in a productive way. Still others (De Winter & Happee 2013) argue that publishing all results—effectively sidestepping the file drawer problem—leads to less efficient science in the long run, as long as false positives are ultimately uncovered, and other work notes that science seems to progress well despite irreproducible results (Shiffrin et al. 2018).

Most of the commentary on replication in our corpus takes a crisis tone—and this is by far the most common rhetorical stance among papers that are proposing solutions. Our perhaps idiosyncratic reading of the breadth of work represented here leads us to a slightly different conclusion. First, we must recognize that to publish a paper proposing a solution, one first needs to convince others that there is a problem—so the prepublication bias in this case weighs in favor of seeing nonreplication as a dire threat that each new widget might alleviate. However, that we see such a vibrant discussion around the issues of accuracy, robustness, and replication suggests in many ways that the system might be working, even if it is messy. If the goal is to facilitate effective scientific progress while maintaining scientific trust, our sense is that neither the dishonest actor nor the broken system narrative helps. Rather, we need to engage the issues in a way that promotes the best of what science can do, by building on rigor and robustness while encouraging creativity, discovery, and novelty, and recognizing that novel findings will sometimes turn out to be dead ends.

The Promise of Prepublication Review

To the extent that reproducibility is hampered by honest errors, one promising alternative is to provide prepublication code and data review. Multiple journals (Christian et al. 2018, Colaresi 2016) now do a simple code review, requiring authors to submit the code and data required to produce key figures and results. These are then checked against the results reported in the paper and adjusted as necessary, which has proven effective at finding errors (Eubank 2016). Whereas most prepublication review models ensure that code runs and produces results as reported, it seems such reviews could be more extensive. For example, such a review could evaluate whether the distribution of significance tests suggests *p*-hacking, by using *p*-curves and related tools (Simonsohn et al. 2014), or conduct model specification (Young 2009, 2018) or sampling-based robustness tests. In cases where authors use large common data sources (such as the General Social Survey or similar databases), such a review could compare descriptive statistics to baselines provided by the data producer. The core advantage of prepublication review is that because it comes prior to publication, the incentives of authors and journals are aligned: Everyone wants the published record to be accurate, as it is much less embarrassing for an author to find an error before the paper is published rather than after.

The obvious downside to prepublication rigor and robustness review is that it is labor intensive and, thus, costly. Journals would have to develop tools and protocols for temporary access to confidential data, for example. However, this should be weighed against the reality that, as a

discipline, we have decided that other publication processes—copyediting, layout, and bibliometrics, for example—are acceptable and worthwhile. It may be time to include data and methods editing in this process. Finally, a prepublication robustness review would, for authors interested in participating, complement open science by ensuring that replication materials work as advertised.

CONCLUSION

We used a formal bibliometric review of 1,947 papers on replication to identify broad trends and patterns in the field. We uncovered a vibrant field, dominated by psychology but with much to say to sociologists. Our review highlighted two broad issues. First, there are substantive field-specific questions about particular results—priming being an archetypical example—that highlight how research practice might be systematically biased in favor of publishing false positives. Sociologists should pay attention to these examples—not because we necessarily care about priming, but because there is no way to ensure that models of class, race, or culture are not subject to the same publication bias that, presumably, underlies these hard-to-replicate results. This facet of the problem speaks to deep incentive structures and publication pressures in the social sciences that are unlikely to change. As such, any solution to this aspect of the replication problem needs to focus on clarity in theory and, we think, effect sizes and substantive import in addition to (or perhaps more than) statistical significance.

The extent to which publication bias problems should be concerning depends on the type of investigation at hand. Researchers who aim to directly inform policy (e.g., intervention trials) or test previously well-established or taken-for-granted theory should be extremely cautious about the dangers of practices that favor false positives and, wherever possible, take evasive action. The sort of action will vary by the type of study, but using preregistration, splitting samples, and performing model robustness checks are all promising options. However, work that aims to discover new aspects of social action or organization should be given greater rein for exploratory evaluations, though we agree that truly novel findings should be held to a high statistical bar (Benjamin et al. 2018). The root of all science is a dance between discovery and validation, and we do not want to discourage discovery in the name of validation any more than we want to pretend any single study is definitive.

The second broad issue that our review uncovered is more about reproducibility than about replicability. That is, errors in the construction and analysis of data appear to be all too common. This means that a new analyst using the same data would likely not arrive at the same result (Auspurg & Brüderl 2021, Silberzahn et al. 2018). Science's public standing as an authoritative voice depends on maintaining a level of rigor that, unfortunately, is often not met. Now that it is regularly possible for third parties to uncover such errors, we risk the authoritative reputation of science by not cleaning house and fixing such mistakes wherever possible. The open science model commonly proposed relies on people volunteering to do this sort of work, and sometimes it certainly helps. But we suspect that if we are serious about finding and correcting such errors, we need to invest in the work itself. Prepublication review aligns the interests of honest researchers with those of the public and eases issues of accessibility when authors have an incentive to help.

Despite the dominant crisis tone of much of the replication literature, the field itself is large and has produced a rich body of scholarship that promises to improve how social scientists work, though formal empirical evaluation of such efforts is challenging. This is work that sociologists have contributed to, though to a lesser degree than our colleagues in psychology. The apparent lack of concern among sociologists has left us somewhat in the dark—we simply do not know how often the results of sociological investigations can be replicated. Importantly, however, indirect evidence about the replicability of sociological work is worrisome. At the very least, sociology needs

significantly more replication attempts to know where the field stands. More broadly, this review highlights important issues regarding evidentiary value that sociologists should engage directly, rather than simply borrowing approaches and drawing conclusions by observing conversations and replication efforts occurring in other fields. Of particular note is the breadth of data types that are common in sociology, including experiments, secondary data analysis, and historical and ethnographic work. This diversity of empirical evidence suggests that, as a field, we should tailor our solutions to the problems most common in each type of analysis.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. 2011. Public availability of published research data in high-impact journals. *PLOS ONE* 6:e24357

Anderson RG, Dewald WG. 1994. Replication and scientific standards in applied economics a decade after the *Journal of Money, Credit and Banking* project. *Fed. Reserve Bank St. Louis Rev.* 76:79–83

Anvari F, Lakens D. 2018. The replicability crisis and public trust in psychological science. *Compr. Results Soc. Psychol.* 3:266–86

Auspurg K, Brüderl J. 2021. Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "Many Analysts, One Data Set" project. *Socius* 7:1–14

Bail CA. 2016. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *PNAS* 113:11823–28

Becker G, Gentleman R, Barr C, Lawrence M. 2017. Enhancing reproducibility and collaboration via management of R package cohorts. *J. Stat. Softw.* 82:1

Bedeian A, Taylor S, Miller A. 2010. Management science on the credibility bubble: cardinal sins and various misdemeanors. *Acad. Manag. Learn. Educ.* 9:715–25

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, et al. 2018. Redefine statistical significance. *Nat. Hum. Behav.* 2:6–10

Bernecker SL, Constantino MJ, Atkinson LR, Bagby RM, Ravitz P, McBride C. 2016. Attachment style as a moderating influence on the efficacy of cognitive-behavioral and interpersonal psychotherapy for depression: a failure to replicate. *Psychotherapy* 53:22–33

Billheimer D. 2019. Predictive inference and scientific reproducibility. *Am. Stat.* 73:291–95

Brown AW, Kaiser KA, Allison DB. 2018. Issues with data and analyses: errors, underlying themes, and potential solutions. *PNAS* 115:2563–70

Bruns SB, Ioannidis JP. 2016. *p*-Curve and *p*-hacking in observational research. *PLOS ONE* 11:e0149144

Butler N, Delaney H, Spoelstra S. 2017. The gray zone: questionable research practices in the business school. *Acad. Manag. Learn. Educ.* 16:94–109

Byington EK, Felps W. 2017. Solutions to the credibility crisis in management science. *Acad. Manag. Learn. Educ.* 16:142–62

Carp J. 2012. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* 63:289–300

Caruso EM, Shapira O, Landy JF. 2017. Show me the money: a systematic exploration of manipulations, moderators, and mechanisms of priming effects. *Psychol. Sci.* 28:1148–59

Cesario J. 2014. Priming, replication, and the hardest science. *Perspect. Psychol. Sci.* 9:40–48

Christian T-ML, Lafferty-Hess S, Jacoby WG, Carsey TM. 2018. *Operationalizing the replication standard: a case study of the data curation and verification workflow for scholarly journals*. Work. Pap., SocArXiv. **https://doi.org/10.31235/osf.io/cfdba**

Colaresi M. 2016. Preplication, replication: a proposal to efficiently upgrade journal replication standards. *Int. Stud. Perspect.* 17:367–78

Cropley A. 2017. The specter of scholarship without novel ideas: replication, hyperauthorship and the danger of stagnation. *Psychol. Aesthet. Creat. Arts* 11:69–76
De Winter JCF, Happee R. 2013. Why selective publication of statistically significant results can be effective. *PLOS ONE* 8:e66463
Dewald WG, Thursby JG, Anderson RG. 1986. Replication in empirical economics: the *Journal of Money, Credit and Banking* project. *Am. Econ. Rev.* 76:587–603
Dörrenberg S, Rakoczy H, Liszkowski U. 2018. How (not) to measure infant theory of mind: testing the replicability and validity of four non-verbal measures. *Cogn. Dev.* 46:12–30
DuBois JM, Antes AL. 2018. Five dimensions of research ethics: a stakeholder framework for creating a climate of research integrity. *Acad. Med.* 93:550–55
Ducke B. 2012. Natives of a connected world: free and open source software in archaeology. *World Archaeol.* 44:571–79
Eberlen J, Scholz G, Gagliolo M. 2017. Simulate this! An introduction to agent-based models and their power to improve your research practice. *Int. Rev. Soc. Psychol.* 30:149–60
Eubank N. 2016. Lessons from a decade of replications at the *Quarterly Journal of Political Science*. *Political Sci. Politics* 49:273–76
Fanelli D, Costas R, Larivière V. 2015. Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLOS ONE* 10:e0127556
Fang FC, Steen RG, Casadevall A. 2012. Misconduct accounts for the majority of retracted scientific publications. *PNAS* 109:17028–33
Fecher B, Friesike S, Hebing M. 2015. What drives academic data sharing? *PLOS ONE* 10:e0118053
Ferguson CJ, Heene M. 2012. A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspect. Psychol. Sci.* 7:555–61
Fiedler K, Prager J. 2018. The regression trap and other pitfalls of replication science—illustrated by the report of the Open Science Collaboration. *Basic Appl. Soc. Psychol.* 40:115–24
Fisher JC. 2019. Data-specific functions: a comment on Kindel et al. *Socius* 5:1–6
Frank KA, Maroulis SJ, Duong MQ, Kelcey BM. 2013. What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educ. Eval. Policy Anal.* 35:437–60
Freese J. 2007. Replication standards for quantitative social science: Why not sociology? *Sociol. Methods Res.* 36:153–72
Freese J, Peterson D. 2017. Replication in social science. *Annu. Rev. Sociol.* 43:147–65
Garijo D, Kinnings S, Xie L, Xie L, Zhang Y, et al. 2013. Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. *PLOS ONE* 8:e80278
Gelman A, Geurts HM. 2017. The statistical crisis in science: How is it relevant to clinical neuropsychology? *Clin. Neuropsychol.* 31:1000–14
Gerber AS, Malhotra N. 2008. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociol. Methods Res.* 37:3–30
Gigerenzer G, Marewski JN. 2015. Surrogate science: the idol of a universal method for scientific inference. *J. Manag.* 41:421–40
Grant S, Mayo-Wilson E, Hopewell S, Macdonald G, Moher D, Montgomery P. 2013. Developing a reporting guideline for social and psychological intervention trials. *J. Exp. Criminol.* 9:355–67
Hansen PR. 2005. A test for superior predictive ability. *J. Bus. Econ. Stat.* 23:365–80
Hardwicke TE, Mathur MB, MacDonald K, Nilsonne G, Banks GC, et al. 2018. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *R. Soc. Open Sci.* 5:180448
Herring C. 2017. Is diversity still a good thing? *Am. Sociol. Rev.* 82:868–77
Hlavac M. 2018. stargazer: Well-formatted regression and summary statistics tables. *R package*, version 5.2.1. **https://CRAN.R-project.org/package=stargazer**
Hopp C, Hoover GA. 2019. What crisis? Management researchers' experiences with and views of scholarly misconduct. *Sci. Eng. Ethics* 25:1549–88
Hopwood CJ, Zanarini MC. 2010. Five-factor trait instability in borderline relative to other personality disorders. *Personal. Disord.* 1:58–66

Hsu PH, Kuan CM. 2005. Reexamining the profitability of technical analysis with data snooping checks. *J. Financ. Econom.* 3:606–28

Hubbard R, Vetter DE, Little EL. 1998. Replication in strategic management: scientific testing for validity, generalizability, and usefulness. *Strateg. Manag. J.* 19:243–54

Huey L, Bennell C. 2017. Replication and reproduction in Canadian policing research: a note. *Can. J. Criminol. Crim. Justice* 59:123–38

Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, et al. 2009. Repeatability of published microarray gene expression analyses. *Nat. Genet.* 41:149–55

Iverson GJ, Lee MD, Wagenmakers EJ. 2009. p_{rep} misestimates the probability of replication. *Psychon. Bull. Rev.* 16:424–29

Jamieson KH. 2018. Crisis or self-correction: rethinking media narratives about the well-being of science. *PNAS* 115:2620–27

Kerr NL. 1998. HARKing: hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* 2:196–217

Killeen PR. 2005. An alternative to null-hypothesis significance tests. *Psychol. Sci.* 16:345–53

King G. 1995. Replication, replication. *Political Sci. Politics* 28:444–52

Klein O, Doyen S, Leys C, Magalhães de Saldanha da Gama PA, Miller S, et al. 2012. Low hopes, high expectations: expectancy effects and the replicability of behavioral experiments. *Perspect. Psychol. Sci.* 7:572–84

Köhler T, Cortina JM. 2019. Play it again, Sam! An analysis of constructive replication in the organizational sciences. *J. Manag.* 47:488–518

Kruschke JK, Liddell TM. 2018. Bayesian data analysis for newcomers. *Psychon. Bull. Rev.* 25:155–77

Kühberger A, Fritz A, Scherndl T. 2014. Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PLOS ONE* 9:e105825

Lacetera N, Zirulia L. 2011. The economics of scientific misconduct. *J. Law Econ. Organ.* 27:568–603

Leeper TJ. 2014. Archiving reproducible research with R and Dataverse. *R J.* 6:151–58

Leung K. 2011. Presenting post hoc hypotheses as a priori: ethical and theoretical issues. *Manag. Organ. Rev.* 7:471–79

Levine TR, Asada KJ, Carpenter C. 2009. Sample sizes and effect sizes are negatively correlated in meta-analyses: evidence and implications of a publication bias against nonsignificant findings. *Commun. Monogr.* 76:286–302

Light R. 2014. From words to networks and back. *Soc. Curr.* 1:111–29

Liu DM, Salganik MJ. 2019. Successes and struggles with computational reproducibility: lessons from the Fragile Families Challenge. *Socius* 5:1–21

Longo DL, Drazen JM. 2016. Data sharing. *N. Engl. J. Med.* 374:276–77

Lucas JW, Morrell K, Posard M. 2013. Considerations on the 'replication problem' in sociology. *Am. Sociol.* 44:217–32

Lundberg I, Narayanan A, Levy K, Salganik MJ. 2019. Privacy, ethics, and data access: a case study of the Fragile Families Challenge. *Socius* 5:1–25

Machado PPP, Grilo CM, Crosby RD. 2018. Replication of a modified factor structure for the Eating Disorder Examination-Questionnaire: extension to clinical eating disorder and non-clinical samples in Portugal. *Eur. Eat. Disord. Rev.* 26:75–80

Mack RW. 1951. The need for replication research in sociology. *Am. Sociol. Rev.* 16:93–94

Madden CS, Easley RW, Dunn MG. 1995. How journal editors view replication research. *J. Advert.* 24:77–87

Makel MC, Plucker JA. 2014. Facts are more important than novelty: replication in the education sciences. *Educ. Res.* 43:304–16

Maner JK. 2014. Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspect. Psychol. Sci.* 9:343–51

Maniadis Z, Tufano F, List JA. 2017. To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *Econ. J.* 127:F209–35

Maraun M, Gabriel S. 2010. Killeen's 2005 p_{rep} coefficient: logical and mathematical problems. *Psychol. Methods* 15:182–91

Mathur A, Lean SF, Maun C, Walker N, Cano A, Wood ME. 2019. Research ethics in inter- and multi-disciplinary teams: differences in disciplinary interpretations. *PLOS ONE* 14:e0225837

McCullough BD, McGeary KA, Harrison TD. 2006. Lessons from the *JMCB* archive. *J. Money Credit Banking* 38:1093–107
Montgomery P, Grant S, Hopewell S, Macdonald G, Moher D, et al. 2013. Protocol for CONSORT-SPI: an extension for social and psychological interventions. *Implement. Sci.* 8:99
Moody J, Light R. 2006. A view from above: the evolving sociological landscape. *Am. Sociol.* 37:67–86
Morey RD, Chambers CD, Etchells PJ, Harris CR, Hoekstra R, et al. 2016. The peer reviewers' openness initiative: incentivizing open research practices through peer review. *R. Soc. Open Sci.* 3:150547
Mueller-Langer F, Fecher B, Harhoff D, Wagner GG. 2019. Replication studies in economics—How many and which papers are chosen for replication, and why? *Res. Policy* 48:62–83
Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, et al. 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* 1:0021
Nelson LK. 2019. To measure meaning in big data, don't give me a map, give me transparency and reproducibility. *Sociol. Methodol.* 49:139–43
Nicholson S. 2000. Raising reliability of web search tool research through replication and chaos theory. *J. Am. Soc. Inf. Sci. Technol.* 51:724–29
Nosek BA, Bar-Anan Y. 2012. Scientific utopia: I. Opening scientific communication. *Psychol. Inq.* 23:217–43
Nuijten MB, Borghuis J, Veldkamp CLS, Dominguez-Alvarez L, Van Assen MALM, Wicherts JM. 2017. Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra Psychol.* 3:31
Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* 48:1205–26
Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349:aac4716
Park JH, Venger O, Park DY, Reid LN. 2015. Replication in advertising research, 1980–2012: a longitudinal analysis of leading advertising journals. *J. Curr. Issues Res. Advert.* 36:115–35
Pashler H, Coburn N, Harris CR. 2012. Priming of social distance? Failure to replicate effects on social and food judgments. *PLOS ONE* 7:e42510
Peterson RR. 1996. Statistical errors, faulty conclusions, misguided policy: reply to Weitzman. *Am. Sociol. Rev.* 61:539–40
Pickett JT, Roche SP. 2018. Questionable, objectionable or criminal? Public opinion on data fraud and selective reporting in science. *Sci. Eng. Ethics* 24:151–71
Piwowar HA, Day RS, Fridsma DB. 2007. Sharing detailed research data is associated with increased citation rate. *PLOS ONE* 2:e308
Powell LJ, Hobbs K, Bardis A, Carey S, Saxe R. 2018. Replications of implicit theory of mind tasks with varying representational demands. *Cogn. Dev.* 46:40–50
Prosperi M, Bian J, Buchan IE, Koopman JS, Sperrin M, Wang M. 2019. Raiders of the lost HARK: a reproducible inference framework for big data science. *Palgrave Commun.* 5:125
Ram K. 2013. Git can facilitate greater reproducibility and increased transparency in science. *Source Code Biol. Med.* 8:7
Reid LN, Soley LC, Wimmer RD. 1981. Replication in advertising research. *J. Advert.* 10:3–13
Rodgers JL. 2010. The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *Am. Psychol.* 65:1–12
Rosenthal R. 1979. The file drawer problem and tolerance for null results. *Psychol. Bull.* 86:638–41
Rozeboom WW. 1960. The fallacy of the null-hypothesis significance test. *Psychol. Bull.* 57:416–28
Rubin M. 2017. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Rev. Gen. Psychol.* 21:308–20
Rubio-Fernández P. 2018. Publication standards in infancy research: three ways to make violation-of-expectation studies more reliable. *Infant Behav. Dev.* 54:177–88
Santana-Perez I, Pérez-Hernández MS. 2015. Towards reproducibility in scientific workflows: an infrastructure-based approach. *Sci. Program.* 2015:243180
Schank JC, Koehnle TJ. 2009. Pseudoreplication is a pseudoproblem. *J. Comp. Psychol.* 123:421–33
Schimmack U. 2012. The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* 17:551–66

Schmidt FL. 1996. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol. Methods* 1:115–29
Schultz DE, Kerr G, Kitchen P. 2019. Replication and George the Galapagos tortoise. *J. Mark. Commun.* **https://doi.org/10.1080/13527266.2019.1658465**
Shiffrin RM, Borner K, Stigler SM. 2018. Scientific progress despite irreproducibility: a seeming paradox. *PNAS* 115:2632–39
Shorey RC, Dawson AE, Haynes E, Strauss C, Elmquist J, et al. 2016. Is general or alcohol-specific perceived social support associated with depression among adults in substance use treatment? *J. Psychoact. Drugs* 48:359–68
Sijtsma K. 2016. Playing with data—or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika* 81:1–15
Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, et al. 2018. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* 1:337–56
Simons DJ. 2014. The value of direct replication. *Perspect. Psychol. Sci.* 9:76–80
Simonsohn U, Nelson LD, Simmons JP. 2014. *p*-Curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* 143:534–47
Smaldino PE, McElreath R. 2016. The natural selection of bad science. *R. Soc. Open Sci.* 3:160384
Stanley TD, Doucouliagos C, Jarrell SB. 2008. Meta-regression analysis as the socio-economics of economics research. *J. Socio-Econ.* 37:276–92
Sterling TD. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* 54:30–34
Sterling TD, Rosenbaum WL, Weinkam JJ. 1995. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *Am. Stat.* 49:108–12
Stockemer D, Koehler S, Lentz T. 2018. Data access, transparency, and replication: new insights from the political behavior literature. *Political Sci. Politics* 51(4):799–803
Stojmenovska D, Bol T, Leopold T. 2017. Does diversity pay? A replication of Herring 2009. *Am. Sociol. Rev.* 82:857–67
Strupler N, Wilkinson TC. 2017. Reproducibility in the field: transparency, version control and collaboration on the Project Panormos Survey. *Open Archaeol.* 3:279–304
Uhlmann EL, Ebersole CR, Chartier CR, et al. 2019. Scientific utopia III: crowdsourcing science. *Perspect. Psychol. Sci.* 14:711–33
Valentine JC, Aloe AM, Lau TS. 2015. Life after NHST: how to describe your data without "*p*-ing" everywhere. *Basic Appl. Soc. Psychol.* 37:260–73
von Hippel PT. 2022. Is psychological science self-correcting? Citations before and after successful and failed replications. *Perspect. Psychol. Sci.* In press
Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, et al. 2018. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* 25:35–57
Waldman ID, Lilienfeld SO. 2016. Thinking about data, research methods, and statistical analyses: commentary on Sijtsma's 2014 "Playing with Data." *Psychometrika* 81:16–26
Warnholz S, Schmid T. 2016. Simulation tools for small area estimation: introducing the R-package saeSim. *Österr. Z. Stat.* 45:55–69
Weinstein S, Drozdenko R, Weinstein C. 1984. Brain wave analysis in advertising research: validation from basic research & independent replications. *Psychol. Mark.* 1:83–95
Weitzman LJ. 1996. The economic consequences of divorce are still unequal: comment on Peterson. *Am. Sociol. Rev.* 61:537
Wernicke J, Li M, Sha P, Zhou M, Sindermann C, et al. 2019. Individual differences in tendencies to attention-deficit/hyperactivity disorder and emotionality: empirical evidence in young healthy adults from Germany and China. *Atten. Deficit Hyperact. Disord.* 11:167–82
Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. 2011. Statistical evidence in experimental psychology: an empirical comparison using 855 *t* tests. *Perspect. Psychol. Sci.* 6:291–98
White H. 2000. A reality check for data snooping. *Econometrica* 68:1097–126
Wicherts JM, Bakker M, Molenaar D. 2011. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE* 6:e26828

Wicherts JM, Borsboom D, Kats J, Molenaar D. 2006. The poor availability of psychological research data for reanalysis. *Am. Psychol.* 61:726–28

Williams LE, Bargh JA. 2008. Keeping one's distance: the influence of spatial distance cues on affect and evaluation. *Psychol. Sci.* 19:302–08

Wilson BA, Dhamapurkar S, Tunnard C, Watson P, Florschutz G. 2013. The effect of positioning on the level of arousal and awareness in patients in the vegetative state or the minimally conscious state: a replication and extension of a previous finding. *Brain Impair.* 14:475–79

Wilson FD, Smoke GL, Martin JD. 1973. The replication problem in sociology: a report and a suggestion. *Sociol. Inquiry* 43:141–49

Wingen T, Berkessel JB, Englich B. 2020. No replication, no trust? How low replicability influences trust in psychology. *Soc. Psychol. Pers. Sci.* 11:454–63

Winship C. 2007. Introduction to the special section on replication and data access. *Sociol. Methods Res.* 36:151–52

Young C. 2009. Model uncertainty in sociological research: an application to religion and economic growth. *Am. Sociol. Rev.* 74:380–97

Young C. 2018. Model uncertainty and the crisis in science. *Socius* 4:1–7

Young C, Horvath A. 2015. Sociologists need to be better at replication. *Orgtheory.net blog*, Aug. 11. **https://orgtheory.wordpress.com/2015/08/11/sociologists-need-to-be-better-at-replication-a-guest-post-by-cristobal-young/**